AU-Aware Vision Transformers for Biased Facial Expression Recognition

Shuyi Mao, Xinpeng Li, Qingyang Wu, and Xiaojiang Peng, *Member, IEEE*

***Abstract*—Studies have proven that domain bias and label bias exist in different Facial Expression Recognition (FER) datasets, making it hard to improve the performance of a specific dataset by adding other datasets. For the FER bias issue, recent researches mainly focus on the cross-domain issue with advanced domain adaption algorithms. This paper addresses another problem: how to boost FER performance by leveraging cross-domain datasets. Unlike the coarse and biased expression label, the facial Action Unit (AU) is fine-grained and objective suggested by psychological studies. Motivated by this, we resort to the AU information of different FER datasets for performance boosting and make contributions as follows. First, we experimentally show that the naive joint training of multiple FER datasets is harmful to the FER performance of individual datasets. We further introduce expression-specific mean images and AU cosine distances to measure FER dataset bias. This novel measurement shows consistent conclusions with experimental degradation of joint training. Second, we propose a simple yet conceptually-new framework, AU-aware Vision Transformer (AU-ViT). It improves the performance of individual datasets by jointly training auxiliary datasets with AU or pseudo-AU labels. We also find that the AU-ViT is robust to real-world occlusions. Moreover, for the first time, we prove that a carefully-initialized ViT achieves comparable performance to advanced deep convolutional networks. Our AU-ViT achieves state-of-the-art performance on three popular datasets, namely 91.10% on RAF-DB, 65.59% on AffectNet, and 90.15% on FERPlus. The code and models will be released soon.**

***Index Terms*—Facial expression recognition, facial action unit, datasets bias, multiple databases, Transformer.**

I. Introduction

Facial Expressions Recognition (FER) aims to identify people's emotions in a facial image. It plays a great role in various applications including surveillance [10], games [35], educations [54], medical treatments [35] and marketing [37]. In the past decade, more and more scholars in computer vision have devoted themselves to this task and contributed a lot. [11], [33], [38], [57], [25] Recently, many studies show one of the main challenges is data bias, which mainly comes from data composition [26], [8], [34] and label ambiguity (e.g. subjective annotation [56], [9] and compound expression [27]).

Dataset bias results in sub-optimal performance and even serious deterioration in joint dataset training. Existing works have achieved a lot in eliminating dataset bias, which can be mainly categorized into (1) label modification [34], [56], [47], [7], [44], [9] and (2) cross-domain adaptation [26], [8], [55]. In label modification, the algorithms use web images, attention mechanisms, EM iteration, or distribution learning to correct biased labels. In cross-domain adaptation, the algorithms adopt regularization terms, adversarial learning, or meta mechanism to minimize domain divergence. However, they neither boost a target dataset's performance by leveraging auxiliary datasets nor bridge different datasets with objective information.

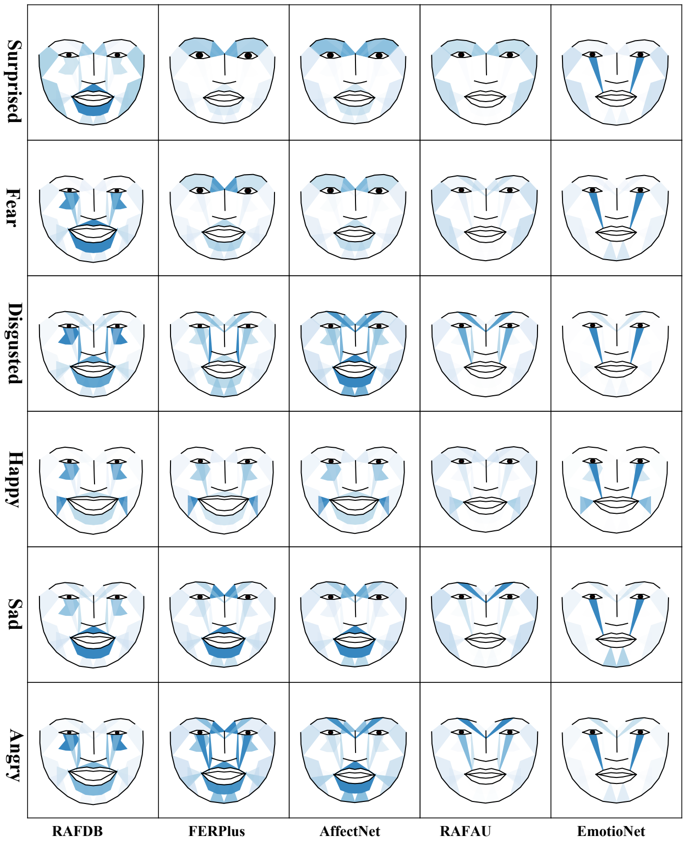

Fig. 1: The mean activated facial AUs of different emotions on popular FER datasets like RAFDB, FERPlus, AffectNet, RAFAU, and EmotioNet. The blue patches indicate AU regions and the color density refers to AU intensity.

Psychological studies show that Facial Action Unit (AU) is an objective and common standard to describe the physical expression of emotions [15]. For instance, $AU12$ and $AU6$ indicate lip corner puller and cheek raiser, which relate to happiness; $AU1$, $AU4$, and $AU6$ denote inner brow raiser, brow lowerer, and cheek raiser, which relate to sadness. Figure 1 shows the mean activated facial AUs (extracted by OpenFace [2] and drawn by Py-FEAT [1]) of different emotions

Shuyi Mao and Xinpeng Li are equally-contributted authors.
Corresponding Author: Xiaojiang Peng(pengxiaojiang@sztu.edu.cn)

[1] https://github.com/cosanlab/py-feat

on several popular datasets. On the one hand, we observe that the AUs of the same expressions are inconsistent in different datasets, indicating dataset bias indirectly. On the other hand, since AUs are semantically universal among all datasets and partially shared in the same expression, we believe that the FER task can benefit from AUs.

In this paper, we first investigate the FER bias issue comprehensively. Existing works mainly suggest the bias existence by cross-dataset FER performance [26], [8], [56], [34], i.e., training on a source dataset and testing on another target dataset. Here, we mainly focus on boosting individual dataset performance with cross-domain datasets. We analyze dataset bias in a fine-grained view by regarding FER as an objective image-to-AU plus a subjective AU-to-emotion procedure. The bias of the image-to-AU process is more likely to be collection bias, and the one of the AU-to-emotion procedure is mainly about annotation bias due to subjectivity. We focus on measuring the later procedure since the large-scale data and deep neural network are expected to eliminate the collection bias. Specifically, we use AU differences to estimate the annotation subjectivity qualitatively and quantitatively. In the quantitative aspect, we show that the performance degrades for a specific dataset by joint training with other FER datasets. We plot the expression-specific mean images and AU cosine distance among datasets in the qualitative aspect. The cosine distance is consistent with joint training performance degradation.

Second, we propose a conceptually-new yet simple network architecture, termed an AU-aware Vision Transformer (AU-ViT), to jointly train a target dataset and cross-domain datasets, which can effectively boost the performance of the target dataset. Since the image-to-AU procedure is objective, the AU-ViT mainly takes AUs or pseudo-AUs as a bridge between target and auxiliary datasets. The AU-ViT consists of three vital branches: a Base branch, a ViT-based Expression classification branch (ViT-Exp branch), and an AU branch. For the Base branch, two alternatives are designed, i.e., vanilla ViT, which conducts Transformers on pixel patches, and convolution-based ViT (CNN-ViT), which applies Transformers on feature maps. Besides, we empirically introduce advanced modules like multi-stage blocks. For the AU branch, we elaborately design several patch-splitting strategies for feature aggregation, motivated by the fact that AUs are mainly defined by local information. Plus, we introduce a symmetric maxout layer to address occlusion and profile issues. The AU branch is expected to boost the Base branch's feature learning and thus lead to better expression recognition.

Finally, we extensively evaluate our approach on three popular FER datasets, two AU datasets, and three occlusion FER test datasets. Our AU-ViT achieves state-of-the-art FER performance with 91.10% on RAF-DB, 65.59% on AffectNet, and 90.15% on FERPlus. The AU-ViT also outperforms current state-of-the-art methods on Occlusion-RAFDB, Occlusion-FERPlus, and Occlusion-FERPlus with clear margins. In addition, we find that a vanilla ViT pipeline with proper initialization (Base branch plus ViT-Exp branch) achieves comparable performance to advanced deep convolutional networks. With visualizations, we observe that our AU-ViT captures more meaningful semantic features than other methods. The code and pre-trained models will be released soon.

