PIDViT: Pose-Invariant Distilled Vision Transformer for Facial Expression Recognition in the Wild

Yin-Fu Huang and Chia-Hsin Tsai

Abstract—Many Facial expression recognition methods have achieved great success, but they only considered front facial images or facial images close to the front. Besides, unlike in-the-laboratory datasets, the facial images in the real world (or in the wild) are without lighting and pose control, so that it is a big challenge to recognize these facial expressions. In this paper, the PIDViT (i.e., Pose-Invariant Distilled Vision Transformer) using the teacher-student model for the probability distributions of facial expressions of frontal and multi-pose faces was proposed and solved the pose variance and occlusion issues on expression recognition. First, the multi-pose face dataset FairFace-3D was built from the original FairFace and then used to train pose-invariance on the PIDViT. The PIDViT was trained in two stages; stage 1 is to train the PIDViT to achieve the consistency of facial expressions between frontal faces and multi-pose faces, and stage 2 is to use the student model pre-trained in stage 1 and train facial expressions further on target datasets. Finally, comprehensive experiments were conducted on three in the wild facial expression datasets, and the results validates the generalization of the PIDViT and its superiority over most state-of-the-art models.

Index Terms—Facial expression recognition, pose-invariant, visual transformer, teacher-student model, knowledge distillation, multi-pose face datasets

1 Introduction

In the past, understanding human emotions has always been a very important subject, especially in psychology [14], [15]. Due to the popularity of modern computer sciences, people try to use machines to identify human emotions. In particular, facial expressions are the most obvious signals affected by emotions in non-verbal languages [6]. Therefore, people have done a lot of research in the wild for facial recognition. Nowadays, facial expression recognition systems are widely used in driver fatigue monitoring [53], mental health assessment [20], and human-computer interaction applications [5]. Many facial expression recognition methods [9], [19], [21], [58] have achieved great success, but they only considered front facial images or facial images close to the front. Besides, unlike in-the-laboratory datasets, the facial images in the real world (or in the wild) are without lighting and pose control, so that it is a big challenge to recognize these facial expressions. At present, the existing methods are still difficult to recognize facial expressions accurately against the influences of pose variance and occlusion, especially for large-pose facial images. Furthermore, the head rotation also causes a self-mask; e.g., half of the face is masked on a profile-face image. Therefore, the robustness in the pose-invariance recognition is the most critical issue addressed by facial expression recognition systems.

Traditional methods using various hand-crafted features, such as Local Binary Pattern (LBP) [58], Histogram of Oriented Gradients (HOG) [11], and Gabor [33], cannot deal with the distortion of the non-linear facial texture effectively, caused by pose variations. In recent years, deep learning has been successfully applied to image classification, object detection, segmentation, and pose estimation. The deep learning for solving facial expression recognition usually makes use of synthesis techniques to facilitate discriminative feature learning or tries to boost the performances by designing new loss functions or network architectures. Recently, Generative Adversarial Networks (GAN) [50], [55], [56] and attention networks [48] have been widely used in facial expression recognition.

For pose variance, the facial expression of any pose should be consistent with the facial expression of the corresponding frontal face. The different expression probability distributions of frontal and multi-pose faces as shown in Fig. 1 are derived by ourselves. We built FairFace-3D to test the DAN model that identifies facial expressions with different probabilities. In other words, all frontal faces in FairFace-3D are identified as soft labels since no facial expressions (or ground-true labels) are supported in FairFace. In general, a frontal face is better than a multi-pose face for facial expression analysis. Thus, the concept is to propose a model learning how the expression probability distributions of multi-pose faces are consistent with (or the same as) those of frontal faces, thereby reducing the impact of multi-pose faces. In this paper, the PIDViT (i.e., Pose-Invariant Distilled Vision Transformer)

• *The authors are with the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Douliou, Yunlin 640, Taiwan. E-mail: {huangyf, m10817003}@yuntech.edu.tw.*

Manuscript received 9 January 2022; revised 15 October 2022; accepted 6 November 2022. Date of publication 9 November 2022; date of current version 29 November 2023.
(Corresponding author: Yin-Fu Huang.)
Recommended for acceptance by A. Dhall.
Digital Object Identifier no. 10.1109/TAFFC.2022.3220972

1949-3045 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

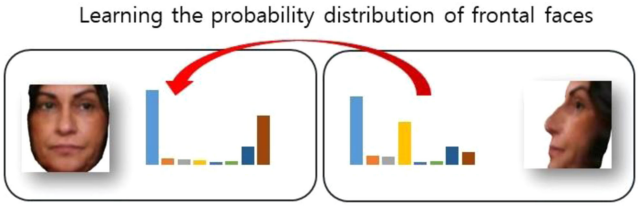

Fig. 1. Learning the probability distribution of consistent facial expressions.

using the teacher-student model for the probability distributions of facial expressions of frontal and multi-pose faces is proposed and then the essence is extracted as the training material for the student model. For a multi-pose face, it can be regarded as a partially occluded face. Here, we use the Vision Transformer as the backbone to focus on each area of a face, in order to reduce the expression recognition error. Finally, to evaluate the pose variance and occlusion impact on the proposed model and its superiority over most state-of-the-art models, three facial expression datasets in the wild are tested in the experiments; i.e., AffectNet [36], Real-world Affective Faces Database (RAF-DB) [29], and Static Facial Expressions in the Wild (SFEW2.0) dataset [12].

In summary, we identify the novelty and contributions in this paper as follows:

1. The multi-pose face dataset FairFace-3D was built using FairFace, which contains 10456 subjects with 9 age groups and 7 races.
2. The teacher-student model for the probability distributions of facial expressions of frontal and multi-pose faces is the first one trained to achieve their consistency in facial expressions.
3. A new indicator "consistency" is proposed to evaluate the impact of multi-pose faces on facial expression recognition.
4. Comprehensive experiments were conducted on three in the wild datasets, and the results validates the generalization of the PIDViT and its superiority over most state-of-the-art models.

The remainder of this paper is organized as follows. In Section 2, the previous research related to facial expression recognition is described. In Section 3, the overview of the proposed facial expression recognition model is introduced. In Section 4, a multi-pose face dataset called FairFace-3D is built from the original FairFace. In Section 5, the two-stage training model is described in detail to solve the pose variance and occlusion issues on expression recognition. In Section 6, to evaluate the generalization of the proposed model and its superiority over most state-of-the-art models, three datasets in the wild are tested in the experiments. Finally, we make conclusions and give future work in Section 7.

2 Related Work

In general, Facial Expression Recognition (FER) consists of three stages, namely face detection, feature extraction, and facial expression classification. In face detection, some of the current mainstream face detectors are MTCNN [54] and Dlib [2], which can detect faces in complex backgrounds and perform face trimming and alignment. Then, the feature extraction methods can generally be divided into the methods based on geometric features and the ones based on texture features. Unlike the geometric feature-based methods [21], [52], the texture feature-based methods used Convolutional Neural Networks (CNNs) to extract image features. Many studies have found that CNNs are robust to face position and scale changes, and are superior to other feature extraction methods [18], [19], [35]. After extracting image features, the features are usually forwarded to supervised classifiers, such as Support Vector Machines (SVMs), Softmax layer, and logistic regression for facial expression recognition. Since eyebrows, eyes, and mouth are closely related to facial expressions, Chen et al. [9] proposed to improve model performance by cutting these areas as the Regions of Interest (ROI). Later, subsequent studies also showed that using attention networks can highlight areas that affect facial emotions [16], [24], [31], [47], [48]. Besides, according to psychological research, different ethnic groups have different ways of expressing facial emotions [8], [10]. For example, Western Caucasians tend to express emotions through their eyes and chins, while Southeast Asians express emotions through the area around their mouths. Besides, recent studies have adopted two-stage training to improve facial expression recognition; that is, learning facial attributes first and then recognizing facial expression can improve the accuracy of facial expression recognition [16], [39].

