The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Robust Lightweight Facial Expression Recognition Network with Label Distribution Training

Zengqun Zhao, Qingshan Liu[*] and Feng Zhou
B-DAT Lab, Nanjing University of Information Science & Technology, Nanjing, China
{zqzhao, qsliu}@nuist.edu.cn

Abstract

This paper presents an efficiently robust facial expression recognition (FER) network, named EfficientFace, which holds much fewer parameters but more accurate and robust to the FER in the wild. Firstly, to improve the robustness of the lightweight network, a local-feature extractor and a channel-spatial modulator are designed, in which the depthwise convolution is employed. As a result, the network is aware of local and global-salient facial features. Then, considering the fact that most emotions occur as combinations, mixtures, or compounds of the basic emotions, we introduce a simple but efficient label distribution learning (LDL) method as a novel training strategy. Experiments conducted on realistic occlusion and pose variation datasets demonstrate that the proposed EfficientFace is robust under occlusion and pose variation conditions. Moreover, the proposed method achieves state-of-the-art results on RAF-DB, CAER-S, and AffectNet-7 datasets with accuracies of 88.36%, 85.87%, and 63.70%, respectively, and a comparable result on the AffectNet-8 dataset with an accuracy of 59.89%. The code and training logs are available at https://github.com/zengqunzhao/EfficientFace.

Introduction

Facial expression plays a vital role in communication, and automatically recognize facial expression is extremely crucial for its applications in various fields. In the field of human-computer interaction (HCI), the environments adaptive systems, socially aware systems, or robots with social skills can be built by detecting user's affective states (Maat and Pantic 2007; DeVault et al. 2014; Corneanu et al. 2016). In the field of education, detecting students' frustration can help improve e-learning experience (Kapoor, Burleson, and Picard 2007). In the field of medicine, the pain detection is used for monitoring patient progress in the clinical setting (Lucey et al. 2010; Kaltwang, Rudovic, and Pantic 2012; Irani et al. 2015), the facial attributes such as expressions, action units, arousal, and valence are used for classifying autism spectrum disorder (ASD) (Wang et al. 2004; Dapretto et al. 2006; Loth et al. 2018; Li et al. 2019a). In the field of driver assistance, monitoring drowsiness or attentive and emotional status of the driver is critical for the safety and comfort of driving (Khan et al. 2018; Jeong and Ko 2018).

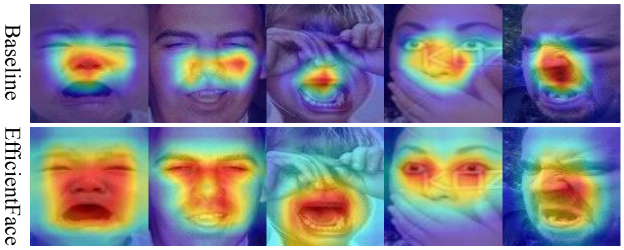

Figure 1: The comparison of CAM between the baseline and proposed EfficientFace. The baseline denotes ShuffleNet-V2. The images are from the test set of the RAF-DB dataset.

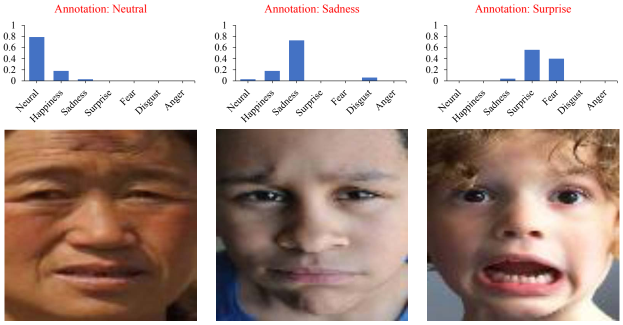

Figure 2: The real-world facial expression can be viewed as a combination of basic expressions. The label distributions on the facial image are the output of the network trained in our framework.

Because deep learning has outstanding capacity in learning visual semantic features (LeCun, Bengio, and Hinton 2015), it was also successfully employed to facial expression recognition (FER) and has achieved excellent progress (Zhao et al. 2016; Yu et al. 2018; Wang, Wang, and Liang 2019; Fu et al. 2020; Wang et al. 2020a; Wang, Shuai, and Liu 2020; Chen et al. 2020b). However, deep learning models generally have excessive parameters and FLOPs, which are inconvenient for practical application. To improve the efficiency of the FER model, some researchers have attempted to design the real-time FER model from the temporal per-

[*] Corresponding author

spective (Miao et al. 2019; Lee and Wong 2020; Koujan et al. 2020). But it is more essential to design the lightweight FER model in the spatial level, namely designing of the lightweight static FER model. To achieve this purpose, it is intuitive to utilize the state-of-the-art lightweight models which are proposed by researchers on model compression, such as the MobileNet series (Howard et al. 2017; Sandler et al. 2018), and ShuffleNet series (Zhang et al. 2018; Ma et al. 2018). However, due to the limited capacity of the lightweight networks in feature learning and the challenges like occlusion and pose variation existed in FER in the wild, employing the aforementioned networks to FER directly may entail poor performance both in accuracy and robustness. Therefore, a few researchers have paid attention to designing lightweight models for static FER. Hewitt and Gunes (Hewitt and Gunes 2018) designed three kinds of lightweight FER models on mobile devices, and Barros *et al.* (Barros, Churamani, and Sciutti 2020) proposed a lightweight FER model named FaceChannel. However, the accuracy and robustness of these methods are yet inferior. Ferro-Pérez and Mitre-Hernandez (Ferro-Pérez and Mitre-Hernandez 2020) proposed ResMoNet for FER, which can be used in resource-limited systems. Regrettably, the ResMoNet is just evaluated on one lab-controlled FER dataset but have not been compared with the state-of-the-art FER methods.