Our contributions can be summarized as follows:

1) We conduct extensive FER datasets joint training experiments. We find naive joint training seriously decreases the FER performance of a particular dataset. Then we measure FER dataset bias in a fine-grained way by computing expression-specific mean images and AU cosine distances. The measurement shows consistent conclusions with experimental degradation.
2) We propose a novel AU-aware Vision Transformer (AU-ViT) architecture to train a target dataset and cross-domain datasets jointly. The AU branch of AU-ViT resorts to AU information and effectively boosts the performance of the target dataset. We also introduce several patch-splitting schemes for the AU branch. Visualizations show that our AU-ViT captures more meaningful semantic features than other methods due to the usage of AUs.
3) To the best of our knowledge, This is the first work to repurpose a vanilla ViT pipeline for FER. We demonstrate that a carefully initialized vanilla ViT obtains comparable performance to advanced deep convolutional networks. Besides, the CNN embedding and advanced modules like multi-stage blocks further boost the performance.
4) We conduct extensive experiments on eight publicly available FER and AU databases to evaluate our AU-ViT and achieve state-of-the-art performance with 91.10% on RAF-DB, 65.59% on AffectNet, and 90.15% on FERPlus.

II. Related work

Numerous scholars in computer vision have devoted themselves to FER and contributed a lot of works, including algorithms [21], [51], [43], [11], [33], [38], [57], [25], [17], [59], [47], [48] and datasets [58], [29], [27], [3]. This paper proposes AU-ViT to address how to boost FER performance by leveraging cross-domain datasets. Therefore, we will show related works closely related to this issue and methods, including cross-domain FER, joint training FER, dataset bias in FER, vision transformers in FER, and AU assistance in FER.

Dataset bias in FER. Dataset bias exists among different FER datasets due to data composition [26], [8], [34] and label ambiguity (e.g. subjective annotation [56], [9] and compound expression [27]). Existing works have achieved a lot in eliminating dataset bias. Wang *et al.* [47] modify a pair of training labels with the self-attention mechanism. Chen *et al.* [7] leverage action unit recognition and facial landmarks detection to guide label distributions. She *et al.* [44] build a multi-branch framework to mine label distribution and uncertainty extent in label space. *Different from these works, we resort to the AU information of different FER datasets for bias measurement.*

Joint training and cross-domain in FER. Some works have explored joint training in FER to improve algorithm generalization. Pan *et al.* [34] collect web images for more diversified image-label pairs. Zeng *et al.* [56] assign each sample several labels and find the latent truth in joint training.

Recent methods regard skipping dataset bias as a cross-domain problem. Li *et al.* [26] propose to minimize an emotion-conditional maximum mean discrepancy to learn domain-invariant features. Chen *et al.* [8] adopt holistic and local features between datasets with an adversarial graph representation learning. Zeng *et al.* [55] utilize face recognition datasets to address FER's class imbalance with an additional adaptation network. *Different from these works, we address how to boost FER performance of a target dataset by leveraging cross-domain datasets.*

Vision Transformer in FER. Vision Transformers have been proved effective in image classification, detection, and segmentation [13], [6], [50], [20], [28]. Recently, some works introduced Transformers to the FER task. Ma *et al.* [30] combine two kinds of feature maps generated by two-branch CNNs, then feed the fused feature into Transformers to explore relationships between visual tokens. Xue *et al.* [52] leverage Transformers upon CNNs to learn rich relations among diverse local patches between two attention droppings. *Different from these works, we build an AU-aware branch for patches after Transformer, which guides the blocks of Transformer to learning finer facial expression features.*

AU assistance in FER. Previous methods utilize AU information for FER since AUs are related to emotions [14]. Deng *et al.* [12] present a distillation strategy to learn from incomplete labels to boost the performance of AU detection, expression classification, and valence-arousal estimation. Pu *et al.* [36] exploit the relationships between AU and expression and choose useful AU representations for FER. Chen *et al.* [9] leverage AUs to understand and mitigate annotation bias and adopt triplet loss to encourage embeddings with similar AUs to get close in feature space. *Different from these works, we leverage AU or pseudo-AU labels from any other facial datasets and embed them into Transformers, aiming to boost the accuracy of target datasets.*

III. Methodology

To bridge the dataset bias and boost the accuracy of the target dataset, we propose an AU-aware Vision Transformer (AU-ViT) for multiple datasets training in FER. We first show the FER dataset bias qualitatively and quantitatively and then detail the components of the AU-ViT.

A. Dataset bias

Extensive datasets usually contribute to better network training in the deep learning era. However, the performance of a specific dataset degrades when training with auxiliary FER datasets due to bias. The existing works measure FER bias by dataset recognition and cross-dataset generalization [56], [34], [8], [26]. Here, we analyze FER bias in a fine-grain view, which regards FER as an image-to-AU and AU-to-expression procedure. The image-to-AU process mainly contains expression-independent bias (e.g., illumination), while the AU-to-expression procedure includes expression-dependent discrimination. We focus on the latter bias since the former is expected to be eliminated by large-scale datasets or robust deep convolutional networks, which is less related to joint training degradation.

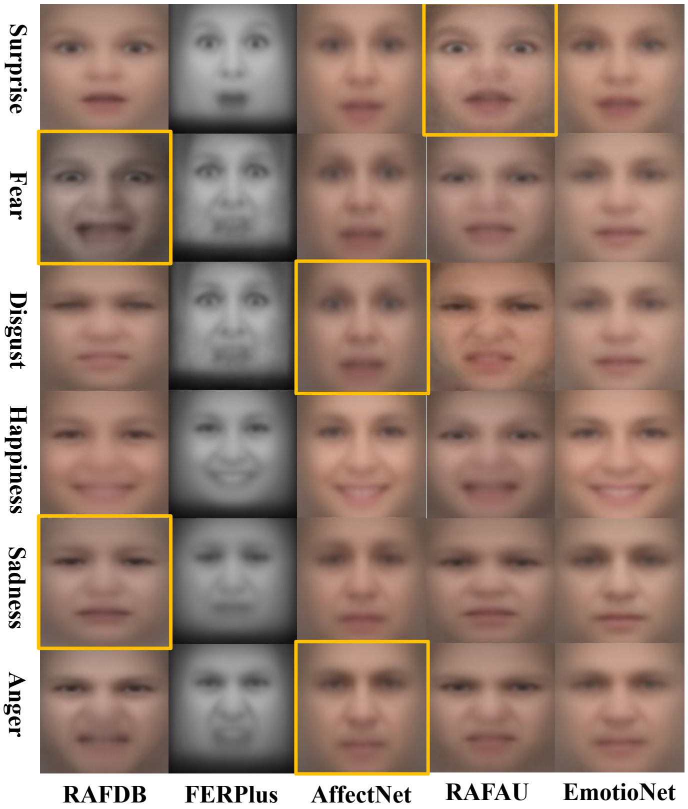

Fig. 2: Illustration of the AU-to-emotion Bias. We calculate the mean images of different emotions and datasets. The yellow box indicates the image with different facial motions.

Dataset mean images. We show the mean images of different expressions among different datasets in Figure 2. For comparison, we crop all images into the face image ratio of RAFDB and resize them into 112 x 112 size. Then, we calculate the mean images, which encode the statistical AUs of expressions in a specific dataset. We can see different relations between AUs and expressions among different datasets from the mean pictures, which indicate the dataset bias qualitatively.

As shown in Figure 2, the mean images of the same expressions vary more or less in different datasets. For example, the surprise average image of FERPlus shows more vigorous facial motion than one of the other datasets. The difference results from intra-class variance and annotation subjectivity. Worth mentioning that happiness shows more consistent facial movement in various datasets than other expressions.