For in the wild datasets, the accuracy of facial expression recognition is restricted, which is affected by poses, light, and unobvious expressions. For poses in the study of Sengupta et al. [41], the performance of most face models from Frontal-Frontal to Frontal-Profile verification reduces the accuracy by at least 10%. Most deep learning methods did not directly consider the effects of multiple poses, but indirectly aligned and regulated faces to reduce the effects of poses; in other words, they did not fully deal with the pose problem [11], [48]. In addition, augmenting datasets by generating multi-pose face images is also a solution to the pose variant problem. Thus, Lai and Lai [28] proposed GANs to turn a human face into a frontal face while maintaining identity and expression. Later, Zhang et al. [55], [56] used GANs to expand FER datasets by generating facial images with different expressions in arbitrary poses, but the generated images affected the performance of facial expression recognition. Xie et al. [50] also used GANs to separate expressions from faces. They built two independent branches used to process faces and expressions, and then the two branches are combined for encoding. Finally, the decoder generates a composite image, and the expression branch is used for classification. DAN proposed by Wen et al. [48] is basically a CNN model which is compose of Feature Clustering Network (FCN), Multi-head cross Attention Network (MAN), and Attention Fusion Network (AFN). The FCN extracts robust features by adopting a large-margin learning objective to maximize class separability. The MAN instantiates a number of attention heads to simultaneously attend to multiple facial areas and build attention maps on these regions. The AFN distracts these attentions to multiple locations before fusing the attention maps to a comprehensive one.

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

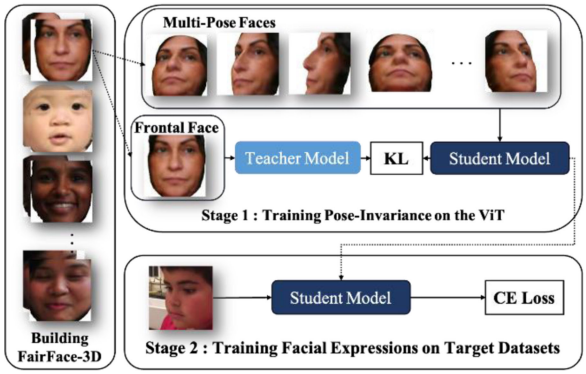

Fig. 2. System overview.

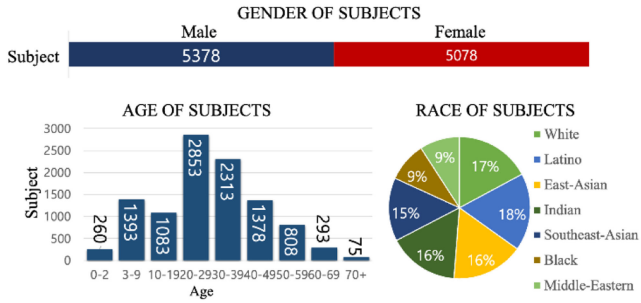

Fig. 3. Gender, ages, and races in FairFace-3D.

The Vision Transformer (ViT) proposed by Dosovitskiy et al. [13] is the first method to apply a transformer model to image classification by dividing an image into non-repetitive 16×16 image patches and inputting these patches into a linear projection, where the classification makes use of an extra learnable "class token". Although the model uses the in-house JFT-300M dataset with far more data than ImageNet to fine-tune the pre-trained model on ImageNet and achieves the most advanced performance, the model has to rely on large-scale data, compared with convolutional neural networks. To solve this problem, Touvron et al. [44] proposed the Data-efficient image Transformer (DeiT) using two classification tokens; i.e., class token and distillation token. The DeiT fine-tunes a pre-trained teacher model using the distillation procedure, and performs better than the ViT, even only trained on ImageNet. Recent studies have begun to use ViT to recognize facial expressions. However, Aouayeb et al. [4] validated that only using ViT for learning is not good enough, and they tried to add the attention block of Squeeze and Excitation in front of the MLP Head to improve ViT learning. Besides, since a CNN model cannot effectively recognize facial expressions from a global perspective, Huang et al. [24] and Ma et al. [34] proposed to extract features through CNNs, and then input the features into ViT to solve the weakness of long-range induction bias in learning on the CNN-based FER model.

3 System Overviews

In this section, we propose a facial expression recognition model to solve the pose and occlusion impact on expression recognition. The model is divided into 3 parts; i.e., building multi-pose face datasets, training pose-invariance on the ViT (Stage 1), and training facial expressions on target datasets (Stage 2), as shown in Fig. 2. For building multi-pose face datasets, face reconstruction technology was used to build a multi-pose face dataset called FairFace-3D from the original FairFace [26] where age and ethnicity are widely distributed. FairFace-3D was used to train pose-invariance on the ViT. Next, the training can be further divided into two stages; stage 1 is to train the ViT to achieve the consistency between frontal faces and multi-pose faces, and stage 2 is to use the student model pre-trained in stage 1 and train facial expressions further on target datasets.

4 Building Datasets

Till now, no multi-pose face datasets covering a wide range of people were used to fulfil the comprehensive study in existing facial expression recognition research. In the past, as shown in Table 1, most multi-pose face datasets contained less than 1000 subjects. Although FairFace (not shown in Table 1) has more than 108501 images (or subjects), each subject has only one pose. Thus, for our study, face reconstruction technology was used to build a multi-pose face dataset called FairFace-3D from the original FairFace. The new multi-pose face dataset has 10456 subjects spanning a wide of people, gender, ages, and races, as shown in Fig. 3.

FairFace-3D was built from the original FairFace using five steps, as shown in Fig. 4. At step 1, the head posture evaluation method proposed by Zhou and Gregson [60] was used to delete the images of the original FairFace where the yaw, pitch, and roll angles are more than 8 degrees. At step 2, we constructed the 3D images in the dataset using the 3D face reconstruction model proposed by Jackson et al. [25]. At step 3, face images were generated each from the

TABLE 1
Multi-pose Face Datasets

Datasets	Subjects	Types	Races	Ages
BU-3DFE [51]	100	3D	Black, East-Asian, Hispanic-Latino, Indian, Middle-Eastern, and White	18∼70
Multi-PIE [22]	337	Images	African-American, Asian, European-American, and others	27.9 (average)
M2FPA [30]	229	Images	Asian	-
ColorFERET [38]	994	Images	Asian, Black, Hispanic, Middle-Eastern, Native-American, Pacific-Islander, Southeast-Asian, White, and others	-
FairFace-3D (ours)	10456	3D& Images	Black, East-Asian, Indian, Latino, Middle-Eastern, Southeast-Asian, and White	0∼70+

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

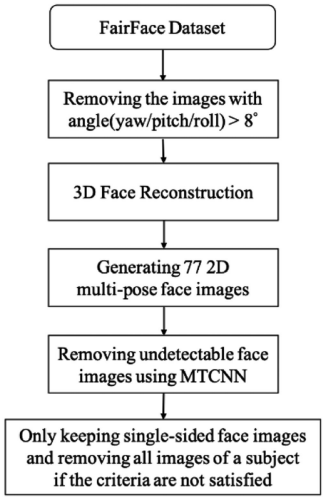

Fig. 4. Building FairFace-3D from FairFace.

combinations of per 15 degrees at the yaw angle of -75 to 75 degrees and the pitch angle of -45 to 45 degrees. Thus, in total, we have $11 \times 7 = 77$ face images for a subject. At step 4, we used MTCNN [54] to detect these 77 face images and removed undetectable images. At step 5, three criteria are used to filter out detectable images. For criterion 1, for a subject, if the face image with the yaw and pitch angles equal to 0 degree does not exist in the detectable images, all images of the subject are removed. For criterion 2, since face images of a subject may have side occlusion problems, if the left and right symmetrical parts both exist in the detectable images, only the side with higher confidences remains. Thus, a subject can have at most $6 \times 7 = 42$ face images. For criterion 3, if the number of images for a subject is less than 30 (i.e., an empirical value, about 70% of 42 face images), all images of the subject are also removed.