To address the dilemma between saving computational overheads and enhancing performance, based on the lightweight backbone network, we propose a more robust and accurate network from the perspective of the feature extraction and training strategy. First of all, to obtain robust facial features, given the characteristics of the human face and issues of occlusion and pose variation in the wild, a channel-spatial modulated and locality-aware lightweight network is proposed, the proposed network is aware of the global-salient and local facial features. Concretely, a local-feature extractor and a channel-spatial modulator are designed and implemented in the backbone network. Regarding the FER in the wild, such a method can enhance the performance of the lightweight model prominently but with a negligible increase of computational overheads. The visualization results of class activation mapping (CAM) (Zhou et al. 2016) are displayed in Figure 1.

Additionally, the psychologist's study (Plutchik 1980) and previous FER work (Zhou, Xue, and Geng 2015; Jia et al. 2019) have shown that the most emotions occur as combinations, mixtures, or compounds of the basic emotions, and multiple emotions always have different intensities in a single facial image, especially in the real world (see Figure 2). Therefore, to further improve the performance of the FER model, training the FER models by label distribution called label distribution learning (LDL) instead of a single label seems more reasonable. Furthermore, there are researches have demonstrated that the LDL can also address the noise problem caused by the subjectiveness of annotators and ambiguous in facial images (Gao et al. 2017; Chen et al. 2020a). For the LDL, a key issue is how to construct label distribution. Previous work acquired label distribution by utilizing distribution annotations (Zhou, Xue, and Geng 2015; Jia et al. 2019), face affinity graph (He et al. 2017), label smoothing (Gao et al. 2017; Ling and Geng 2019) or auxiliary label space graphs (Chen et al. 2020a). Different from these methods, we propose a simple but efficient method by training a deep convolutional neural network (DCNN) named label distribution generator (LDG) to generate the label distribution directly. The experiments demonstrate that training with generated label distribution can enhance the performance of the lightweight networks remarkably.

In summary, this paper has the following contributions:

- A novel lightweight network named EfficientFace for practical facial expression recognition is presented. The proposed EfficientFace with few parameters and FLOPs, designed from the view of the feature extraction and training strategy, is robust and accurate for FER in the wild.
- A local-feature extractor and a channel-spatial modulator are designed from the view of the feature extraction to learn local facial features and global-salient features. As a result, the network can learn comprehensive facial features and is robust under occlusion and pose variation conditions. In the view of the training strategy, a simple but efficient LDL method is proposed by designing the LDG, which enhanced the performance of the EfficientFace remarkably.
- Experiments conducted on realistic occlusion and pose variation datasets indicate that the proposed method is robust towards occlusion and pose variation problems. With few parameters and FLOPs, our method achieves state-of-the-art results on RAF-DB, CAER-S, and AffectNet-7 datasets, and a comparable result on the AffectNet-8 dataset.

Related Work

Lightweight FER Models

For efficient FER models, only a little work has focused on the design of the lightweight FER model. Hewitt and Gunes (Hewitt and Gunes 2018) proposed three variants of established CNN architectures, namely the AlexNet variant, VGGNet variant, and MobileNet variant, which are utilized to FER on mobile devices. Barros *et al.* (Barros, Churamani, and Sciutti 2020) proposed a lightweight FER model, named FaceChannel, which has ten convolutional layers including four pooling layers and applied shunting inhibitory fields in the last layer. Ferro-Pérez and Mitre-Hernandez (Ferro-Pérez and Mitre-Hernandez 2020) proposed a network named ResMoNet for FER in resource-limited systems. Specifically, the proposed ResMoNet is composed of five types of block, namely Stem Block, Mobile Block, Residual Block, Transition Block, and Dense Block, and generates a less number of parameters and multi-add operations. However, these methods or have inferior performance or are not evaluated on in-the-wild FER datasets.

Methods of Label Distribution Learning

Zhou *et al.* (Zhou, Xue, and Geng 2015) introduced the label distribution learning into FER and achieved better performance than the single-label learning method. In their study,