Dataset cosine distance. To get a further accurate measurement for dataset bias, we plot the cosine distance of expression-specific facial AU representations between datasets in Figure 3. Specifically, we adopt AU labels from dataset annotations or OpenFace predictions. For comparison, we choose commonly available AUs: AU01, AU02, AU04, AU05, AU06, AU09, AU10, AU12, AU15, AU17, AU20, AU25, and AU26. Therefore, a 13-dimension mean AU vector represents the facial information of an expression in a dataset. Then, we compute their cosine distance, ranging from 0.0 to 1.0, to show dataset bias. It is worth mentioning that two vectors are regarded as similar enough when the space is less than 0.3

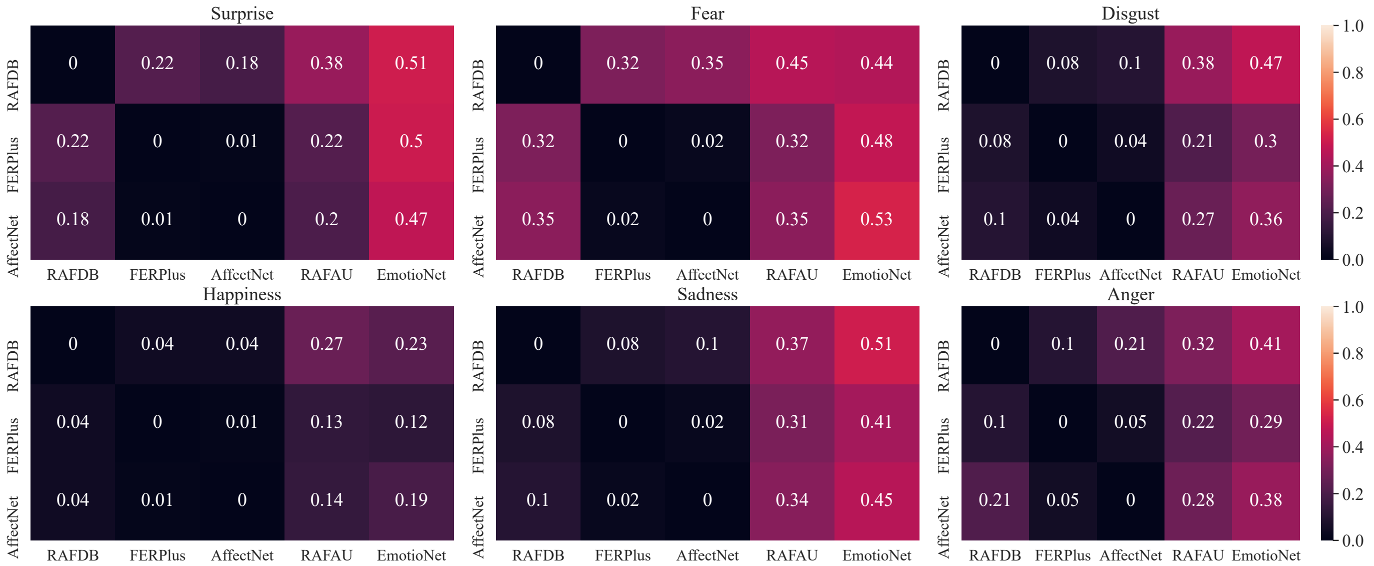

Fig. 3: Illustration of the quantitative AU-to-expression bias. We calculate the mean AU vectors of all emotions in different datasets and then plot their cosine distance. It is worth mentioning that two vectors are regarded as similar enough when the distance is less than 0.3 and different enough when more than 0.5.

Accuracy (%) Auxiliary Target	RAFDB	FERPlus	AffectNet	RAFAU	EmotioNet	None
RAFDB	/	84.84 (-2.46)	86.67 (-0.63)	82.72 (-4.58)	85.10 (-2.2)	**87.3**
FERPlus	86.54 (-1.37)	/	87.69 (-0.22)	87.19 (-0.87)	86.93 (-0.98)	**87.91**
AffectNet	58.19 (-1.86)	59.91 (-0.41)	/	59.24 (-0.81)	58.88 (-1.17)	**60.05**

TABLE I: The results of joint training on dataset pairs with FER labels. The results are reported on target datasets. The bracket presents the performance degradation compared to traditional training on individual datasets (i.e., the last column).

and different enough when more than 0.5.

From a view of general expressions, the EmotioNet is the most different dataset and contains around 0.5 distance from others. The distance between the RAFDB, FERPlus, and AffectNet is less than 0.3. From a view of a specific expression, fear is the most contrasting expression, and it includes more than 0.3 distance from other datasets. Happiness is the most consistent expression and holds less than 0.2 distance. These biases are constant with the following performance degradation of joint-training results.

Joint training exploration. We train different dataset pairs jointly to show how dataset bias affects the performance of FER. Specifically, we choose from three FER datasets: RAFDB, FERPlus, and AffectNet, and two AU datasets: RAFAU and EmotioNet. We use the deepface[2] [41], [42] to predict expressions and OpenFace [2] to predict AUs if labels are unavailable in the datasets. To keep the dominated role of target FER datasets, we set the mixing ratio of target and auxiliary images as 4 : 1. We pretrain the pure ViT on VGGFace2 in experiments and present the results in Table I.

From Table I, we can see that the results of joint training methods are consistently inferior to traditional training methods. For example, the accuracy of RAFDB drops by 2.46%, 0.63%, 4.58%, and 2.2% when combining FERPlus, AffectNet, RAFAU, and EmotioNet, respectively. Joint training with RAFAU results in the largest performance degradation for RAFDB, corresponding to the most significant quantitative difference in Figure 3. Besides, the performance of FERPlus decreases by 1.37%, 0.22%, 0.87%, and 0.98% when combining RAFDB, AffectNet, RAFAU, and EmotioNet, respectively. We also find that joint training FERPlus with AffectNet only suffers a slight degradation, consistent with the small quantitative difference in Figure 3. These results demonstrate that joint training with expression labels is not helpful in FER due to dataset bias and suggest the effectiveness of our quantitative expression bias scheme.

B. *AU-aware Vision Transformer*

From the above analysis, dataset bias is the critical problem for performance boosting in joint training. Existing methods mainly eliminate the dataset bias from an image-to-expression view, e.g., changing labels or regularizing embeddings. In this paper, regarding FER as an objective image-to-AU procedure

[2] https://github.com/serengil/deepface

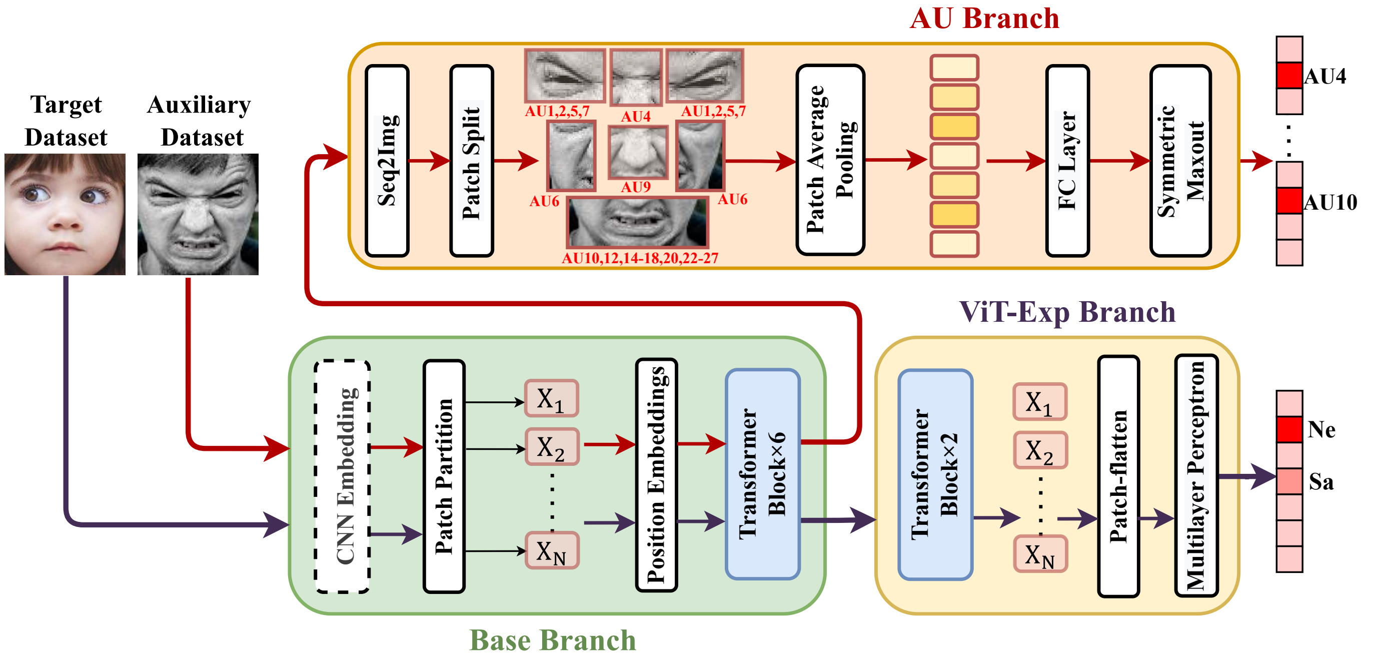

Fig. 4: Overview of our AU-ViT. It consists of three essential branches: Base branch, ViT-Exp branch, and AU branch. Specifically, the AU-ViT first takes as input the images from both target and auxiliary datasets, then encodes them into middle-level feature maps by the Base branch, and finally feeds the feature maps into either the ViT-Exp branch for facial expression classification or the AU branch for AU recognition. The CNN Embedding part is optional in the Base branch.

plus a biased AU-to-expression procedure, we resort to the AU objectivity yet avoid the expression label bias to boost the performance of individual target datasets. Specifically, we propose the AU-ViT model for jointly training a target FER dataset and auxiliary datasets with their AUs or pseudo-AUs.