For face alignment and highlighting the face part in an image, MTCNN was also used to obtain five facial landmarks; i.e., the center point of the left eye, the center point of the right eye, the tip point of the nose, the left corner point of the mouth, and the right corner point of the mouth. For face alignment, both eye coordinators are used to calculate for rotating the inclined face to a correct angle. For highlighting the face part in an image, the central point of a clipping image is defined as (c_x, c_y) where c_x is the x-coordinate of the middle point between both eyes and c_y is the y-coordinate of the nose tip. Then, the clipping range from (c_x, c_y) is defined as follows

$$\underset{p \in \textit{five facial landmarks}}{Max} \sqrt{(c_x - p_x)^2 + (c_y - p_y)^2} * 3 \quad (1)$$

Since the similarity of the images with different pitch angles, which are generated by face alignment, is very high when their yaw angles are greater than 45 degrees, only one image needs to be retained in the dataset; i.e., only the image with the highest confidence among the seven images is retained. Finally, a subject can have at most $42 - 6 \times 2 = 30$ multi-pose images. In short, the statistics of the yaw and pitch angles of the face images in FairFace-3D are shown in Fig. 5.

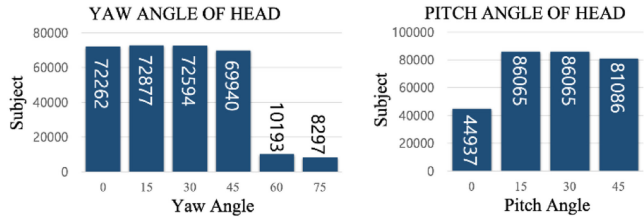

Fig. 5. Yaw and pitch angles of the face images in FairFace-3D.

5 Proposed Models

In the section, a facial expression recognition model with the generalization capability suitable for in the wild datasets is proposed, where the processing can be divided into two stages. First, the Vision Transformer (ViT) [13] is adopted as the backbone of our model. Then, in stage 1, knowledge distillation is used to train pose-invariance on the ViT where the consistency of facial expressions between frontal faces and multi-pose faces can be achieved. Finally, in stage 2, the facial expression recognition model is trained on target datasets.

5.1 Vision Transformer Backbone

A transformer has been proven to achieve good performances on many visual applications, and the ViT is the first method to apply a transformer model to image classification. For a face image I_in with the size of $H \times W \times C$ where (H, W) is the image resolution and C is the number of channels, we reshape it into a sequence of 2D I_p with the size of $N \times (P \times P \times C)$ where (P, P) is the resolution of each image patch and $N = H \times W / P \times P$ is the number of patches. In our model, P is set to 16 and the image dimension is $224 \times 224 \times 3$. These patches serve as the effective input sequence for the transformer. The transformer flattens the patches and maps to a D-dimensional vector Z_p through a trainable linear projection, where D is the constant latent vector size for each patch in all the layers. Besides, the ViT borrows BERT and adds a special class token which is a learnable embedding patch Z_0. Thus, the input of the transformer encoder is a sequence of patch embedding, and the output is a sequence of patch features of the same length. Here, avoiding to learn the puzzle through the semantics of patches, the position embedding of a patch is also considered in the model; i.e., adding the position embedding E_{pos} to the D-dimensional vector Z_p. Thus, the size of input embedding vector of the transformer encoder is $(N + 1) \times D$ (including the special class token). Within the transformer encoder, the self-attention is permutation-invariant; i.e., shuffling the order of tags in the sequence would not change the results. Besides, the advantage of the self-attention does not have a fixed and limited receptive field like convolution; i.e., it can obtain long-range information. Finally, instead of using the simple pooling strategy, the ViT can realize the image classification based on the final output features plus a linear classifier. The classification head is implemented by an MLP with one hidden layer.

5.2 Training Pose-Invariance on the ViT

According to the previous studies [11], [28], [41], [48], [50], [55], [56], multi-pose faces affected the accuracy in the real-

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

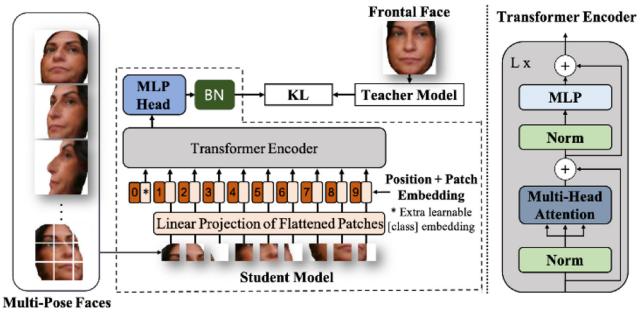

Fig. 6. Stage 1: ViT using knowledge distillation to train pose-invariance.

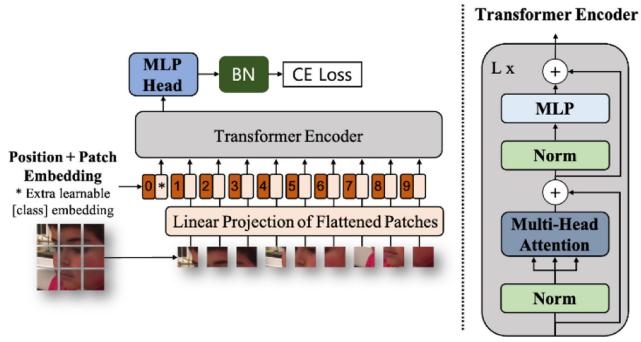

Fig. 8. Stage 2: ViT applied to facial expression recognition applications.

world facial expression recognition. To reduce the influences of multi-pose faces, our model uses knowledge distillation in stage 1 to train pose-invariance and achieves the consistency of facial expressions between frontal faces and multi-pose faces. Knowledge distillation was first proposed by Bucila et al. [7] and initially used in model compression. Later, Hinton et al. [23] proposed the teacher/student concept framework to realize knowledge distillation. Different from common knowledge distillation using hard labels, we use the teacher model to generate soft labels and adopt semi-supervised learning. The weights of the teacher model are trained first and then the essence is extracted as the training material for the student model. This enables the student model to compete in accuracy with the teacher model or even to be better than the teacher model.

For stage 1 as shown in Fig. 6, a Vision Transformer is used in the student model, the classification head is implemented by an MLP with a hidden layer, and batch normalization (BN) is added after the MLP. FairFace-3D is divided into two parts: frontal faces (i.e., yaw angle and pitch angle are both 0 degree) used in the teacher model and multi-pose faces used in the student model, respectively. During training the student model, a face image is selected at random from multi-pose faces as the input of the student model, and data augmentation and background filling (background - white area) are done at random on the face image, as shown in Fig. 7. The data augmentation on multi-pose faces, including horizontal flipping, affine transformations, and random erasing, is used to learn more and/or avoid overfitting during training. Besides, the weights of the teacher model are frozen, and data augmentation and background filling are not done on frontal faces, since it acts as the teacher.

For the candidates of the teacher model, they should be robust enough to provide the soft labels of frontal faces. Here, the DAN model [48] pre-trained on AffectNet or RAF-DB is chosen as the teacher model where the training accuracy on AffectNet-8 is 61.79%. Besides, negative labels also contain a lot of hidden information, which can bring more information to train the student model.

Fig. 7. Data augmentation and background filling at random on multi-pose faces.