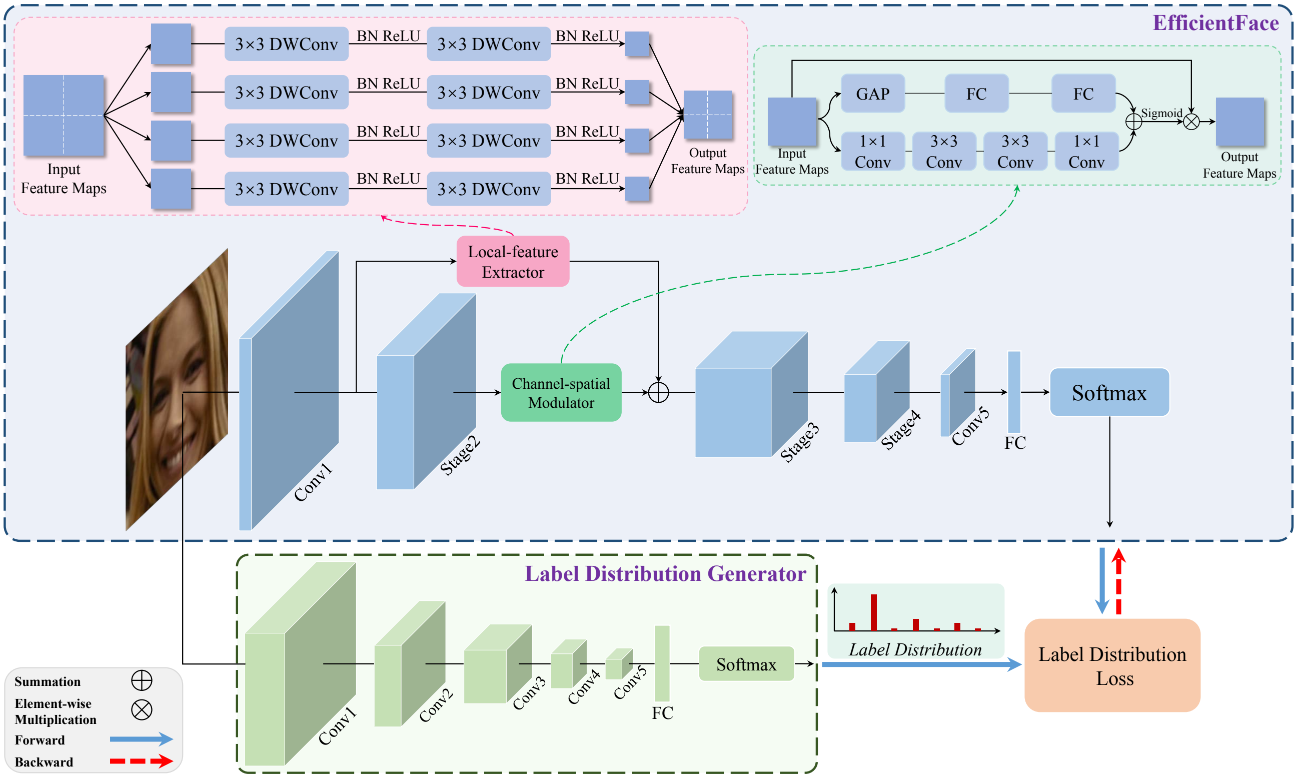

Figure 3: The overall structure of the proposed method. The proposed method is composed of two parts: the EfficientFace employed to recognize facial expression and the Label Distribution Generator (LDG) employed to generate label distribution as a ground-truth for training EfficientFace. DWConv denotes depthwise convolution, GAP denotes global average pooling, and FC denotes a fully-connected network.

the label distributions come from the annotations. Due to the difficulty of annotating label distribution for each image, some researchers have attempted to construct label distribution for images automatically. Gao *et al.* (Gao et al. 2017) utilized the normal distribution to construct label distribution based on the single label, and they applied such a method in age estimation, head pose estimation, multi-label classification, and semantic segmentation to address the label ambiguity. For the particularity of the human face, He *et al.* (He et al. 2017) utilized a face affinity graph to construct label distribution for age estimation. Ling *et al.* (Ling and Geng 2019) constructed label distribution by the discretized Gaussian distribution with adaptive variance. The latest work (Chen et al. 2020a) proposed auxiliary label space graphs based on facial action units and landmark for label distribution learning.

Method

To decrease the computational overheads of the FER model, the state-of-the-art lightweight network ShuffleNet-V2, consists of Conv1, Stage2, Stage3, Stage4, and Conv5, is employed as a backbone network in our method. To handle the problems of occlusion and pose variation existed in the real-world scenario, a local-feature extractor and a channel-spatial modulator are designed. Furthermore, a novel label distribution learning method is proposed, which is consistent with the psychologist's theory (Plutchik 1980). In the following subsections, we first introduce the proposed local-feature extractor and channel-spatial modulator, then we give the details of the proposed label distribution generator and label distribution loss. The overall structure of the proposed method is shown in Figure 3.

Channel-Spatial Modulated and Locality-Aware Lightweight Network

Previous work has indicated that learning local facial features is favorable to FER in the wild (Li et al. 2019c; Wang et al. 2020b). However, those methods acquiring local features based on the facial landmarks, which is inefficient. Therefore, we propose an efficient local-feature extractor to learn local region features, and the local region features are obtained at a feature level and supplemented into global features in a residual form. The architecture of the local-feature extractor is shown in the upper left of Figure 3. Specifically, given a input facial image with size of $224 \times 224 \times 3$, after the Conv1, the low-level global feature maps $F_{conv1} \in \mathbb{R}^{H \times W \times C}$ can be obtained, where $H = W = 56, C = 29$. The F_{conv1} is first split into four feature patches and each feature patch denoted by $F^i_{conv1} \in \mathbb{R}^{H' \times W' \times C}$, where $i \in \{1, 2, 3, 4\}$, $H' = W' = 28$. Then, each feature patch

F^i_{conv1} is processed by two corresponding 3×3 convolutions. The depthwise convolution is used to decrease computation overheads. After two convolution layers, the learned local facial features $F^i_{local} \in \mathbb{R}^{H'/2 \times W'/2 \times C'}$ can be obtained, where $C' = 116$. The four local facial features are finally concatenated along the spatial axis and then the local features $F_{local} \in \mathbb{R}^{H' \times W' \times C'}$ are obtained. It should mention that dividing the feature maps into four patches conforms the action units related to expression.

Because there is a lot of redundant information in global facial features (Zhong et al. 2012, 2014), the channel-spatial modulator is introduced to highlight the crucial global features after the Stage2. Allow for computational overheads, the channel-spatial modulator is designed based on BAM (Park et al. 2018). The architecture of the channel-spatial modulator is shown in the upper right of Figure 3. Concretely, after the Stage2, the high-level global feature maps $F_{stage2} \in \mathbb{R}^{H' \times W' \times C'}$ are obtained, firstly, the channel heatmaps $M_c(F_{stage2}) \in \mathbb{R}^{C'}$ and the spatial heatmaps $M_s(F_{stage2}) \in \mathbb{R}^{H' \times W'}$ are computed at two paralleled branches respectively, then the whole heatmaps $M(F_{stage2})$ can be computed and normalized as:

$$M(F_{stage2}) = \sigma(M_c(F_{stage2}) + M_s(F_{stage2})) \quad (1)$$

where σ is a sigmoid function. The channel heatmaps $M_c(F_{stage2})$ and spatial heatmaps $M_s(F_{stage2})$ are all resized to $\mathbb{R}^{H' \times W' \times C'}$ before addition. Finally, the modulated global feature maps can be computed as:

$$F_{modulated} = F_{stage2} \otimes M(F_{stage2}) \quad (2)$$

where $\otimes$ denotes element-wise multiplication.