AU-ViT Overview. As shown in Figure 4, the AU-ViT consists of three essential branches, namely the Base branch, ViT-Exp branch, and AU branch. The AU-ViT first takes the images from both target and auxiliary datasets as input. The Base branch then encodes images into middle-level feature maps and feeds them into either the ViT-Exp branch for facial expression classification or the AU branch for AU recognition. Specifically, we sample the images from target and auxiliary datasets with a 4:1 ratio. The ViT-Exp branch only handles the target dataset, and the AU branch only deals with the auxiliary datasets which own AU or pseudo-AU labels.

Base branch. The Base branch is designed into two alternatives, i.e., the vanilla ViT and CNN-ViT. Note that both ViTs of the Base branch exclude the classification part of a standard ViT pipeline. For the vanilla ViT, we split the image into N patches with 16×16 size and then reshape them into a 256-dimension embedding X. Then, we combine X and a learnable position embedding P through element-wise addition. For the CNN-ViT, we adopt feature maps from the third stage of IR50 [49] and conduct the above operations. The feature maps are split into non-overlapping patches with 1×1 size, and the dimension of each patch embedding is 256. Note that we concatenate a class token with X in the vanilla ViT while discarding it in the CNN-ViT. The above operations of can be formulated as follows,

$$Z = [X_1; X_2; ...; X_N] + P, \tag{1}$$

where Z is the sequence to be fed into the Transformers.

Existing works show the Multi-stage Transformer performs better than primitive ones in computer vision task [50], [20], [28]. To this end, we adopt it to build AU-ViT blocks experimentally. As shown in Figure 5, Multi-stage Transformers gradually increase the channel size while decreasing the spatial size. Moreover, the number of tokens is reduced by patch merging layers as the network gets deeper. At the end of each stage, the pooling layer concatenates the tokens of each 2×2 neighbor patch and doubles their dimensions. Therefore, the size of feature map in the end of the k-th stage of Transformer is $\frac{H_f}{2k} \times \frac{W_f}{2k} \times 2kC_f$, where H_f, W_f, and C_f are the height, width and dimension of the input feature map.

Inside the Multi-stage Transformer, a Multi-Head Self-Attention (MHSA) module processes the input embedding and models the complex interactions. Then, a Convolution in Feed-Forward Network (Conv-FFN) is employed to capture the connectivity of local regions. In the Conv-FFN module, we rearrange the sequence of tokens into a 2D lattice and convert it to 2D feature maps. The number of channels of the feature maps increases firstly, and then a depth-wise convolution with a kernel size of 3×3 performs on them. Finally, we restore the channels of feature maps and flatten them into sequences with the initial dimension. The local convolution enhances the representation correlation with neighboring eight tokens.

ViT-Exp branch. The ViT-exp branch handles the middle-level feature maps of a target dataset from the Base branch.

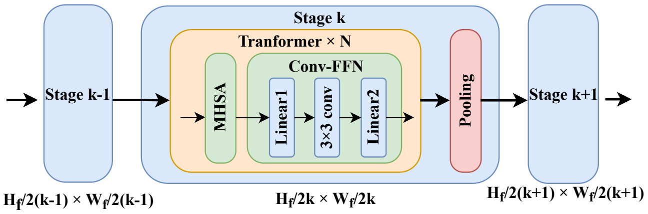

Fig. 5: Illustration of Multi-stage Transformers. The feature maps from the former stage are fed into MHSA, Conv-FFN, and a pooling layer. The size of feature map in the end of the k-th stage of Transformer is $\frac{H_f}{2k} \times \frac{W_f}{2k} \times 2kC_f$, where H_f, W_f, and C_f are the height, width and dimension.

For the vanilla ViT base branch, we implement a Multi-Layer Perception on the CLS Token to obtain the final classification. For the CNN-ViT base branch, two more Multi-stage Transformer blocks are further used upon the feature maps. Then, we flatten all the patch embeddings in the Patch-flatten layer and implement a Multi-Layer Perception for expression classification. Compared to pooling, the flattened operation and fully connected layer can retain all spacial information and the strong relationships among the features.

AU branch. The AU branch processes the middle-level feature maps of the auxiliary dataset from the Base branch. A Seq2Img operation [28] is first used to transform the patch tokens into 2D feature maps. Then, we crop the 2D feature maps into several overlapped patches according to AUs.

The top part of Figure 4 plots our AU-patch splitting strategy. We split the feature maps into seven patches: the left eye, right eye, left cheek, right cheek, between-eyebrow, nose, and mouth. The left and right-eye patches are responsible for AU1, AU2, AU5, and AU7. Likewise, the left and right-cheek patches are both responsible for AU6. The between-eye and nose patches are in charge of AU4 and AU9, respectively. In particular, the mouth patch is the largest part among these patches, which is related to 14 kinds of AUs: AU10, AU12, AU14, AU15, AU16, AU17, AU18, AU20, AU22, AU23, AU24, AU25, AU26, AU27. After splitting the feature maps, we conduct average pooling in each patch (Patch Average Pooling in Figure 4) to get patch representations and feed them into the fully-connected layers for AU recognition.

It is worth noting that the left and right patches, e.g., eye and cheek, are responsible for the same AUs. However, the occlusion and profile may lead to an asymmetric face and thus inconsistent AU prediction. To address this issue, we introduce a *symmetric maxout* layer to select the maximum of the left and right prediction. The formulation is as follows,

$$z = max(z_{left}, z_{right}), \tag{2}$$

where z_{left} and z_{right} denote the AU predictions of the left and right face, and z is the final AU prediction.

Loss Function. We adopt the cross-entropy loss (L_{ce}) for FER and binary cross-entropy loss (L_{bce}) for AU recognition. Let z_{fer} and z_{au} be the final output of the ViT-Exp branch and the AU branch, y_{fer} and y_{au} be the ground truth labels of the FER dataset and AU dataset, respectively. The former is fed into the softMax function while the latter inputs the Sigmoid function. The joint loss of AU-ViT is as follows:

$$\begin{aligned} L = \alpha L_{\text{ce}}(softmax(z_{fer}), y_{fer}) \\ + \beta L_{\text{bce}}(sigmoid(z_{au}), y_{au}), \end{aligned} \tag{3}$$

where α and β are hyper-parameters to balance two tasks.

IV. Experiments

In this section, we first describe the implementation details and datasets. Then we show quantitative and qualitative experimental results of eight datasets, which demonstrate the superiority of the AU-ViT in joint training. Finally, we conduct ablation studies of advanced modules and compare the AU-ViT with state-of-art methods.

A. Implementation details

The vanilla ViT is pre-trained on VGGFace2 [5] with an image size of 224×224. The CNN Embedding module is initialized by IR50 [49] pre-trained on MS1M [19]. We generate pseudo-AU labels for the FER datasets by the AU detector OpenFace [2]. All the images are aligned and resized to 112×112. For data augmentation, we use Random-Horizontal Flip, Random-Grayscale, and Gaussian Blur for all datasets. We use Adam optimizer for RAFDB and FERPlus, and SGD optimizer for AffectNet. The weight decay is 5e-4, and the batch size is 64 for RAFDB and 128 for FERPlus and AffectNet. The learning rate is initialized as 0.0001, and we use a linear learning rate warm-up of 10 epochs and a cosine learning rate decay of 10 epochs. All the experiments are implemented by Pytorch with 4 NVIDIA V100 GPUs. The default α and β in Equation (3) are 1.0 and 1.0, respectively.

B. Datasets

To evaluate our method, we use three popular in-the-wild facial expression datasets, namely RAF-BD [27], FERPlus [3] and AffectNet [32], and two popular in-the-wild facial AU datasets, namely RAFAU [53] and EmotioNet [16]. Besides, we also use the occlusion test datasets [48] from RAFDB, FERPlus, and AffectNet.

RAFDB [27]. It consists of 30,000 facial images downloaded from the Internet and annotated with basic or compound expressions by 40 trained human students. It is a large-scale and real-world dataset. In our experiment, we only use images with seven basic expressions (neutral, happiness, surprise, sadness, anger, disgust, fear, neutral), including 12,271 images for training and 3,068 for testing. We mainly report the overall accuracy of the test set.