As mentioned, the teacher model has been pre-trained on target datasets, and the learned parameters are denoted as

$$\hat{\theta}^t = \arg\min_{\theta} \sum_{i=1}^{n} L_{CE}^t(x_i, y_i; [\theta_{DAN}, W]) \tag{2}$$

where the superscript t denotes parameters in the teacher model, L_{CE}^t denotes the cross-entropy loss for the teacher training, and θ_{DAN} denotes parameters of the DAN. The output probability of a given input x_i can be expressed as

$$\hat{y}_i = P^t(y_i|x_i) = softmax\left(\frac{W * DAN\left(x_i; \hat{\theta}^t\right)}{T}\right) \tag{3}$$

$$\text{with } softmax(x)_j = \frac{e^{x_j}}{\sum_{i=1}^{n} e^{x_i}} \tag{4}$$

where $P^t(\cdot|\cdot)$ represents the output probability of the teacher model passing through the softmax, $\hat{y}_i$ is a soft label, and T normally set to 1 is the temperature used in knowledge distillation. Using a higher value for T (set to 2 in our experiments) produces a softer probability distribution over classes [17]. The output probability of the student model is similar to the above formula. Let θ^s represent the learning parameters of the student model, and $P^s(\cdot|\cdot)$ represent the output probability of the student model. The loss between the teacher prediction and the student prediction is called Kullback-Leibler (KL) divergence loss [23] as shown in Fig. 6 and defined as

$$L_{KL} = -\sum_{i=1}^{n}\sum_{c=1}^{C}\left[P^t(y_i = c|x_i; \hat{\theta}^t)\log\frac{P^s(y_i = c|x_i; \theta^s)}{P^t(y_i = c|x_i; \theta^t)}\right] \tag{5}$$

where C is the number of all categories and n is the number of subjects.

5.3 Training Facial Expressions on Target Datasets

In stage 2, the facial expression recognition model is trained on target datasets; i.e., further training the student model on target datasets such as AffectNet [36]. Thus, as depicted in Fig. 8. Since AffectNet is a highly imbalance dataset where happiness occupies 47.2%, and contempt occupies only 1.2%, the imbalance issue should be addressed in the loss function. The cross-entropy loss is defined as

$$L_{CE}^s = -\sum_{i=1}^{n}\sum_{c=1}^{C}[1[y_i = c]\log P^s(y_i = c|x_i; \theta^s)] \tag{6}$$

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

Then, the categories with fewer face images can be weighted in the cross-entropy loss function as follows.

$$weight_i = \frac{number\ of\ images\ in\ category_{max}}{number\ of\ images\ in\ category_i} \quad (7)$$

where catagotymax is the category with the most face images.

6 Experiments

In this section, to evaluate the pose variance and occlusion impact on the proposed model and its superiority over most state-of-the-art models, three facial expression datasets in the wild are selected and tested in the experiments; i.e., AffectNet [36], Real-world Affective Faces Database (RAF-DB) [29], and Static Facial Expressions in the Wild (SFEW2.0) dataset [12]. In the following subsections, we introduce the datasets used in the experiments, the experimental environments, and the evaluation indicators used to compare the proposed model with the state-of-the-art models. Finally, we present the experimental results of the model on AffectNet, RAF-DB, and SFEW2.0, and validate its superiority over most state-of-the-art models.

6.1 Datasets

AffectNet is the largest facial expression dataset, which contains more than 1 million face images collected from the Internet, and about half of the images (∼440K) were manually annotated. As collected from the Internet, the face images in AffectNet are with different backgrounds, light, poses, and races. This brings more challenges to facial expression recognition, compared with other laboratory-generated datasets. In this study, we used 8 classes and 7 classes of facial expressions in the experiments, respectively. The 8 classes of facial expressions include neutral, happy, sad, surprise, fear, disgust, anger, and contempt, whereas the 7 classes include all but contempt, like most studies. Since the test set is unavailable, the official validation set is used for testing purposes. For the 8-class experiments, 287K images are for training and 4000 images for test; for the 7-class experiments, 283K images are for training and 3500 images for test. Although AffectNet has an imbalance training set (47.2% for happy, 26.2% for neutral, 8.8% for sad, 8.6% for anger, 4.8% for surprise, 2.1% for fear, 2.1% for disgust, and 1.2% for contempt), the test set is balanced.

RAF-DB is a large-scale facial expression database with more than 29670 face images with uncontrolled poses and lighting. Based on crowdsourced annotations, each image was independently marked by about 40 annotators. The dataset contains 7 classes of basic emotions and a dual-label subset with 12 classes of compound expressions. In the experiments, we used the images with 7 basic emotions (i.e., surprise, fear, disgust, happy, sad, anger, and neutral), and 12271 images are for training and 3068 images for test.

SFEW2.0 was constructed by extracting frames from Acted Facial Expressions in the Wild (AFEW) database, covering unconstrained facial expressions. It contains 958 training samples, 436 verification samples, and 372 test samples. Each image was marked by two independent labelers as one of 7 expressions, including anger, disgust, fear, happy, sad, surprise, and neutral. Since the labels of the test samples are private and unavailable, we used the validation set instead in the experiments.

TABLE 2
Number of Images Used in the Training, Validation, and Test

Datasets	Training	Validation	Test
AffectNet-7	283901	3500	-
AffectNet-8	287651	4000	-
RAF-DB	12271	-	3068
SFEW2.0	958	436	372

Finally, the number of images used in the training, validation, and test of the three datasets is summarized and shown in Table 2.

6.2 Experimental Environments

Our model was implemented using the Pytorch toolbox [37], and all the experiments were conducted on a single NVIDIA Titan RTX 24GB GPU card where about 11G GPU memory was required during training. The Adam optimizer [27] was adopted in the model where the Kullback-Leibler divergence loss [23] was measured in the pose-invariant training (i.e., stage 1), and the categorical cross-entropy loss with weights was measured in the target expression training (i.e., stage 2), as mentioned in Section 5. The batch sizes of the two training stages were set to 64, but the epochs are different; i.e., stage 1 is 25 for FairFace-3D, and stage 2 is 10 for AffectNet-7 (i.e., 7 classes) and AffectNet-8 (i.e., 8 classes), 80 for RAF-DB, and 3 for SFEW2.0. The learning rate was set to 0.00001, and the learning rate adjustment strategy was adopted only for AffectNet-7, AffectNet-8, and RAF-DB. For AffectNet-7 and -8, the current learning rate was reduced by 60% per epoch, and for RAF-DB, the current learning rate was reduced by 60% per 15 epochs. For the input image processing as mentioned in Section 4, if the facial landmarks of an image in the validation or test set cannot be obtained through MTCNN, the original facial landmarks provided in the dataset are used instead. Besides, as mentioned in Section 5, the augmentation method conducting horizontal flipping, random transformations, and random erasing on images was used to learn more and/or avoid overfitting in different epochs during training. DAN was pre-trained on dataset Microsoft Celeb. Before using DAN as the teacher model in stage 1, we pre-train it on datasets AffectNet or RAF-DB for our model, depending on which is the current target dataset. For most state-of-the-art models, pre-training was first performed on face-related datasets. Instead, we use the ViT with the initial weights from the timm repository "vit_base_patch16_224" (trained on ImageNet-21k and ImageNet2012) which is a deep-learning library created by Ross Wightman [49].

