Facial Chirality: From Visual Self-reflection to Robust Facial Feature Learning

Ling Lo, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng, *Senior Member, IEEE*

***Abstract*—As a fundamental vision task, facial expression recognition has made substantial progress recently. However, the recognition performance often degrades significantly in real-world scenarios due to the lack of robust facial features. In this paper, we propose an effective facial feature learning method that takes the advantage of facial chirality to discover the discriminative features for facial expression recognition. Most previous studies implicitly assume that human faces are symmetric. However, our work reveals that the facial asymmetric effect can be a crucial clue. Given a face image and its reflection without additional labels, we decouple the emotion-invariant facial features from the input image pair to better capture the emotion-related facial features. Moreover, as our model aligns emotion-related features of the image pair to enhance the recognition performance, the value of precise facial landmark alignment as a pre-processing step is reconsidered in this paper. Experiments demonstrate that the learned emotion-related features outperform the state of the art methods on several facial expression recognition benchmarks as well as real-world occlusion datasets, which manifests the effectiveness and robustness of the proposed model.**

***Index Terms*—facial expression, visual chirality, feature disentanglement, deep learning, vision transformer**

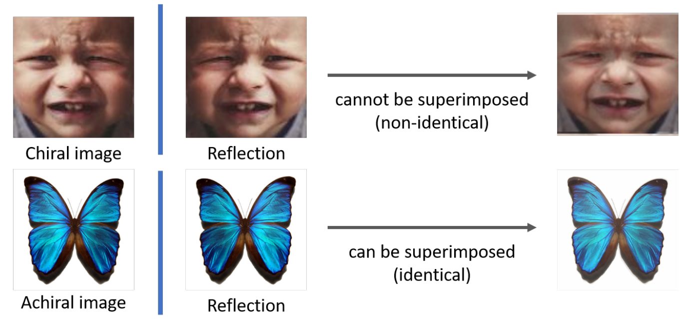


Fig. 1. An illustration of *facial chirality*. The human face is chiral and its horizontal reflection cannot be superimposed to make the same image, while an achiral object can be perfectly overlapped with its horizontal reflection.

I. Introduction

FACIAL expressions can serve as crucial clues in human-to-human communications and play an important role in human-machine interaction (HMI) systems. To capture the universal signal of human-being and interpret the emotional state of people, robust facial expression recognition (FER) has become vital and essential in HMI systems [1]–[13].

While researchers have made significant progress on automatic FER, most of the facial analysis works are explored with the assumption that a human face is a symmetric structure [14]–[18], *i.e.*, the left and the right halves of a human face are the same. The symmetric assumption implicitly implies that the features extracted from a facial image and its mirror-reflection are highly similar or even identical. However, faces are not so symmetrical as many think. In fact, psychological studies of facial asymmetry have revealed that the left and the right halves of the face differ in emotional attributes due to different dominant hemispheres of the brain [19], indicating that the faces are not horizontally symmetric under most circumstances. Meanwhile, chirality is a chemistry term used to describe the situation where two objects that appear to be similar are not symmetrical when folded over onto their own mirror-reflection. Especially, an emotional face is a concrete chiral object since the face should look similar to its reflection as they are with same facial expression, but cannot be overlapped onto each other since human faces are not horizontally symmetric. Previously in our work [20], we define the asymmetric characteristic of a face image and its reflection as *facial chirality*. Fig 1 shows an example of chiral object with chirality and achiral object without chirality.

Existing FER approaches did not take facial chirality into consideration, resulting in instability as the model may not recognize the expression of a well-trained sample after reflected due to facial asymmetry. Therefore, in this paper, we extend our previous work [20] and exploit the fact of *facial chirality* to discover discriminative features from the facial asymmetry for effective and robust expression feature learning. Since an image and its reflection are with identical emotion states, we believe that only the features they share are the crucial clues for FER. As a result, we propose a simple but effective feature disentanglement learning approach to distinguish the decisive emotion-related information from the emotion-invariant distortion using facial chirality. A Convolution Neural Network (CNN) feature extractor is first employed to generate feature representations. Then, the extracted representations are split along the channel dimension into emotion-related features and chirality-related features. Afterward, the split features are sliced into patches and processed along with a class token using a transformer encoder, whose self attention layers allow

L. Lo, H.-X. Xie are with the Institute of Electronics, National Yang Ming Chiao Tung Univresity, Hsinchu, 300 Taiwan.
E-mai: {linglo.ee08, hongxiaxie.ee08}@nycu.edu.tw.

H.-H. Shuai is with the Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung Univresity, Hsinchu, 300 Taiwan.
E-mail: hhshuai@nycu.edu.tw.