FERPlus [3]. The FERPlus contains 28,709 training images, 3,589 validation images, and 3,589 test images collected by the Google search engine. All of the pictures of FERPlus are aligned and resized to 48×48, each of which is annotated by ten annotators. We re-align the faces in the aligned set of FERPlus to increase the area ratio of faces in an image. Apart from seven basic emotions in FER2013, contempt is included in FERPlus. We report the overall accuracy of the test set.

Accuracy (%) Auxiliary Target	RAFDB	FERPlus	AffectNet	RAFAU	EmotioNet
RAFDB	/	88.41 (+3.57)	87.84 (+1.17)	**88.8** (+6.08)	87.9 (+2.8)
FERPlus	88.4 (+1.86)	/	88.34 (+0.56)	88.44 (+1.25)	**88.53** (+1.6)
AffectNet	61.57 (+3.38)	62.03 (+2.12)	/	**62.33** (+3.09)	61.17 (+2.83)

TABLE II: The results of joint training on dataset pairs with our basic AU-ViT. The numbers in brackets present the improvements of our basic AU-ViT compared to the results of ViT in Table I.

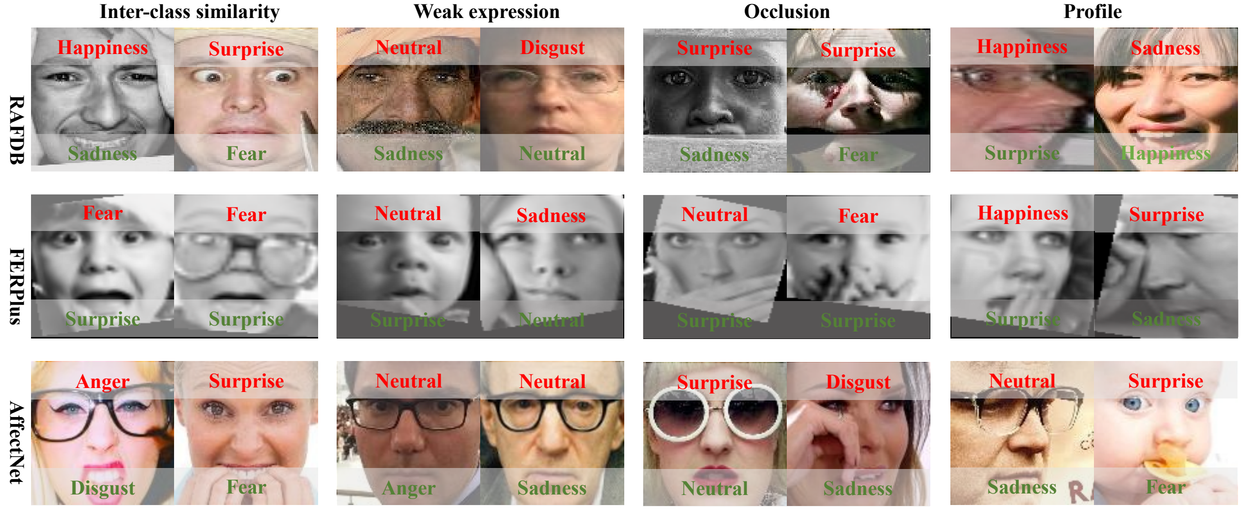

Fig. 6: The illustration of predicted samples by vanilla ViT trained on individual datasets and our basic AU-ViT jointly-trained with RAFAU. The red texts indicate incorrect predictions by vanilla ViT while the green ones indicate correct predictions by the AU-ViT. We can see the AU-ViT performs well in inter-class similarity, weak expression, occlusion, and profile situations.

AffectNet [32]. The AffectNet is the largest dataset that provides both categorical and Valence-Arousal annotations. It includes over one million images from the Internet by querying expression-related keywords in three search engines. In AffectNet, about 420,000 images are manually annotated with eight basic expression labels like RAFDB and 3500 validation images. Note that AffectNet has an imbalanced training set, a test set, and a balanced validation set. So we use an oversampling strategy to produce balanced batch samples in the training phase. The overall accuracy of the validation set is used for measurement.

RAFAU [53]. The RAFAU is an extended dataset of RAF-ML collected from the Internet with blended emotions. It varies in subjects' identity, head poses, lighting conditions, and occlusions. The face images in RAFAU are FACS-coded by two experienced coders independently. It contains 4601 real-world images with 26 kinds of AUs: AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU16, AU17, AU18, AU20, AU22, AU23, AU24, AU25, AU26, AU27, AU28, AU29, AU35, AU43. In our experiments, we only use the first 21 AUs, which strongly relate to expressions.

EmotioNet [16]. The EmotioNet database includes 950,000 images with annotated AUs. The images are downloaded from the Internet by searching with related facial expression keywords. A specific algorithm automatically annotates these images with AUs, AU intensities, and expression categories. It contains 23 AUs, including AU1, AU2, AU4, AU5, AU6, AU9, AU10, AU12, AU15, AU17, AU18, AU20, AU24, AU25, AU26, AU28, AU43, AU51, AU52, AU53, AU54, AU55, AU56. In our experiments, we only use the first 16 AUs, which strongly relate to facial expressions.

Occlusion FER datasets [48]. The total numbers of occlusion samples in FERPlus (test set), AffectNet (validation set), and RAF-DB (test set) are respectively 605, 682, and 735, which are 16.86%, 17.05%, and 23.9% of their original sets. These real-world test sets are annotated manually with different occlusion types, which include wearing a mask, wearing glasses, objects on the left/right, objects in the upper face and objects in the bottom face, and non-occlusion. Images in the occlusion test sets have at least one type of occlusion. We report the overall accuracy of the three occlusion datasets.

C. Joint training with AU-ViT

We conduct extensive joint training experiments among three FER datasets: RAFDB, FERPlus, and AffectNet, and two AU datasets: RAFAU and EmotioNet. In particular, we predict pseudo-AUs for FER datasets via OpenFace[2]. The expected AUs include AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25,

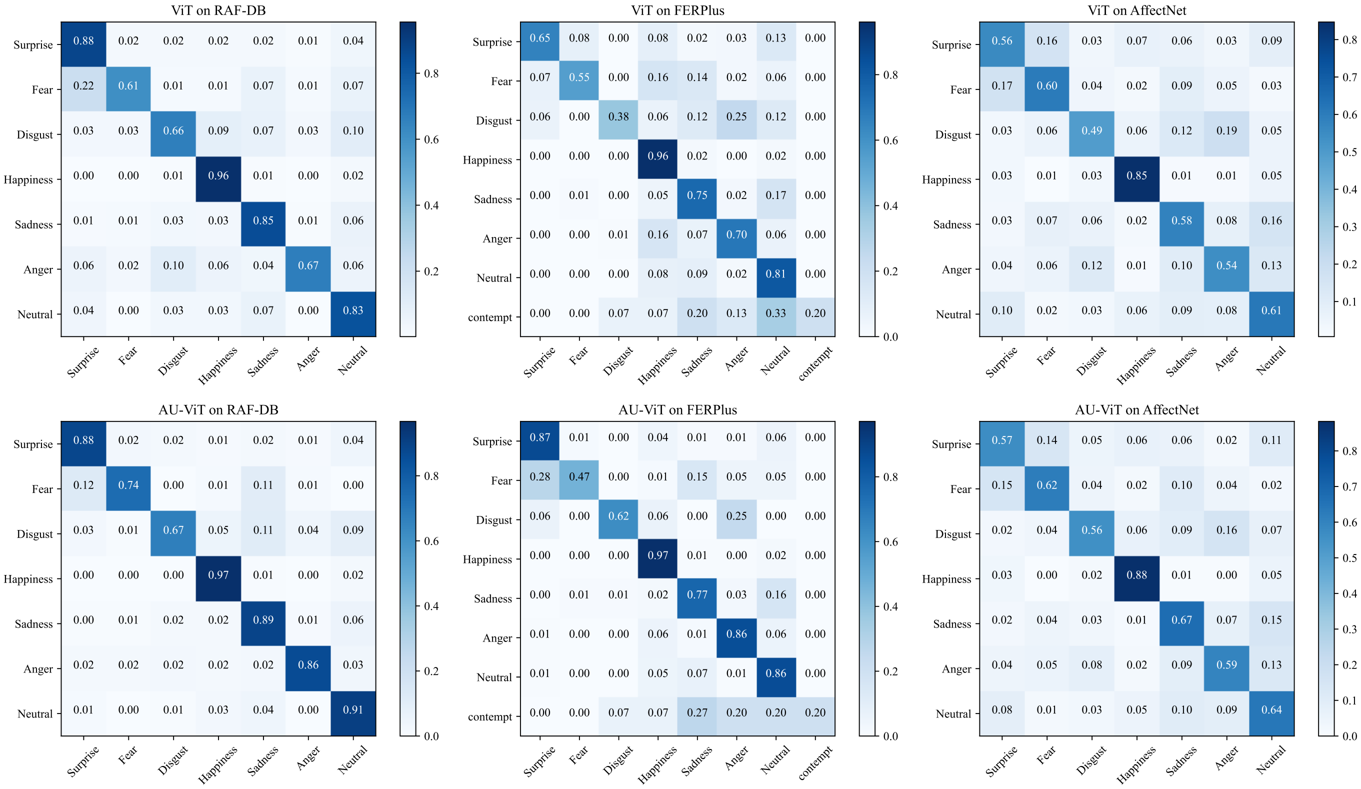

Fig. 7: The confusion matrices of the vanilla ViT and our AU-ViT on RAF-DB, FERPlus, and AffectNet.

and AU26, which are all used for training. Besides, we select available expression-related AUs if auxiliary AU datasets are used. The specific AUs are mentioned in the datasets section. In the experiments, similar to Table I, the base branch adopts the vanilla ViT and is pre-trained on the VGGFace2 [5]. The results are shown in Table II.