6.3 Evaluation Indicators

To compare the proposed model with the state-of-the-art models, the commonly used indicators, namely overall accuracy and average accuracy, are used and defined as follows:

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
Comparisons Among the Models on Different Datasets

Models	Pre-trained Datasets	Year	AffectNet-7 Overall	AffectNet-8 Overall	RAF-DB		SFEW2.0 Overall
					Overall	Average	
RAN [47]	MS Celeb	2020	-	0.5950	0.8690	-	0.5640
ESR-9 [43]	-	2020	-	0.5930	-	-	-
R-FENet [46]	-	2020	-	0.6095	-	-	0.5597
PSR [45]	ImageNet	2020	0.6377	0.6068	0.8898	0.7957*	-
CVT [34]	MS Celeb +ImageNet	2021	-	0.6185	0.8814	0.8186*	-
FER-VT [24]	-	2021	-	-	0.8826	0.7583*	-
SPWFA-SE [31]	ImageNet	2021	-	0.5923	0.8631	0.7843*	-
EfficientFace [59]	MS Celeb	2021	0.6370	0.5989	0.8836	-	-
ARM [42]	ImageNet	2021	0.6520	0.6133	0.9042	0.8277*	0.5871
MViT [32]	ImageNet	2021	0.6457	0.6140	0.8862	0.8038*	-
Emotion-GCN [3]	Wikipedia	2021	**0.6646**	0.6194	-	-	-
DACL [17]	MS Celeb	2021	0.6520	-	0.8778	0.8044*	-
EfficientNet-B2 [39]	ImageNet	2021	0.6634	0.6242	-	-	-
DAN [48]	MS Celeb	2021	0.6569	0.6209	0.8970	0.8532	0.5788
PIDViT (Ours)	FairFace-3D	-	0.6580	**0.6252**	**0.9071**	**0.8533***	**0.6008**

The average accuracy is calculated using Equation (9), not cited from the original papers.

$$overall\ accuracy = \frac{number\ of\ correctly\ predicted\ items}{total\ number\ of\ items\ to\ be\ predicted} \tag{8}$$

$$average\ accuracy = \frac{sum\ of\ predicted\ accuracy\ for\ each\ class}{number\ of\ classes} \tag{9}$$

Besides, the face expression of a subject should have the same expression in all his/her profile faces. In the past, postures were seldom considered in the literature, but postures are indeed a big problem in the facial expression recognition. In other words, recognizing facial expression should not be affected by the yaw and pitch angles of face images, and it is very crucial for a model to be able to recognize emotions of multi-pose faces. Here, we propose another new indicator "consistency" to evaluate the impact of multi-pose faces on facial expression recognition, defined as follows:

$$consistency = \frac{number\ of\ faces\ with\ the\ same\ expression\ as\ front\ faces}{total\ number\ of\ faces\ (no\ front\ faces)} \tag{10}$$

6.4 Comparisons With the State-of-the-Art Models

As shown in Table 3, our model is compared with the state-of-the-art models on AffectNet (overall accuracy), RAF-DB (overall and average accuracy), and SFEW2.0 (overall accuracy). For the state-of-the-art models, CVT, FER-VT, and MViT also used transformers as the backbones, like our model. However, different technologies were used in their models. Models CVT and FER-VT 1) extract features through CNNs, 2) then transform features into a sequence of visual tokens, and 3) input visual tokens into the transformers. Model MViT filters out complex backgrounds and occlusion of face images using a mask generation network, and rectifies incorrect labels in the wild datasets. Here, our model solves the face pose variant problem to improve accuracy. As shown in Table 3, our model is in the 3rd place (65.80%) on AffectNet-7, in the 1st place (62.52%) on AffectNet-8, in the 1st place (overall 90.71%) on RAF-DB, in the 1st place (average 85.33%) on RAF-DB, and in the 1st place (60.08%) on SFEW2.0. Obviously, our model is superior to most state-of-the-art models.

6.4.1 Experimental Results of Each Category on AffectNet

For AffectNet-7, as shown in Table 4 although our model is not the best for each category, the standard deviation of accuracy is lower than the other models; in other words, our model is more stable than the other models. For AffectNet-8, as shown in Table 5 our model is still not the best for most categories, but the average accuracy is the best and the standard deviation is much lower than the other models. We found that recent models are difficult in identifying contempt (less than 50% accuracy in most studies), while our model can reach 60% accuracy in identifying contempt. Besides, as shown in Fig. 9a, the confusion matrix on AffectNet-7 shows that fear-surprise and disgust-anger are more likely to confuse in our model. The same cases also occur on AffectNet-8, as shown in Fig. 9b, besides happy-contemp. Nevertheless, these situations also happen in the other models. For the common prediction errors as shown in Fig. 10, the indistinguishableness between surprise and fear expressions is due to widening eyes and opening mouths, the indistinguishableness between disgust and anger expressions is due to frowning, and the indistinguishableness between happy and contempt expressions is due to all face parts are similar to each other, except contempt with the upward unilateral corner of the mouth. In short, these expressions are sometimes difficult for humans to distinguish.

6.4.2 Experimental Results of Each Category on RAF-DB

As shown in Table 6 our model is the best for most categories, except for neutral and happy. Similarly, the average accuracy is the best and the standard deviation is much lower than the other models. We also found that our model, like the other models, is more difficult to recognize fear and disgust. Besides, as shown in Fig. 11, the confusion matrix

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

TABLE 4
Comparisons of Each Category on AffectNet-7

Models	Year	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Average	Std.
PSR [45]	2020	0.6300	0.8400	0.5900	**0.6500**	**0.6500**	0.5800	0.5200	0.6371	0.1006
SPWFA-SE [31]	2021	**0.7300**	**0.9500**	0.5900	0.4500	0.4900	0.5900	0.3500	0.5929	0.1985
ARM [42]	2021	0.6500	0.8700	0.6400	0.6100	0.6200	0.5300	**0.6400**	0.6514	0.1045
Emotion-GCN [3]	2021	0.6400	0.8800	0.6800	0.6000	**0.6500**	**0.6400**	0.5800	**0.6671**	0.0995
DACL [17]	2021	0.6420	0.8780	**0.6820**	0.6100	0.6100	0.5620	0.5800	0.6520	0.1072
PIDViT (Ours)	-	0.6240	0.8360	0.6660	0.6440	0.6320	0.5780	0.6260	0.6580	**0.0829**

TABLE 5
Comparisons of Each Category on AffectNet-8

Models	Year	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	Average	Std.
ESR-9 [43]	2020	0.5800	0.7740	0.6140	0.5540	0.6360	0.5380	0.5900	0.4540	0.5925	0.0919
R-FENet [46]	2020	0.4800	0.7100	0.6300	0.6500	0.6300	**0.6200**	0.5500	**0.6200**	0.6113	0.0688
PSR [45]	2020	0.6200	0.8100	0.6200	0.5100	**0.7100**	0.5500	0.5300	0.4900	0.6050	0.1098
CVT [34]	2021	0.5300	0.7800	**0.6400**	0.6500	0.6200	0.5500	**0.6300**	0.5600	0.6200	0.0789
ARM [42]	2021	**0.6300**	**0.8600**	**0.6400**	0.6100	0.6200	0.5200	**0.6300**	0.4000	0.6138	0.1289
PIDViT (Ours)	-	0.5380	0.7100	0.6300	**0.6720**	0.6480	0.6000	0.6020	0.6010	**0.6251**	**0.0525**

on RAF-DB shows that fear-surprise is still more likely to confuse in our model, like on AffectNet-7. For the common prediction errors as shown in Fig. 12, the indistinguishableness between surprise and fear expressions is still due to widening eyes and opening mouths.

6.4.3 Experimental Results of Each Category on SFEW2.0

As mentioned in Section 6.1, since SFEW2.0 has only 958 training samples, similar to some state-of-the-art models, our model was first trained using RAF-DB and then fine-tuned using SFEW2.0. As shown in Table 3, our model has the best overall accuracy 60.08%. Besides, to observe the impact of the cross-dataset on our model, the model was only trained using RAF-DB (i.e., not fine-tuned using SFEW2.0) and then directly tested using SFEW2.0. As a result, our model gets the overall accuracy 54.82%, obviously lower than the former overall accuracy 60.08%. As shown in Table 7, our model without fine-tuning is the best for most categories, except for fear and disgust, and the average accuracy is also the best. Besides, as shown in Fig. 13, We found that our model is more difficult to recognize fear and disgust.

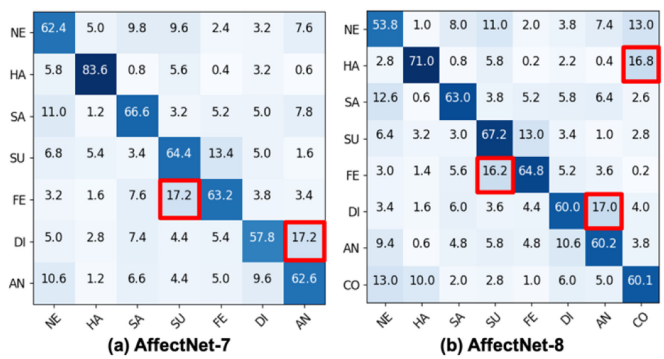

Fig. 9. Confusion matrices of our model on AffectNet-7 and AffectNet-8.