After the extraction of local features and modulation of global features, the final global-local features can be computed as:

$$F_{final} = F_{local} + F_{modulated} \quad (3)$$

As a result, the network achieves the ability to learn both global-salient and local facial features but with a slight increase in computation overheads. It should be noted that the local-feature extractor and channel-spatial modulator are adopted only after Stage2. The reason why employing such a strategy is: on the one hand, excessive use of designed modules will increase the overheads of the FER model; on the other hand, the size of the deeper feature maps is tiny, and those feature maps are not feasible for extracting local facial features.

The comparison of CAM between the baseline and EfficientFace shown in Figure 1. Compared with the baseline network, the EfficientFace-based CAM results pay attention to broader regions related to facial expression under frontal and non-frontal conditions. Moreover, the proposed model can focus on the non-occlusion regions and neglect the occlusion regions under occlusion conditions.

Constructing Label Distribution and Loss Function

Due to the difficulty of annotation, the emotion distributions of the facial images are often unavailable. To this end, we design a label distribution generator (LDG) to generate label distribution for training. The structure of the LDG is presented in Figure 3. The proposed LDG is pre-trained on FER datasets and fixed in training phase.

Given a facial image s with label $l \in \{0, 1, ..., c-1\}$, where c is the number of the expression categories. The function of the LDG is to generate a distribution $\boldsymbol{d} = (d_0, d_1, ..., d_{c-1})$, where $\sum_{i=0}^{c-1} d_i = 1$. After the FC layer of LDG, a feature vector $\boldsymbol{v} = (v_0, v_1, ..., v_{c-1})$ can be obtained, then the $\boldsymbol{d}$ can be computed by a softmax function:

$$d_i = \frac{exp(v_i)}{\sum_{j=0}^{c-1} exp(v_j)} \quad (4)$$

where $i \in \{0, 1, ..., c-1\}$

Regarding the conventional FER method, the single label l is used as a ground-truth called single label learning (SLL). While in our LDL method, the generated label distribution is adopted as a ground-truth. The cross-entropy is employed to measure the distance between the prediction of EfficentFace and the output of LDG. Hence, a label distribution loss can be defined as:

$$\mathcal{L} = -\frac{1}{N \times c} \sum_{i=0}^{N-1} \sum_{j=0}^{c-1} d^i_j log(\widetilde{d^i_j}) \quad (5)$$

where N is the number of samples; $\widetilde{\boldsymbol{d}} = (\widetilde{d_1}, \widetilde{d_2}, ..., \widetilde{d_{c-1}})$ is the predicted label distribution of EfficientFace and $\sum_{j=0}^{c-1} \widetilde{d_j} = 1$; the superscript i and subscript j denote sample and expression category respectively.

It should be noted that the LDG is trained using a single label and employed only in the training phase. After the EfficientFace is fully trained, only the EfficientFace is needed and abandoning the rest of the part. In the inference phase, the FER result of a facial image is the index of the maximum probability in $\widetilde{\boldsymbol{d}} = (\widetilde{d_0}, \widetilde{d_1}, ..., \widetilde{d_{c-1}})$.

Experiments

Datasets

To verify the effectiveness of the proposed method, we conduct the experiments on three popular in-the-wild facial expression datasets: RAF-DB (Li and Deng 2018), CAER-S (Lee et al. 2019), and AffectNet (Mollahosseini, Hasani, and Mahoor 2017), and five realistic occlusion and pose variation datasets: FED-RO (Li et al. 2019c), Occlusion-AffectNet, Occlusion-RAF-DB, Pose-AffectNet and Pose-RAF-DB (Wang et al. 2020b).

RAF-DB The RAF-DB dataset contains 30,000 facial images annotated with basic or compound expressions by 40 trained human coders. Consistent with the most previous work, only images with seven basic emotions, i.e., neutral, happiness, sadness, surprise, fear, disgust, and anger, are used, including 12,271 images as training data and 3,068 images as test data.

Datasets	Methods	# Params (M)	# MFLOPs	Accuracy (%)
RAF-DB	Baseline	1.26	147.79	82.23
	Baseline + Local-Feature Extractor	1.27	152.79	83.57
	Baseline + Channel-Spatial Modulator	1.27	153.79	83.12
	Baseline + Local-Feature Extractor + Channel-Spatial Modulator	1.28	154.18	**83.83**
CAER-S	Baseline	1.26	147.79	79.52
	Baseline + Local-Feature Extractor	1.27	152.79	80.68
	Baseline + Channel-Spatial Modulator	1.27	153.79	80.48
	Baseline + Local-Feature Extractor + Channel-Spatial Modulator	1.28	154.18	**81.48**

Table 1: Evaluation of each component in EfficientFace on RAF-DB and CAER-S. The models are trained from scratch.

Datasets	Methods	# Params (M)	# MFLOPs	Accuracy (%)
RAF-DB	LDG (SLL)	23.52	4109.48	86.93
	Baseline (SLL)	1.26	147.79	84.58
	Baseline (LDL)	1.26	147.79	87.87
	EfficientFace (SLL)	1.28	154.18	85.66
	EfficientFace (LDL)	1.28	154.18	**88.36**
CAER-S	LDG (SLL)	23.52	4109.48	84.81
	Baseline (SLL)	1.26	147.79	84.06
	Baseline (LDL)	1.26	147.79	85.74
	EfficientFace (SLL)	1.28	154.18	84.51
	EfficientFace (LDL)	1.28	154.18	**85.87**

Table 2: Evaluation of the proposed label distribution learning on RAF-DB and CAER-S. The models are pre-trained on the MS-Celeb-1M dataset. The SLL and LDL denote single label learning and label distribution learning respectively.

CAER-S The CAER-S dataset was created by selecting static frames from the CAER (Lee et al. 2019) dataset with 65,983 images and has been divided into two sets: training set (44,996 samples) and test set (20,987 samples). Each image is assigned to one of seven basic expressions.