Compared to joint training with FER labels (i.e., Table I), our AU-ViT significantly enhances the performance, e.g., we achieve a 6.08% gain for RAFDB when combining RAFDB and RAFAU. These results demonstrate the effectiveness of our AU-ViT. We also find two other observations as follows. First, when combining AU datasets, the gain from RAFAU is generally more significant than that from EmotioNet, which is consistent with the AU distance in Figure 3. Second, the auxiliary AU datasets usually bring more improvements, which may be explained that the manual AU annotations have better quality than the pseudo ones.

To investigate the superiority of our AU-ViT, we show the samples correctly recognized by AU-ViT but not by ViT in Figure 6. From these samples, we find the AU-ViT performs better in four aspects of examples: inter-class similarity, weak expression, occlusion, and profile. Different expressions sometimes share similar facial motions, e.g., a sad face may raise the mouth corner and show the teeth as in a happy face. The ViT trained with FER labels is confused by these samples, while the AU-ViT is not. It may be explained that the AU branch of AU-ViT makes the model pay attention to all the local regions related to action units. Some expressions (e.g., sadness, surprise, and anger) have weak intensity and look similar to neutral. Occlusion and profile faces make it hard for ViT since critical clues, like front faces, are missing. However, our AU-ViT occasionally captures dominated features for expression recognition in non-occlusion facial regions thanks to the AU recognition branch. We further compare ViT and our basic AU-ViT by illustrating confusion matrices in Figure 7. Our AU-ViT achieves remarkable improvements for the categories ‘Fear’ and ‘Anger’ on RAF-DB, ‘Disgust’ and ‘Anger’ on FERPlus, ‘Disgust’ and ‘Sadness’ on AffectNet.

TABLE III: Evaluation of advanced modules in AU-ViT. The results are reported on RAFDB, and the AU-ViT models are jointly trained on RAFDB and RAFAU.

CNN Embedding	Multi-stage	Conv-FFN	Patch-flatten	AU Branch	Accuracy
					87.30
				√	88.80
√					89.86
√	√				90.12
√	√	√			90.22
√	√	√	√		90.42
√	√	√	√	√	**91.10**

D. Advanced modules for AU-ViT

Though our basic AU-ViT has already obtained comparable performance to state-of-the-art methods, recently advanced

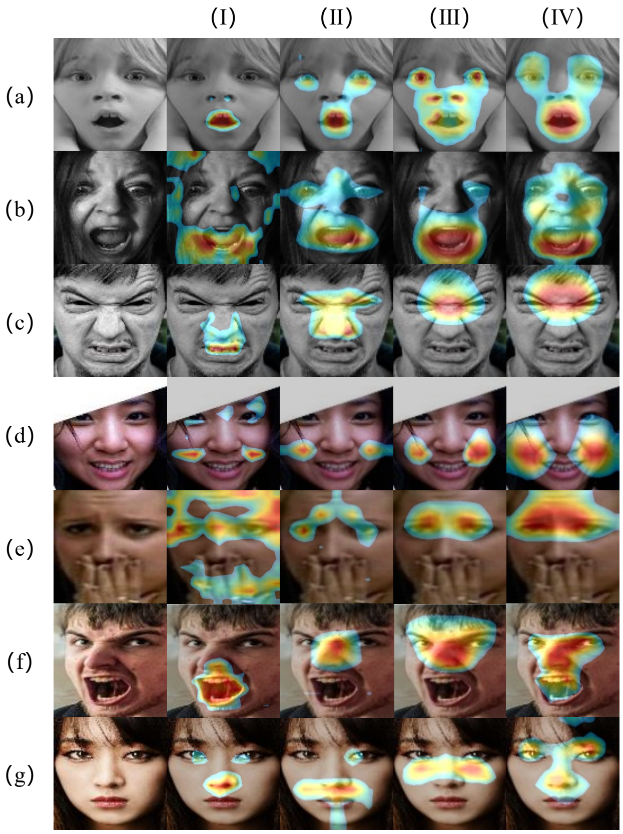

Fig. 8: Different attention feature maps from (I) the vanilla ViT, (II) ViT with CNN Embedding, (III) ViT with CNN Embedding and Multi-stage blocks, and (IV) our final version of AU-ViT. The rows (a)-(g) show seven facial images with different facial expressions from the RAFDB.

modules for Transformers can further be embedded into our AU-ViT. Thus, an evaluation is designed to investigate the effects of CNN Embedding block, Multi-stage block, Conv-FFN, and Patch-flatten operations. The results are presented in Table III. For the CNN Embedding block in our base branch, we resort to the IR50 pre-trained on the MS1M dataset [49] and use the output feature maps from the third stage as input of Transformer blocks. Adding the CNN Embedding block brings the most significant gain, i.e., 2.56%, which can be due to the large-scale pre-trained face dataset or the CNN architecture itself. The Multi-stage Transformer block can capture features from more significant regions than standard ViT blocks, which improves the result with CNN Embedding by 0.56%. We flatten all patch embeddings to keep more information from the last fully-connected layer instead of simply projecting the class token. This simple change also brings 0.2% performance gain. Note that our critical AU branch consistently obtains visible improvements whether advanced modules are used or not. Our final version of AU-ViT achieves 91.10% on RAFDB.

TABLE IV: Comparison to the state-of-the-art results on the FERPlus dataset.

Type	Method	Year	Public	Performance
CNN	PLD[4]	2016	ICMI	85.10
	SCN[47]	2020	CVPR	88.01
	RAN[48]	2020	CVPR	88.55
	DMUE[44]	2021	CVPR	89.51
Transformer	VTFF[30]	2021	TAC	88.81
	MVT[23]	2021	/	89.22
	Ours	2022	/	**90.15**

Visualization. To investigate the effects of each module, we employ ScoreCAM [45] to visualize the attention feature maps from different models in Figure 8. From (a) to (g), the categories are surprise, fear, disgust, happiness, sadness, anger, and neutral. The first column shows the aligned input facial images. The second to fifth columns show the results of four different models, namely (I) the vanilla ViT, (II) ViT with CNN Embedding, (III) ViT with CNN Embedding and Multi-stage blocks, and (IV) our final version of AU-ViT.

Comparing different columns, we find that adding CNN Embedding can eliminate unrelated attention regions on the face(e.g., (I), (II) in (b) and (d)). Multi-stage Transformer guides to learn proper and correct features (e.g. (II), (III) in (c)). The AU branch shrinks the attention maps and focuses on more key areas. Take column (IV) in (b) as an example. The model with an AU branch pays attention to the mouth, brow, and eyes, while the model without an AU branch only focuses on the mouth and brow. This visualization can further explain why our AU-ViT outperforms vanilla ViT in the situations of weak expression, occlusion, profile, etc.

E. Comparison with the state-of-the-art methods

We compare our final version of AU-ViT to the state-of-the-art methods on RAFDB, FERPlus, and AffectNet.

Comparison on FERPlus. Table IV shows the comparison between our AU-ViT and state-of-the-art methods on FERPlus. Among them, PLD [4], SCN [47] and RAN [48] are CNN-based methods. PLD uses VGG13 as the backbone, which is very simple and light. SCN tends to solve the uncertain problem in the FER task. RAN proposes a region-biased loss to encourage high attention weights on the most critical regions. VTFF [30] and MVT [23] are Transformer-based methods. VTFF also utilizes CNN embeddings and a multi-layer transformer encoder to recognize expressions. MVT uses two Transformer modules to filter out useless patches and classify the remaining features. Our elaborate AU-ViT outperforms these methods and achieves 90.15% accuracy.