6.4.4 Experimental Results on Occlusion and Pose Variance

To show that our model is robust to occlusion and pose variant in the wild environment, we conducted the experiments on Occlusion-AffectNet-8, Pose^{30+}-AffectNet-8, Pose^{45+}-AffectNet-8, Occlusion-RAF-DB, Pose^{30+}-RAF-DB, and Pose^{45+}-RAF-DB. For the compared targets, RAN adaptively captures the importance of facial regions for occlusion and pose variant, CVT and FER-VT combine CNN and visual transformer technology, and DAN. As shown in Table 8, our model performs better than RAN, CVT, and DAN on both Occlusion-AffectNet-8 and Pose^{30+}-AffectNet-8. It is a little worse than CVT, but still better than RAN and DAN on Pose^{45+}-AffectNet-8. As for on Occlusion-RAF-DB, Pose^{30+}-RAF-DB, and Pose^{45+}-RAF-DB, our model performs all better than RAN, CVT, FER-VT, and DAN. This verifies that our model is definitely robust to occlusion and pose variant. In the paper, our model is not trained particularly for occlusion like MViT filtering out complex backgrounds and occlusion of face images. Instead, we only use the most common random erasing to handle images with occlusion. We believe that processing pose variant can effectively deal with occlusion problems because the head rotation also causes a self-mask; i.e., pose variants can also cause occlusion problems. Besides, as shown in Fig. 14a, the confusion matrix on Occlusion-AffectNet-8 shows that neutral-surprise, happy-contempt, fear-surprise, disgust-anger, and anger-surprise are more likely to confuse in our model. The same cases except anger-surprise also occur on Pose^{30+}-AffectNet-8, as shown in Fig. 14b. Finally, the confusion matrix on Pose^{45+}-AffectNet-8 shows that neutral-surprise, sad- neutral, fear-surprise, and disgust-anger are more likely to confuse in our model, as shown in Fig. 14c.

6.5 Consistency in Facial Expressions of Multi-Pose Faces

In generally, the face expression of a subject should have the same expression in all his/her profile faces. As mentioned

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

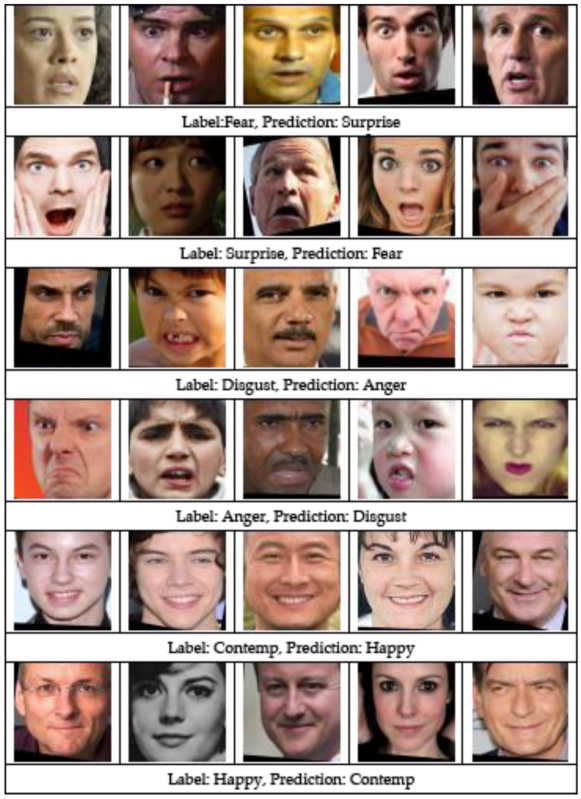

Fig. 10. Common prediction errors in AffectNet-8.

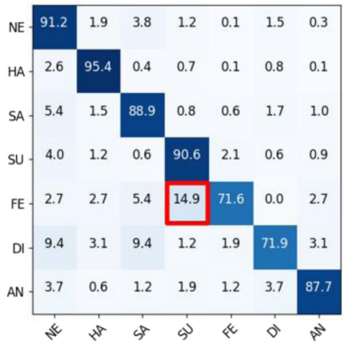

Fig. 11. Confusion matrix of our model on RAF-DB.

in Section 6.3, the new indicator "consistency" is proposed to evaluate the impact of multi-pose faces. Here, we conducted the experiments to evaluate the consistency on FairFace-3D for DAN [48] and PIDViT, based on fair training. In other words, DAN and PIDViT are trained on FairFace-3D

Fig. 12. Common prediction errors in RAF-DB.

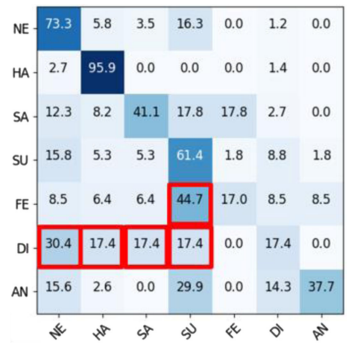

Fig. 13. Confusion matrix of our model on SFEW2.0.

and tested on FairFace-3D. Since the target dataset is FairFace-3D and the metric is "consistency", our model used here is only trained using FairFace-3D in stage 1, not the final model trained using the original target datasets AffectNet or RAF-DB in stage 2. As shown in Table 9, we found that the consistency of our model is more than 70% at yaw angles less than 75 degrees and better than that of DAN at most yaw angles. Our model is better than DAN at the yaw angle of 75 degrees, although most features are lost in facial expression recognition. Similarly, our model is still better than DAN at most pitch angles. Absolutely, this also verifies that our model can effectively solve the pose variant problem in the stage 1 training.

6.6 Visualization

To observe intuitively the solution to the pose variant problem in our model, we randomly select a subject from FairFace-3D and use Gradient-weighted Class Activation Mapping [40] (i.e., Grad-CAM) to view the attention parts of pose faces, as shown in Fig. 15. The red areas in the faces are the attention parts learned in our model. Besides, the occlusion problem is also effectively solved in our model, as shown in Fig. 16. For an occluded face, our model can

TABLE 6
Comparisons of Each Category on RAF-DB

Models	Year	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Average	Std.
PSR [45]	2020	0.9000	**0.9600**	0.8700	0.8600	0.5900	0.5400	0.8500	0.7957	0.1624
CVT [34]	2021	0.8750	0.9409	0.8724	0.8541	0.6486	0.6812	0.8580	0.8186	0.1093
FER-VT [24]	2021	0.8118	0.9249	0.8239	0.7835	0.6486	0.5562	0.7593	0.7583	0.1214
SPWFA-SE [31]	2021	0.8600	0.9300	0.8400	0.8800	0.5900	0.5900	0.8000	0.7843	0.1384
ARM [42]	2021	**0.9790**	0.9540	0.8390	0.9030	0.7030	0.6440	0.7720	0.8277	0.1272
MViT [32]	2021	0.8912	0.9561	0.8745	0.8754	0.6081	0.6375	0.7840	0.8038	0.1337
DACL [17]	2021	0.8706	0.9392	0.8431	0.8693	0.6622	0.6500	0.7963	0.8044	0.1098
PIDViT (Ours)	-	0.9120	0.9540	**0.8890**	**0.9060**	**0.7160**	**0.7190**	**0.8770**	**0.8533**	**0.0958**