AffectNet The AffectNet dataset contains about 450,000 images that are manually annotated with 11 expression categories. The seven expression categories denoted by AffectNet-7 contain seven basic expressions, while the eight expression categories denoted by AffectNet-8 with the addition of contempt expression. For AffectNet-7, there are 283,901 images as training data and 3,500 images as test data, and for AffectNet-8, there are 287,568 images as training data and 4,000 images as test data.

Realistic Occlusion and Pose Variation Datasets To examine the performance of the FER model under real-world occlusion and pose variation conditions, Li *et al.* (Li et al. 2019c) collected and annotated a facial expression dataset with real occlusion (FED-RO) in the wild. The FED-RO contains 400 images in total, and the images were categorized into seven basic expressions. Wang *et al.* (Wang et al. 2020b) built four subsets, Occlusion-AffectNet, Pose-AffectNet, Occlusion-RAF-DB, and Pose-RAF-DB, from the validation set of AffectNet-8 and the test set of RAF-DB respectively. The Occlusion-AffectNet and Occlusion-RAF-DB contain 683 and 735 images in total respectively. The Pose-AffectNet contains 1,948 and 985 faces with an angle larger than 30° and 45° respectively in total, and the Pose-RAF-DB contains 1,248 and 558 faces with an angle larger than 30° and 45° respectively in total.

Experiment Setting

For images on all the datasets, the face region is detected and aligned using Retinaface (Deng et al. 2020) and then cropped and resized to 224×224 pixels. The random cropping and random horizontal flipping are employed to avoid over-fitting. The proposed EfficientFace and LDG are all pre-trained on the face recognition dataset MS-Celeb-1M (Guo et al. 2016). For LDG, the 50-layer Residual Network is adopted as the backbone network. For EfficientFace, parameters were optimized via the SGD optimizer with an initial learning rate of 0.1 and a mini-batch size of 128. All the models are trained on the NVIDIA GeForce Titan Xp GPU based on the open-source PyTorch (Paszke et al. 2019) platform.

Ablation Studies

Effectiveness of Each Component in EfficientFace To verify the effectiveness of each component in EfficientFace, we conduct ablation studies on RAF-DB and CAER-S datasets. Specifically, the ShuffleNet-V2 is employed as a baseline in experiments, then the local-feature extractor and channel-spatial modulator are added to the baseline, respectively. The four kinds of methods are trained with the same setting and using the conventional single label for training. As shown in Table 1, due to the use of depthwise convolution in the local-feature extractor and the lightweight of the channel-spatial modulator, the proposed EfficientFace achieves a tiny increase of computation overheads compared

Methods	# Params (M)	# MFLOPs	Accuracy (%)
IPA2LT (Zeng, Shan, and Chen 2018)	>23.52	>4109.48	86.77
Separate-Loss (Li et al. 2019b)	11.18	1818.56	86.38
gACNN (Li et al. 2019c)	>134.29	>15479.79	85.07
RAN (Wang et al. 2020b)	11.19	14548.45	86.90
LDL-ALSG (Chen et al. 2020a)	23.52	4109.48	85.53
DDA-Lose (Hossein and Qi 2020)	11.18	1818.56	86.90
SCN (Wang et al. 2020a)	∼11.18	∼1818.56	87.03
SCN* (Wang et al. 2020a)	∼11.18	∼1818.56	88.14
EfficientFace (Ours)	**1.28**	**154.18**	**88.36**

Table 3: Comparison with state-of-the-art methods on RAF-DB. * RAF-DB and AffectNet are jointly used for training.

Methods	# Params (M)	# MFLOPs	Accuracy (%)
ResNet-18 (He et al. 2016)	11.18	1818.56	84.67
ResNet-50 (He et al. 2016)	23.52	4109.48	84.81
MobileNet-V2 (Sandler et al. 2018)	2.23	312.86	79.23
Res2Net-50 (Gao et al. 2019)	23.66	4278.60	85.35
CAER-Net-S (Lee et al. 2019)	∼2.12	∼1717.65	73.51
EfficientFace (Ours)	**1.28**	**154.18**	**85.87**

Table 4: Comparison with state-of-the-art methods on CAER-S.

with the baseline network. When the local region features are considered by implementing local-feature extractor, the recognition accuracy outperforms the baseline network by 1.34% and 1.16% on RAF-DB and CAER-S, respectively. Moreover, the channel-spatial modulator enhances the accuracy of the baseline by 0.89% and 0.96% on RAF-DB and CAER-S, respectively. The reason why EfficientFace achieves better performance lies in the supplement of local facial features and the attention of the global-salient facial features, such two techniques not only contribute to the comprehensiveness of the facial feature but also can handle the occlusion and pose variance problems well.

Effectiveness of Label Distribution Learning To verify the effectiveness of the proposed label distribution learning method, we conduct ablation studies on RAF-DB and CAER-S datasets. The comparison between label distribution learning (LDL) in which label distribution is obtained by the LDG and conventional single label learning (SLL) is presented. And the comparison between LDL-based baseline and LDL-based EfficientFace is presented as well. As shown in Table 2, regarding the proposed EfficientFace, the LDL-based accuracies outperform the SLL-based accuracies by 2.70% and 1.36% on RAF-DB and CEAR-S datasets, respectively. Even for the baseline network, the performance utilizing LDL is much better than SLL. Moreover, it is gratifying that the performance of the LDL-based EfficientFace outperforms the LDG even though the parameters and FLOPs of the EfficientFace far less than LDG. The reason why LDL-based methods achieve excellent performance lies in two aspects: one is that label distribution is more close to the real-world human expression which occurs as combinations of the basic emotions, the other is that the label distribution is generated by the LDG which is capable of relieving label noise problem (Gao et al. 2017). Hence, the LDL-based method is adopted in the following experiments.