Comparison on AffectNet. Table V compares our method to several state-of-the-arts methods on AffectNet. DDA-Loss [17] optimizes the embedding space for extreme class imbalance scenarios. EfficientFace [59] focuses on label distribution learning. KTN [22] designs a distillation strategy to classify highly similar representations. DACL [18] focuses on the most discriminative facial regions and features.

TABLE V: Comparison to the state-of-the-art results on the AffectNet-7 dataset.

Type	Method	Year	Public	Performance
CNN	IPA2LT [56]	2018	ECCV	57.31
	LDL-ALSG [7]	2020	CVPR	59.35
	DDA-Loss [17]	2020	CVPR	62.34
	DMUE [44]	2021	CVPR	63.11
	EfficientFace [59]	2021	AAAI	63.7
	KTN [22]	2021	TIP	63.97
	Meta-Face2Exp [55]	2022	CVPR	64.23
	DACL [18]	2021	WACV	65.20
Transformer	MVT [23]	2021	/	64.57
	VTFF [30]	2021	TAC	64.80
	Ours	2022	/	**65.59**

TABLE VI: Performance comparison with the state-of-the-art methods on RAF-DB.

Type	Method	Year	Public	Performance
CNN	DLP-CNN [24]	2016	CVPR	84.13
	LDL-ALSG [7]	2020	CVPR	85.53
	IPA2LT [56]	2018	ECCV	86.77
	RAN [48]	2020	CVPR	86.90
	SPDNet [1]	2018	CVPRW	87.00
	DDL [39]	2020	CVPR	87.71
	SCN [47]	2020	CVPR	88.14
	Meta-Face2Exp [55]	2022	CVPR	88.54
	DMUE [44]	2021	CVPR	89.42
	FDRL [40]	2021	CVPR	89.72
Transformer	VTFF [30]	2021	TAC	88.81
	MVT [23]	2021	/	89.22
	TransFER [52]	2021	ICCV	90.91
	Ours	2022	/	**91.10**

Due to the gap between the imbalanced training set and the balanced validation set, we also use the traditional oversampling strategy and increase the proportion of auxiliary data in each iteration from 4:1 to 3:1. We finally obtain 65.59% on AffectNet, which exceeds the VTFF by 0.79%.

Comparison on RAFDB. Table VI compares our approach with previous state-of-the-art methods on RAFDB. Among all the competing methods, DDL [39] and RAN pay attention to disentangling the disturbing factors in facial expression images. SPDNet [1] introduces a new network architecture, and DLP-CNN [24] uses a novel loss function to explore intra-class variations. IPA2LT [56] and SCN [47] try to reduce the effects of noise labels. TransFER [52] is the first Transformer-based model for the FER task. It is worth noting that IPA2LT also resorts to multiple facial datasets for performance boosting via multiple predicted pseudo-FER labels from varied models. All the above methods improve FER performance by focusing on facial expression images, model architectures, and FER labels, neglecting the importance of AUs. Our AU-ViT finally achieves 91.10% on RAFDB, which surpasses TransFER slightly, establishing a new state of the art.

TABLE VII: Comparison between our AU-ViT and other methods with occlusion conditions.

Test datasets	Model	Acc.
Occlusion-RAFDB	ViT(Baseline)	81.63
	RAN [48]	82.72
	ASF-CVT [31]	83.95
	AU-ViT	**88.02**
Occlusion-FERPlus	ViT(Baseline)	81.48
	RAN [48]	83.63
	ASF-CVT [31]	**84.79**
	AU-ViT	**84.79**
Occlusion-AffectNet	ViT(Baseline)	57.26
	RAN [48]	58.50
	ASF-CVT [31]	62.98
	AU-ViT	**63.92**

Comparison on occlusion datasets As shown in Figure 6, our AU-ViT works better than ViT in occlusion and profile faces, thanks to the guidance of the AU branch. To further verify this, we evaluate our basic AU-ViT (without advanced modules) on the occlusion test sets of RAFDB, FERPlus, and AffectNet. All the models are pre-trained on VGGFace2 [5] and fine-tuned on the corresponding training set. The evaluation results are shown in Table VII.

Two observations can be concluded as follows. First, our AU-ViT improves the baseline ViT on all datasets, with gains of 6.39%, 3.31%, and 6.66% on Occlusion-RAFDB, Occlusion-FERPlus, and Occlusion-AffectNet, respectively. Second, though ASF-CVT utilizes Transformers on CNN feature maps, our basic AU-ViT outperforms it on Occlusion-RAFDB and Occlusion-AffectNet with large margins. Worth mentioning that the symmetric maxout layer in the AU branch can effectively capture AU activation even if half of the face is occluded. We experimentally find that the symmetric maxout

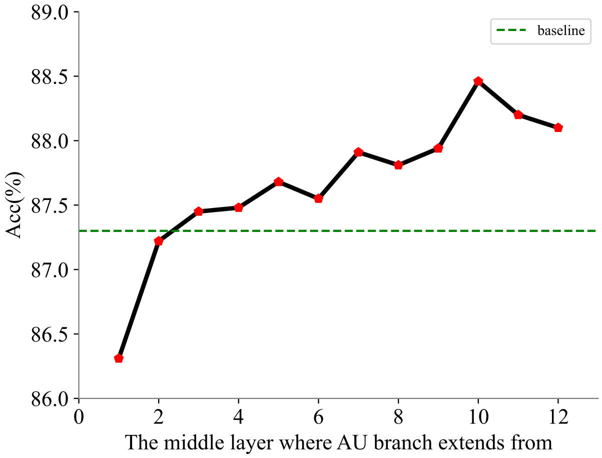

Fig. 9: Evaluation of the AU branch position in our basic AU-ViT.

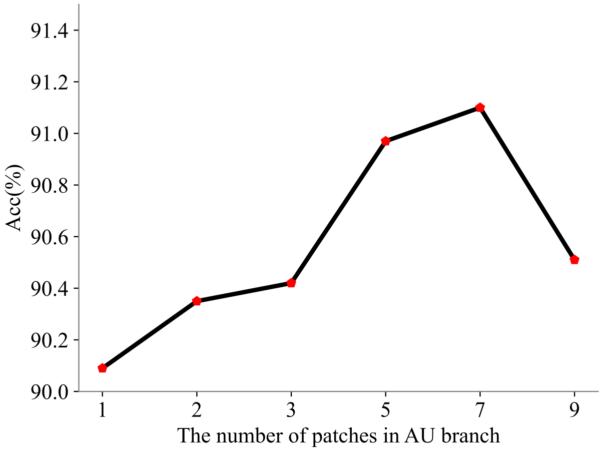

Fig. 10: Evaluation of the splitting strategy in AU branch of our final AU-ViT.

layer can consistently bring around 0.5% performance gain.

F. Ablation study

Adding AU information in FER training is previously used for cross-domain facial expression recognition [46] or compound expression recognition [36]. Indeed, it is not trivial to boost performance on individual FER datasets with AUs. Here, we make ablation experiments to show our efforts on this issue.

The usage of AUs. With AUs, the simple idea is to improve a FER model with a multi-task training strategy. Following the idea, we add an AU Token in the baseline ViT and conduct multi-task training on RAFDB with other auxiliary datasets. Specifically, we concatenate the AU token with patch embeddings and the class Token after the linear projection. We implement a multi-layer perception on the AU Token to estimate AUs, and the others are the same as baseline ViT. Table VIII compares these two usage of AUs. It is worth noting that we can also use the pseudo-AUs of RAFDB for training. From Table VIII, several observations can be concluded as follows. First, compared to the baseline, the 'AU token' multi-task training method can also slightly improve performance except for training with EmotioNet. Second, our AU branch consistently improves the baseline with visible margins. Third, our AU branch not only boosts performance with all auxiliary datasets but also outperforms the 'AU token' strategy. Thus, we believe that our elaborately-designed AU branch with patch splitting and symmetric maxout can be more precise than just adding an AU token for AU estimation.

The position of AU branch. In our AU-ViT model, the base and Exp-ViT branches own 12 Transformer blocks in succession, and our AU branch is extended from the default 10-th block. To investigate the effect of AU branch position, we evaluate RAFDB with auxiliary RAFAU by setting the position from 1 to 12 and show the results in Figure 9. Adding the AU branch in early blocks can be harmful, which may be explained by the fact that the early blocks do not have enough semantic information for AU estimation, thus disturbing the low-level feature learning for FER. As the AU branch position goes deep, the performance is gradually improved and saturated at the default position. Extending from the last two blocks may slightly damage the semantic features of FER, leading to performance degradation.