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

TABLE 7
Comparisons of Each Category on SFEW2.0

Models	Training set	Year	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Average
Zhang et al. [56]	BU-3DFE	2020	0.2000	0.5119	0.2920	0.1870	0.1765	0.2488	0.2909	0.2724
Zhang et al. [57]	BU-3DFE	2020	0.2500	0.4915	0.2600	0.2174	0.2157	**0.3115**	0.2909	0.2910
Almeida et al. [1]	FER2013	2021	0.3500	0.8200	0.0000	0.0800	**0.3100**	0.1000	0.4000	0.2943
PIDViT (Ours)	FairFace-3D +RAF-DB	-	**0.7330**	**0.9590**	**0.4110**	**0.6140**	0.1700	0.1740	**0.3770**	**0.4911**

TABLE 8
Comparisons on Occlusion and Pose-Variance

Models	AffectNet-8			RAF-DB		
	Occlusion	Pose (> 30)	Pose (> 45)	Occlusion	Pose (> 30)	Pose (> 45)
RAN [47]	0.5850	0.5390	0.5319	0.8272	0.8674	0.8520
CVT [34]	0.6298	0.6061	**0.6100**	0.8395	0.8797	0.8835
FER-VT [24]	-	-	-	0.8432	0.8803	0.8608
DAN [48]	0.6149	0.5849	0.5860	0.8624	0.8909	0.8692
PIDViT (Ours)	**0.6530**	**0.6075**	0.6012	**0.8747**	**0.9006**	**0.8978**

effectively identify the occluded area and focus on the unoccluded face area.

6.7 Data Imbalance Impact

Since most in the wild datasets have the data imbalance issue such as 47.2% for happy in Affectnet-8, dealing with data imbalance is particularly important. In general, sampling methods were frequently used to solve the data imbalance issue. However, in this paper, the imbalance issue is addressed in the loss function, as mentioned in Section 5.3; in other words, the categories with fewer face images are weighted in the cross-entropy loss function. Here, the experiments are conducted to explore the effectiveness of weighting cross-entropy loss and different sampling methods. In general, sampling methods can be divided into oversampling, undersampling, and mix-sampling. The oversampling method works by resampling the instances of each category (except happy) so that the sample size is equal to the data size of category happy. The undersampling method works by resampling the instances of each category (except contempt) so that the sample size is equal to the data size of category contempt. The mix-sampling method works by oversampling or undersampling the instances of each category so that the sample size is equal to 12.5% of the original size. Thus, the number of samples in different sampling methods as shown in Fig. 17 is 3.76 times the original size for oversampling, 0.096 times the original size for undersampling, and the same

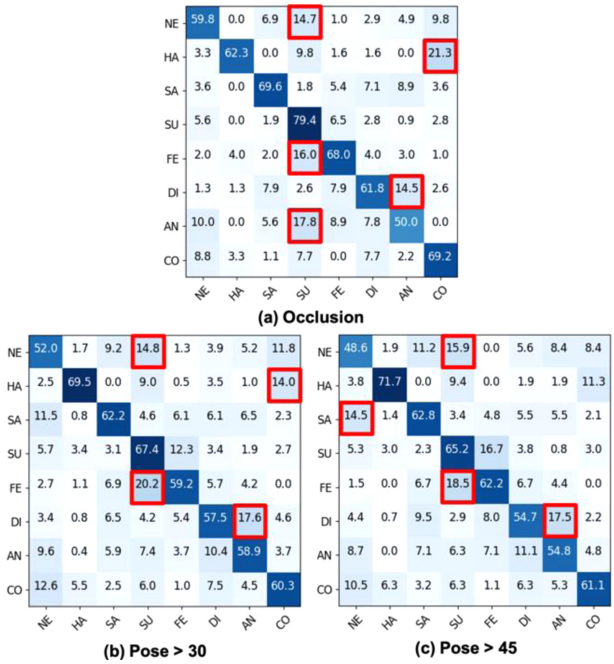

Fig. 14. Confusion matrices of our model on AffectNet-8 with occlusion, pose^{30+}, and pose^{45+}.

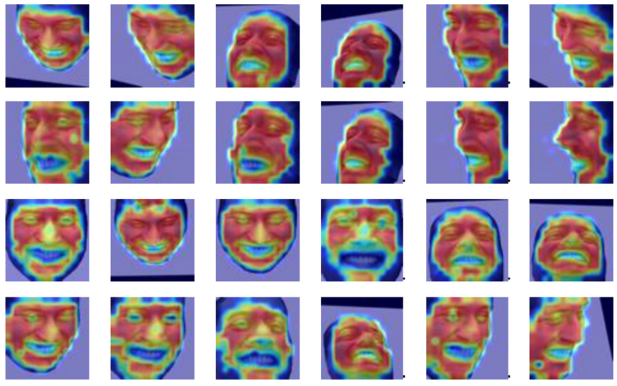

Fig. 15. Grad-CAM on various yaw and pitch angles of the same face.

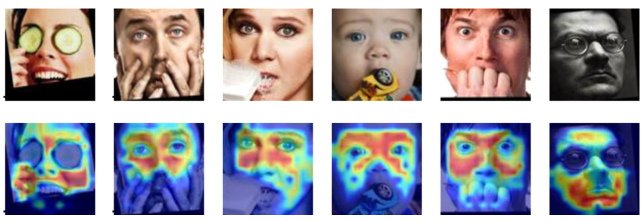

Fig. 16. Grad-CAM on occluded faces.

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

TABLE 9
Pose Consistency on FairFace-3D

Models	Yaw Angle							Pitch Angle					Total
	0	15	30	45	60	75	Std.	0	15	30	45	Std.	
DAN [48]	0.8590	0.8370	0.8210	0.7825	0.7631	0.6721	0.0672	0.8900	0.8404	0.8151	0.7592	0.0544	0.8191
PIDViT (Ours)	0.8718	0.8502	0.8256	0.7926	0.7445	0.6897	0.0686	0.8856	0.8472	0.8254	0.7784	0.0447	**0.8285**

as the original size for mix-sampling. As a result, the proposed method weighting cross-entropy loss verifies its superiority over all the sampling methods in solving the data imbalance issue, as shown in Table 10.

6.8 Ablation Study

As shown in Fig. 2, PIDViT is divided into two stages of training, where the student model uses a Visual Transformer (ViT) as the backbone and batch normalization (BN) is added after the MLP. To verify the effectiveness of the student model and observe the benefit of stage 1 training, we conducted ablation experiments, as shown in Table 11. Comparing PIDViT_a and PIDViT_b, we found that adding BN can greatly improve the accuracy of the model by 0.93%. Besides, comparing PIDViT_c and PIDViT_d, we found that using Kullback-Leibler (KL) divergence loss in stage 1 training is better than using cross-entropy loss (CE) loss, improving the accuracy by 0.26%. Finally, we also found that the stage 1 training can help the stage 2 training, improving the accuracy by 0.8% when comparing PIDViT_b and PIDViT_d.

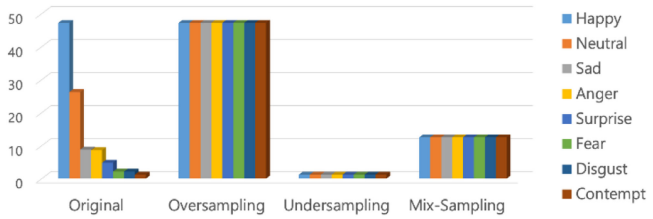

Fig. 17. Number of samples in different sampling methods.