Comparison with State of the Arts

We compare the proposed method to several state-of-the-art methods on RAF-DB, CAER-S, AffectNet-7, and AffectNet-8 datasets. Due to the fact that most previous work in FER do not pay attention to the computation overheads, those methods do not present the metric of computation complexity. Hence, we compute the number of parameters and FLOPs of all the models compared in experiments using the same setting. For the methods of IPA2LT (Zeng, Shan, and Chen 2018) and gACNN (Li et al. 2019c), only the lower limit values are provided (the symbol > is used). Due to the methods of SCN (Wang et al. 2020a), CAER-Net-S (Lee et al. 2019), SNA-DFER (Fu et al. 2020), VGGNet-Variant (Hewitt and Gunes 2018), and Mobile-Variant (Hewitt and Gunes 2018) do not have publicly available implementations, the codes of those methods are first reproduced to the best of our understanding and then computed the number of parameters and FLOPs (the symbol ∼ is used). The remainder is all computed according to the open-source code.

Table 3 and Table 4 show the results on RAF-DB and CAER-S, respectively, and Table 5 shows the results on AffectNet-7 and AffectNet-8. Due to the CAER-S dataset is proposed recently, and only (Lee et al. 2019) evaluates their method on it, we conduct several experiments utilizing some state-of-the-art networks on it, such as ResNet-18, ResNet-50, MobileNet-V2, and Res2Net-50. For AffectNet, some work verifies their methods on AffectNet-7 but others on AffectNet-8, while we conduct experiments and compare with other state-of-the-art methods both on AffectNet-7 and AffectNet-8. Due to the AffectNet dataset has an imbalanced training set but a balanced validation set, consistent

Datasets	Methods	# Params (M)	# MFLOPs	Accuracy (%)
AffectNet-7	IPA2LT (Zeng, Shan, and Chen 2018)	>23.52	>4109.48	57.31
	gACNN (Li et al. 2019c)	>134.29	>15479.79	58.78
	IPFR (Wang, Wang, and Liang 2019)	21.80	5729.12	57.40
	Separate-Loss (Li et al. 2019b)	11.18	1818.56	58.89
	FMPN (Chen et al. 2019)	21.80	5729.117	61.52
	VGG-FACE (Kollias et al. 2020)	145.00	15490.46	60.00
	SNA-DFER (Fu et al. 2020)	∼2.47	∼763.09	62.70
	LDL-ALSG (Chen et al. 2020a)	23.52	4109.48	59.35
	DDA-Loss (Hossein and Qi 2020)	11.18	1818.56	62.34
	EfficientFace (Ours)	**1.28**	**154.18**	**63.70**
AffectNet-8	Weighted-Loss (Mollahosseini, Hasani, and Mahoor 2017)	57.03	710624.57	58.00
	VGGNet-Variant (Hewitt and Gunes 2018)	∼6.54	∼80.44	58.00
	MobileNet-Variant (Hewitt and Gunes 2018)	∼**0.074**	∼**13.56**	56.00
	RAN (Wang et al. 2020b)	11.19	14548.45	59.50
	ESR-9 (Siqueira, Magg, and Wermter 2020)	0.37	1164.43	59.30
	SCN (Wang et al. 2020a)	∼11.18	∼1818.56	**60.23**
	EfficientFace (Ours)	1.28	154.18	59.89

Table 5: Comparison with state-of-the-art methods on AffectNet-7 and AffectNet-8.

Datasets	Methods	# Params (M)	# MFLOPs	Accuracy (%)
FED-RO	VGG-16 (Simonyan and Zisserman 2014)	134.29	15479.79	60.15
	ResNet-18 (He et al. 2016)	11.18	1818.56	64.25
	gACNN (Li et al. 2019c)	>134.29	>15479.79	66.50
	RAN (Wang et al. 2020b)	11.19	14548.45	67.98
	EfficientFace (Ours)	**1.28**	**154.18**	**68.25**
Occlusion-AffectNet	ResNet-18 (He et al. 2016)	11.18	1818.56	49.48
	RAN (Wang et al. 2020b)	11.19	14548.45	58.50
	EfficientFace (Ours)	**1.28**	**154.18**	**59.88**
Occlusion-RAF-DB	ResNet-18 (He et al. 2016)	11.18	1818.56	80.19
	RAN (Wang et al. 2020b)	11.19	14548.45	82.72
	EfficientFace (Ours)	**1.28**	**154.18**	**83.24**

Table 6: Comparison with state-of-the-art methods on FED-RO, Occlusion-AffectNet, and Occlusion-RAF-DB.

Datasets	Methods	# Paprams (M)	# MFLOPs	Accuracy (%)	
				Pose $\geqslant 30^\circ$	Pose $\geqslant 45^\circ$
Pose-AffectNet	ResNet-18 (He et al. 2016)	11.18	1818.56	50.10	48.50
	RAN (Wang et al. 2020b)	11.19	14548.45	53.90	53.19
	EfficientFace (Ours)	**1.28**	**154.18**	**57.36**	**56.87**
Pose-RAF-DB	ResNet-18 (He et al. 2016)	11.18	1818.56	84.04	83.15
	RAN (Wang et al. 2020b)	11.19	14548.45	86.74	85.20
	EfficientFace (Ours)	**1.28**	**154.18**	**88.13**	**86.92**

Table 7: Comparison with state-of-the-art methods on Pose-AffectNet and Pose-RAF-DB.

with RAN (Wang et al. 2020b) and SCN (Wang et al. 2020a), an oversampling strategy [1] is employed.

From Table 3, Table 4, and Table 5 it can be witnessed that the proposed method outperforms all of these state-of-the-art methods in terms of accuracy on RAF-DB, CAER-S, and AffectNet-7. Furthermore, our method possesses the minimum number of parameters and FLOPs among all compared methods. That is to say, our method is excellent both in saving computation overheads and recognition accuracy. For AffectNet-8 our method achieves a comparable result. It is noteworthy that the proposed method has a large gap of accuracy between AffectNet-7 and AffectNet-8, in which the AffectNet-8 added expression of contempt based on AffectNet-7. We argue that there exists a lot of noise annotations of the eighth expression categories in the AffectNet test data (see Figure 4).