The patch splitting strategy. Since AUs are defined as the movement of muscles in the face, they are naturally related to local regions. We believe that splitting a face image into different patches helps recognize AUs. However, which splitting scheme is the best for the final performance is unknown. To this end, we design five strategies for our default AU branch, as shown in Figure 11. The splitting schemes generate 2, 3, 5, 7, and 9 regions for each face image. We do not split the bottom half face for most schemes since 14 AUs are related to the lower part of the face, and it isn't easy to separate individual AUs. All the patches are slightly overlapped for the necessity of complete AU-related features.

We evaluate with our final version of AU-ViT and present the results in Figure 10. Without splitting (i.e., one patch), our AU-ViT obtains 90.09%. It slightly improves performance by dividing it into 2 or 3 patches in a row. Further separating the upper face in row and column dramatically boost the accuracy. We get the best performance by the default 7-patch splitting strategy. We observe a visible performance degradation when further splitting the bottom face. All in all, breaking the feature maps into multiple patches keeps spatial information and benefits AU estimation and the final FER performance.

V. Conclusion

In this paper, we first introduce the expression-special mean images and mean activated AU images to show the FER dataset bias, which are consistent with the quantitative results of joint training experiments. Meanwhile, we present an AU-aware ViT model for facial expression recognition in the wild. It mainly resorts to the AU information, and effectively boosts the performance of individual FER datasets by leveraging other biased datasets. We carefully design the AU-ViT model with several advanced modules and observe that our model

TABLE VIII: Performance comparison between the AU Token and the AU branch on RAFDB.

Accuracy (%) Auxiliary Target		RAFDB	FERPlus	AffectNet	RAFAU	EmotioNet	Baseline
RAFDB	AU token	87.35	87.71	87.65	87.97	87.03	87.3
	AU branch	87.81	88.41	87.84	88.8	87.9	

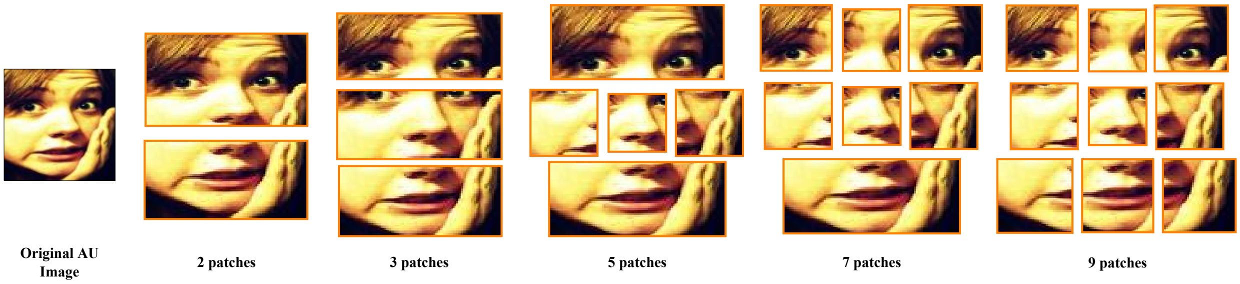

Fig. 11: Evaluation of different patch-splitting strategies.

can capture more local region features than traditional ViT. We finally achieve state-of-the-art performance in most of the popular FER datasets including occlusion datasets.

References

[1] D. Acharya, Z. Huang, D. P. Paudel, and L. V. Gool. Covariance pooling for facial expression recognition. In *2018 CVPRW*, 2018. 10

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 1, 4, 6, 7

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016. 2, 6

[4] Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. *ACMMM*, 2016. 9

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6, 8, 10

[6] C. F. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. 2021. 3

[7] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *CVPR*, pages 13984–13993, 2020. 1, 2, 10

[8] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, Lingbo Liu, and Liang Lin. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *TPAMI*, 2021. 1, 2, 3

[9] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *ICCV*, pages 14980–14991, 2021. 1, 2, 3

[10] Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, and Thibaut Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008. 1

[11] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *TPAMI*, 38(8):1548–1568, 2016. 1, 2

[12] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020. 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 3

[15] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 1

[16] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016. 6, 7

[17] Amir Hossein Farzaneh and Xiaojun Qi. Discriminant distribution-agnostic loss for facial expression recognition in the wild. In *CVPRW*, June 2020. 2, 9, 10

[18] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2402–2411, 2021. 9, 10

[19] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic imaging*, 2016(11):1–6, 2016. 6

[20] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *ICCV*, pages 11936–11945, 2021. 3, 5

[21] Pvc Hough. Method and means for recognizing complex patterns. *U.s.patent*, 1962. 2

[22] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao. Adaptively learning facial expression representation via c-f labels and distillation. *TIP*, PP(99):1–1, 2021. 9, 10

[23] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. Mvt: Mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021. 9, 10

[24] S. Li, W. Deng, and J. P. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 CVPR*, 2017. 10

[25] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020. 1, 2

[26] Shan Li and Weihong Deng. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, 2020. 1, 2, 3

[27] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2584–2593. IEEE, 2017. 1, 2, 6

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision

transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3, 5, 6

[29] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 2010. 2

[30] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021. 3, 9, 10

[31] Fuyan Ma, Bin Sun, and Shutao Li. Robust facial expression recognition with convolutional visual transformers. *arXiv preprint arXiv:2103.16854*, 2021. 10

[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, 2017. 6, 7

[33] Fatemeh Noroozi, Dorota Kaminska, Ciprian Corneanu, Tomasz Sapinski, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 2018. 1, 2

[34] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *ECCV*, pages 579–595, 2018. 1, 2, 3

[35] Giovanni Pioggia, Roberta Igliozzi, Marcello Ferro, Arti Ahluwalia, Filippo Muratori, and Danilo De Rossi. An android for enhancing social skills and emotion recognition in people with autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(4):507–515, 2005. 1

[36] Tao Pu, Tianshui Chen, Yuan Xie, Hefeng Wu, and Liang Lin. Au-expression knowledge constrained representation learning for facial expression recognition. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11154–11161. IEEE, 2021. 3, 11

[37] Fuji Ren and Changqin Quan. Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. *Information Technology and Management*, 13(4):321–332, 2012. 1

[38] Philipp V Rouast, Marc Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 2019. 1, 2

[39] D. Ruan, Y. Yan, S. Chen, J. H. Xue, and H. Wang. Deep disturbance-disentangled learning for facial expression recognition. In *MM '20: ACMMM*, 2020. 10

[40] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *CVPR*, pages 7660–7669, 2021. 10

[41] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020. 4

[42] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 4

[43] Caifeng Shan, Shaogang Gong, and Peter W. Mcowan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. 2

[44] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *CVPR*, pages 6248–6257, 2021. 1, 2, 9, 10

[45] H. Wang, Z. Wang, M. Du, F. Yang, and X. Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPRW*, 2020. 9

[46] Kai Wang, Yuxin Gu, Xiaojiang Peng, Panpan Zhang, Baigui Sun, and Hao Li. Au-guided unsupervised domain adaptive facial expression recognition. *arXiv preprint arXiv:2012.10078*, 2020. 11

[47] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, pages 6897–6906, 2020. 1, 2, 9, 10

[48] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *TIP*, 29:4057–4069, 2020. 2, 6, 7, 9, 10

[49] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. Face.evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621*, 2021. 5, 6, 9

[50] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021. 3, 5

[51] J. Whitehill and C. W. Omlin. Haar features for facs au recognition. In *International Conference on Automatic Face and Gesture Recognition*, 2006. 2

[52] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *ICCV*, pages 3601–3610, 2021. 3, 10

[53] Wenjing Yan, Shan Li, Chengtao Que, JiQuan Pei, and Weihong Deng. Raf-au database: In-the-wild facial expressions with subjective emotion judgement and objective au annotations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 6, 7

[54] Dongri Yang, Abeer Alsadoon, PW Chandana Prasad, Ashutosh Kumar Singh, and Amr Elchouemi. An emotion recognition model based on facial recognition in virtual learning environment. *Procedia Computer Science*, 125:2–10, 2018. 1

[55] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *CVPR*, pages 20291–20300, 2022. 1, 3, 10

[56] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *ECCV*, pages 222–237, 2018. 1, 2, 3, 10

[57] Ligang Zhang, Brijesh Verma, Dian Tjondronegoro, and Vinod Chandran. Facial expression analysis under partial occlusion: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–49, 2018. 1, 2

[58] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. Best of Automatic Face and Gesture Recognition 2013. 2

[59] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3510–3519, 2021. 2, 9, 10