TABLE 10
Data Imbalance Impact on PIDViT

Methods	AffectNet-8
Oversampling	0.6094
Undersampling	0.6149
Mix-sampling	0.6037
Weighting cross-entropy loss	0.6252

TABLE 11
Ablation Study on PIDViT

Models	AffectNet-8
PIDViT_a (only stage 2 without BN)	0.6079
PIDViT_b (only stage 2 with BN)	0.6172
PIDViT_c (stage 1 with CE + stage 2)	0.6226
PIDViT_d (stage 1 with KL + stage 2)	0.6252

7 Conclusion and Future Work

The PIDViT proposed in this paper is a pose-invariant distilled ViT model for facial expression recognition in the wild. First, for our study, face reconstruction technology with 5 steps was used to build a multi-pose face dataset called FairFace-3D from the original FairFace. FairFace-3D covers 10456 subjects spanning a wide of people, gender, ages, and races, and is used to train pose-invariance on the PIDViT. The training on the PIDViT can be divided into two stages. In stage 1, knowledge distillation using FairFace-3D is to train pose-invariance on the ViT to achieve the consistency of facial expressions between frontal faces and multi-pose faces. In stage 2, the facial expression recognition model is trained on target datasets; i.e., further training the student model on target datasets. The two-stage training can effectively reduce the impact of pose and occlusion on facial expression recognition. Finally, we present the experimental results of PIDViT on AffectNet, RAF-DB, and SFEW2.0, and validate its superiority over most state-of-the-art models. Besides, we also evaluate the impact of multi-pose faces using the new indicator "consistency", address the data imbalance issue on PIDViT, and finally conduct the ablation study on PIDViT.

In the future, two research issues can be extended and explored. The first is to consider compound emotions in facial expression recognition because faces could have more than one emotion. The second is to reduce the uncertainties of in the wild datasets where ambiguous facial expressions, low-quality facial images, and the subjectivity of annotators lead to poor accuracy.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. **Declarations: Availability of data and material:** Datasets related to this article can be found at the following URL link. FairFace-3D (2D images), FairFace-3D (3D modeling), and Labels: https://github.com/mayiprint/PIDViT-Create-FairFace3D-dataset, **Code availability:** https://github.com/mayiprint/PIDViT-Create-FairFace3D-dataset

References

[1] J. Almeida, L. Vilaça, I. Teixeira, and P. Viana, "Emotion identification in movies through facial expression recognition," *Appl. Sci.*, vol. 11, no. 15, Jul. 2021, Art. no. 6827.
[2] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-16-118, Jun. 2016.
[3] P. Antoniadis, P. P. Filntisis, and P. Maragos, "Exploiting emotional dependencies with graph convolutional networks for facial expression recognition," 2021, *arXiv:2106.03487v2*.

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

[4] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," 2021, *arXiv:2107.03107v4*.

[5] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2003, pp. 53–58.

[6] M. Batty and M. J. Taylor, "Early processing of the six basic facial emotional expressions," *Cogn. Brain Res.*, vol. 17, no. 3, pp. 613–620, Oct. 2003.

[7] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 535–541.

[8] C. Chen, C. Crivelli, O. G. B. Garrod, P. G. Schyns, J. M. Fernández-Dols, and R. E. Jack, "Distinct facial expressions represent pain and pleasure across cultures," *Proc. Nat. Acad. Sci. USA*, vol. 115, pp. E10 013–E10 021, 2018.

[9] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Inf. Sci.*, vol. 428, pp. 49–61, Feb. 2018.

[10] M. N. Dailey et al., "Evidence and a computational explanation of cultural differences in facial expression recognition," *Emotion*, vol. 10, no. 6, pp. 874–893, 2010.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[12] A. Dhall, R. Göcke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2106–2112.

[13] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929v2*.

[14] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[15] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Proc. Nebraska Symp. Motivation*, 1971, pp. 207–283.

[16] Y. Fan, V. Li, and J. C. K. Lam, "Facial expression recognition with deeply-supervised attention network," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 1–16, Third Quarter 2020.

[17] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2402–2411.

[18] B. Fasel, "Robust face analysis using convolutional neural networks," in *Proc. Int. Conf. Pattern Recognit.*, 2002, pp. 40–43.

[19] B. Fasel, "Head-pose invariant facial expression recognition using convolutional neural networks," in *Proc. IEEE 4th Int. Conf. Multimodal Interfaces*, 2002, pp. 529–534.

[20] Z. Fei et al., "Deep convolution network based emotion analysis towards mental health care," *Neurocomputing*, vol. 388, pp. 212–227, May 2020.

[21] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, Jun. 2013.

[22] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531v1*.

[24] Q. Huang, C. Huang, X. Wang, and F. Jiang, "Facial expression recognition with grid-wise attention and visual transformer," *Inf. Sci.*, vol. 580, pp. 35–54, Nov. 2021.

[25] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1031–1039.

[26] K. Kärkkäinen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age," 2019, *arXiv:1908.04913v1*.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980v9*.

[28] Y. H. Lai and S. H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 263–270.

[29] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2852–2861.

[30] P. Li, X. Wu, Y. Hu, R. He, and Z. Sun, "M2FPA: A multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10 043–10 051.

[31] Y. Li, G. Lu, J. Li, Z. Zhang, and D. Zhang, "Facial expression recognition in the wild using multi-level features and attention mechanisms," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.3031602.

[32] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu, "MViT: Mask vision transformer for facial expression recognition in the wild," 2021, *arXiv:2106.04520v2*.

[33] M. J. Lyons, S. Akamatsu, M. G. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. IEEE 3rd Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.

[34] F. Ma, B. Sun, and S. Li, "Robust facial expression recognition with convolutional visual transformers," 2021, *arXiv:2103.16854v2*.

[35] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Netw.*, vol. 16, no. 5-6, pp. 555–559, Jun./Jul. 2003.

[36] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, First Quarter 2019.

[37] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[38] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[39] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," 2021, *arXiv:2103.17107v3*.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. EEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[41] S. Sengupta, J. C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[42] J. Shi, S. Zhu, and Z. Liang, "Learning to amend facial expression representation via de-albino and affinity," 2021, *arXiv:2103.10189v3*.

[43] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 5800–5809.

[44] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10 347–10 357.

[45] T. H. Vo, G. S. Lee, H. J. Yang, and S. H. Kim, "Pyramid with super resolution for in the wild of facial recognition facial expression recognition," *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.

[46] C. Wang, K. Lu, J. Xue, and Y. Yan, "R-FENet: A region-based facial expression recognition method inspired by semantic information of action units," in *Proc. 1st Int. Workshop Hum.-Centric Multimedia Anal.*, 2020, pp. 43–51.

[47] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, Jan. 2020.

[48] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," 2021, *arXiv:2109.07270v3*.

[49] R. Wightman, "PyTorch image models," 2022. [Online]. Available: https://github.com/rwightman/pytorch-image-models

[50] S. Xie, H. Hu, and Y. Chen, "Facial expression recognition with two-branch disentangled generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2359–2371, Jun. 2021.

[51] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 211–216.

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.

[52] A. A. A. Youssif and W. A. A. Asker, "Automatic facial expression recognition system based on geometric and appearance features," *Comput. Inf. Sci.*, vol. 4, no. 2, pp. 115–124, Mar. 2011.
[53] Y. Zhang and C. Hua, "Driver fatigue recognition based on facial expression analysis using local binary patterns," *Optik*, vol. 126, no. 23, pp. 4501–4505, Dec. 2015.
[54] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[55] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3359–3368.
[56] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Geometry guided pose-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4445–4460, 2020.
[57] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "A unified deep model for joint facial expression recognition, face synthesis, and face alignment," *IEEE Trans. Image Process.*, vol. 29, pp. 6574–6589, 2020.
[58] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
[59] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 3510–3519.
[60] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," 2020, *arXiv:2005.10353v2*.

Yin-Fu Huang received the BS degree in computer science from National Chiao Tung University, in 1979, and the MS and PhD degrees in computer science from National Tsing Hua University, in 1984 and 1988, respectively. From July 1988 to July 1992, he was with the Chung Shan Institute of Science and Technology as an assistant researcher. He is currently a professor Emeritus with the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology. His research interests include database systems, multimedia systems, data mining, and data analytics.

Chia-Hsin Tsai received the BS degree in computer science from HungKuang University, in 2019, and the MS degree in computer science from the National Yunlin University of Science and Technology, in 2022. His major areas of interests include database systems, multimedia systems, and data mining.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on May 30,2024 at 03:51:43 UTC from IEEE Xplore. Restrictions apply.