[1] https://github.com/ufoym/imbalanced-dataset-sampler

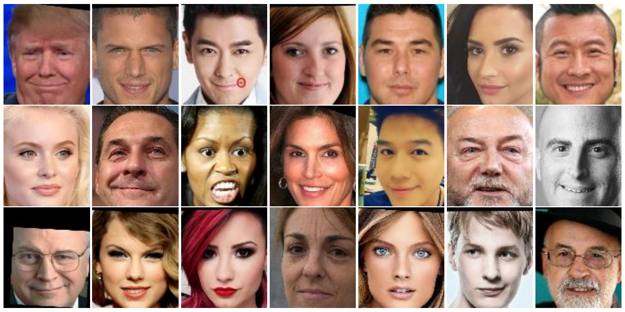

Figure 4: Facial images with the annotation of contempt were selected randomly from the test data of the AffectNet dataset. It can be seen that many facial images do not belong to the expression of contempt.

Robustness Under Realistic Occlusion and Pose Variation Conditions

Occlusion and pose variation are two vital issues for FER in the real world. To evaluate the proposed method under the real-world scenario, we conduct several experiments on datasets with realistic occlusion and pose variation.

Evaluation Under Realistic Occlusion To evaluate the proposed method under the realistic occlusion condition, several experiments are conducted on FED-RO, Occlusion-AffectNet, and Occlusion-RAF-DB datasets, and the experiment setting is the same as the previous work. As shown in Table 6, the proposed method achieves the best performance on FED-RO, Occlusion-AffectNet, and Occlusion-RAF-DB datasets both in computation overheads and accuracy. These results prove that the proposed method has superior robustness under occlusion conditions.

Evaluation Under Realistic Pose Variation To evaluate the proposed method under realistic pose variation conditions, we conduct experiments on Pose-AffectNet and Pose-RAF-DB. From Table 7 it can be caught that the performance of the proposed method is superior to compared methods. Moreover, comparing the accuracy between angle larger than $30°$ and $45°$, the proposed method achieves a tinier gap of accuracy, which indicates that our method possesses fantastic robustness to pose variation.

Conclusion

This paper proposes a lightweight and robust facial expression recognition network named EfficientFace for practical FER in the wild. Specifically, in the view of the feature extraction, a local-feature extractor and a channel-spatial modulator are proposed. As a result, the learned facial features stay robust under occlusion and pose variation conditions. Owing to the real-world facial expression is more like a distribution instead of a single emotion, and the noise problem existed on facial expression datasets, a novel LDL method is proposed by designing of the LDG. Even with few parameters and FLOPs, the proposed method achieves state-of-the-art results on several datasets compared with those methods which have enormous parameters and FLOPs. And extensive experiments conducted on realistic occlusion and pose variation datasets also indicate that the excellent robustness of the proposed method.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61825601 and in part by the Natural Science Foundation of Jiangsu Province (NSF-JS) under Grant BK20192004B.

References

Barros, P.; Churamani, N.; and Sciutti, A. 2020. The FaceChannel: A Light-Weight Deep Neural Network for Facial Expression Recognition. In *FG*, 449–453.

Chen, S.; Wang, J.; Chen, Y.; Shi, Z.; Geng, X.; and Rui, Y. 2020a. Label Distribution Learning on Auxiliary Label Space Graphs for Facial Expression Recognition. In *CVPR*, 13984–13993.

Chen, W.; Zhang, D.; Li, M.; and Lee, D.-J. 2020b. STCAM: Spatial-Temporal and Channel Attention Module for Dynamic Facial Expression Recognition. *IEEE Transactions on Affective Computing* .

Chen, Y.; Wang, J.; Chen, S.; Shi, Z.; and Cai, J. 2019. Facial motion prior networks for facial expression recognition. In *VCIP*, 1–4.

Corneanu, C. A.; Simón, M. O.; Cohn, J. F.; and Guerrero, S. E. 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(8): 1548–1568.

Dapretto, M.; Davies, M. S.; Pfeifer, J. H.; Scott, A. A.; Sigman, M.; Bookheimer, S. Y.; and Iacoboni, M. 2006. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* 9(1): 28–30.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *CVPR*, 5203–5212.

DeVault, D.; Artstein, R.; Benn, G.; Dey, T.; Fast, E.; Gainer, A.; Georgila, K.; Gratch, J.; Hartholt, A.; Lhommet, M.; et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *FG*, 1061–1068.

Ferro-Pérez, R.; and Mitre-Hernandez, H. 2020. ResMoNet: A Residual Mobile-based Network for Facial Emotion Recognition in Resource-Limited Systems. *arXiv preprint arXiv:2005.07649* .

Fu, Y.; Wu, X.; Li, X.; Pan, Z.; and Luo, D. 2020. Semantic Neighborhood-Aware Deep Facial Expression Recognition. *IEEE Transactions on Image Processing* .

Gao, B.-B.; Xing, C.; Xie, C.-W.; Wu, J.; and Geng, X. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26(6): 2825–2838.

Gao, S.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; and Torr, P. H. 2019. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *ECCV*, 87–102.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, Z.; Li, X.; Zhang, Z.; Wu, F.; Geng, X.; Zhang, Y.; Yang, M.-H.; and Zhuang, Y. 2017. Data-dependent label distribution learning for age estimation. *IEEE Transactions on Image processing* 26(8): 3846–3858.

Hewitt, C.; and Gunes, H. 2018. CNN-based facial affect analysis on mobile devices. *arXiv preprint arXiv:1807.08775* .

Hossein, F. A.; and Qi, X. 2020. Discriminant Distribution-Agnostic Loss for Facial Expression Recognition in the Wild. In *CVPRW*, 406–407.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* .

Irani, R.; Nasrollahi, K.; Simon, M. O.; Corneanu, C. A.; Escalera, S.; Bahnsen, C.; Lundtoft, D. H.; Moeslund, T. B.; Pedersen, T. L.; Klitgaard, M.-L.; et al. 2015. Spatiotemporal analysis of RGB-DT facial images for multimodal pain level recognition. In *CVPRW*, 88–95.

Jeong, M.; and Ko, B. C. 2018. Driver's facial expression recognition in real-time for safe driving. *Sensors* 18(12): 4270.

Jia, X.; Zheng, X.; Li, W.; Zhang, C.; and Li, Z. 2019. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *CVPR*, 9841–9850.

Kaltwang, S.; Rudovic, O.; and Pantic, M. 2012. Continuous pain intensity estimation from facial expressions. In *ISVC*, 368–377.

Kapoor, A.; Burleson, W.; and Picard, R. W. 2007. Automatic prediction of frustration. *International Journal of Human-Computer Studies* 65(8): 724–736.

Khan, S. A.; Hussain, S.; Xiaoming, S.; and Yang, S. 2018. An effective framework for driver fatigue recognition based on intelligent facial expressions analysis. *IEEE Access* 6: 67459–67468.

Kollias, D.; Cheng, S.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision* 1–30.

Koujan, M.; Alharbawee, L.; Giannakakis, G.; Pugeault, N.; and Roussos, A. 2020. Real-Time Facial Expression Recognition" In The Wild" by Disentangling 3D Expression from Identity. In *FG*, 539–546.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553): 436–444.

Lee, J.; Kim, S.; Kim, S.; Park, J.; and Sohn, K. 2019. Context-aware emotion recognition networks. In *ICCV*, 10143–10152.

Lee, J. R. H.; and Wong, A. 2020. TimeConvNets: A Deep Time Windowed Convolution Neural Network Design for Real-time Video Facial Expression Recognition. In *CRV*, 9–16.

Li, B.; Mehta, S.; Aneja, D.; Foster, C.; Ventola, P.; Shic, F.; and Shapiro, L. 2019a. A Facial Affect Analysis System for Autism Spectrum Disorder. In *ICIP*, 4549–4553.

Li, S.; and Deng, W. 2018. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* 28(1): 356–370.

Li, Y.; Lu, Y.; Li, J.; and Lu, G. 2019b. Separate Loss for Basic and Compound Facial Expression Recognition in the Wild. In *ACML*, 897–911.

Li, Y.; Zeng, J.; Shan, S.; and Chen, X. 2019c. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing* 28(5): 2439–2450.

Ling, M.; and Geng, X. 2019. Indoor Crowd Counting by Mixture of Gaussians Label Distribution Learning. *IEEE Transactions on Image Processing* 28(11): 5691–5701.

Loth, E.; Garrido, L.; Ahmad, J.; Watson, E.; Duff, A.; and Duchaine, B. 2018. Facial expression recognition as a candidate marker for autism spectrum disorder: how frequent and severe are deficits? *Molecular Autism* 9(1): 7.

Lucey, P.; Cohn, J. F.; Matthews, I.; Lucey, S.; Sridharan, S.; Howlett, J.; and Prkachin, K. M. 2010. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41(3): 664–674.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. ShuffleNet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 116–131.

Maat, L.; and Pantic, M. 2007. Gaze-X: Adaptive, affective, multimodal interface for single-user office scenarios. In *Artifical Intelligence for Human Computing*, 251–271. Springer.

Miao, Y.; Dong, H.; Jaam, J. M. A.; and Saddik, A. E. 2019. A Deep Learning System for Recognizing Facial Expression in Real-Time. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15(2): 1–20.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10(1): 18–31.

Park, J.; Woo, S.; Lee, J.-Y.; and Kweon, I.-S. 2018. BAM: Bottleneck Attention Module. In *BMVC*, 1–14.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 8026–8037.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, 3–33. Elsevier.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNet v2: Inverted residuals and linear bottlenecks. In *CVPR*, 4510–4520.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *ICLR* 1–14.

Siqueira, H.; Magg, S.; and Wermter, S. 2020. Efficient Facial Feature Learning with Wide Ensemble-based Convolutional Neural Networks. In *AAAI*, 5800–5809.

Wang, A. T.; Dapretto, M.; Hariri, A. R.; Sigman, M.; and Bookheimer, S. Y. 2004. Neural correlates of facial affect processing in children and adolescents with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* 43(4): 481–490.

Wang, C.; Wang, S.; and Liang, G. 2019. Identity-and Pose-Robust Facial Expression Recognition through Adversarial Feature Learning. In *MM*, 238–246.

Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020a. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, 6897–6906.

Wang, K.; Peng, X.; Yang, J.; Meng, D.; and Qiao, Y. 2020b. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing* 29: 4057–4069.

Wang, S.; Shuai, H.; and Liu, Q. 2020. Phase Space Reconstruction Driven Spatio-Temporal Feature Learning for Dynamic Facial Expression Recognition. *IEEE Transactions on Affective Computing* .

Yu, Z.; Liu, G.; Liu, Q.; and Deng, J. 2018. Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing* 317: 50–57.

Zeng, J.; Shan, S.; and Chen, X. 2018. Facial expression recognition with inconsistently annotated datasets. In *ECCV*, 222–237.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 6848–6856.

Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; and Yan, S. 2016. Peak-piloted deep network for facial expression recognition. In *ECCV*, 425–442.

Zhong, L.; Liu, Q.; Yang, P.; Huang, J.; and Metaxas, D. N. 2014. Learning multiscale active facial patches for expression analysis. *IEEE Transactions on Cybernetics* 45(8): 1499–1510.

Zhong, L.; Liu, Q.; Yang, P.; Liu, B.; Huang, J.; and Metaxas, D. N. 2012. Learning active facial patches for expression analysis. In *CVPR*, 2562–2569.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.

Zhou, Y.; Xue, H.; and Geng, X. 2015. Emotion distribution recognition from facial expressions. In *MM*, 1247–1250.