W.-H. Cheng is with the Institute of Electronics, National Yang Ming Chiao Tung Univresity, Hsinchu, 300 Taiwan, and the Artificial Intelligence and Data Science Program, National Chung Hsing University, Taichung, 400 Taiwan.
E-mail: whcheng@nycu.edu.tw.

the class token to interact with feature patches and learn useful information for FER. Since emotion-related features and chirality-related features are mutually exclusive, the model cannot learn both features simultaneously with the same set of attention weights of the class token and disentangle the features successfully. Consequently, we propose a new token, chirality token, to capture and process the emotion-invariant features separately and further eliminate the noise for the robustness of class token. Finally, the class token would be fed into the emotion classifier for the desired FER. Compared to our previous approach [20], in addition to separating the latent features into emotion-related and chirality-related subspaces, the hybrid model proposed in this paper further utilizes the self-attention mechanism in the transformer encoder and introduces attention weights to different feature patches. Moreover, the proposed chirality token and class token are updated independently through two sets of attention weights, preventing distortion from each other. As a result, the final classification of both expression and reflection could be more accurate and robust.

Our main contributions can be summarized as follows:

- We propose a hybrid model for feature disentanglement that explicitly considers *facial chirality* to discover discriminative facial features for facial expression recognition.
- Our model stands on the psychological basis of facial asymmetry, and the learned features are robust not only to mirror-reflection but also other real-world occlusion and pose variation. Additionally, we discuss the effectiveness of facial chirality and reconsider the value of a deliberate facial landmark alignment as pre-processing procedure of FER.
- Experiments illustrate that the proposed model can achieve accuracy of 0.9120 on RAF-DB and 0.6640 on Affectnet with a smaller model than other methods that achieve comparable results, demonstrating its effectiveness as well as efficiency.

The rest of this paper is organized as follows. Section II reviews the related works, and Section III presents the proposed method. Section IV presents the evaluations and conclusions are offered in Section V.

II. Related Work

A. Facial Expression Recognition

1) CNN-based FER: In the past decade, many efforts have been made to explore the potential of applying deep learning on FER task. The majority of them extract discriminative expression features by using different kinds of CNN-based models [20]–[27]. Based on psychological theories, Yang *et al.* [28] proposed the idea of an Emotion Circle, which includes emotion polarity, emotion type, emotion intensity, similarity, and additivity. Unlike most FER models supervised by the softmax loss, DACL [21] proposed a Deep Attentive Center Loss to improve intra-class compactness and inter-class separation in the embedding space.

2) Transformer-based FER: At the same time, the transformer-based FER method [29], [30] draws inspiration from the tremendous recent success of transformer in computer vision, *e.g.*ViT [31] and Swin Transformer [32]. Transformers can be utilized to model long dependencies between input sequences by the global self-attention mechanism. TransFER [29] modeled the relations between different facial parts via ViT, and proposed a multi-head self-attention dropping to learn relations among local patches. Visual Transformers with Feature Fusion (VTFF) [30] integrated LBP features and CNN features with the global-local attention and global self-attention for FER.

3) Robustness and efficiency in FER: Faces can suffer from occlusions, variations in the way the head is posed, motion blur, etc., which results in significant changes to their appearance [33], [34]. The FER algorithm therefore faces a great challenge with robustness [24], [30], [35]. RAN [24] adaptively adjusted the importance of the facial region to address the occlusion and pose-variant problems. Ma *et al.* [30] designed a global self-attention enables the network to learn the relationships between elements of visual feature sequences and ignore the information-deficient regions, in case of occlusions scenarios. Since deep models with excessive parameters and FLOPs are impractical for real-world applications, some works propose light-weight FER models [26], [36]. For example, EfficientFace [26] designed a light-weight FER architecture from the perspective of the feature extraction and label distribution learning strategy. However, there are relatively few FER works that consider both robustness and efficiency.

B. Facial Symmetric Learning

Most face related studies consider face as a symmetric structure [14], [15], [37]–[39]. For example, Zhang *et al*. [14] encoded facial symmetry to provide additional supervision for AU intensity estimation, where facial symmetry involves one face image and its horizontally-flipped image. Wu *et al*. [15] assumed the face structure is weakly symmetric, and proposed a symmetry probability map to measure the degree of symmetric uncertainty.

However, these previous works overlook the importance of differentiating the two sides of the face. The general observation in psychology is that emotional expressions are more intense on the left side of faces since the right cerebral hemisphere is dominant for the expression [19]. Lin *et al*. [40], for example, explored visual chirality in images of people and faces using a self-supervised learning approach, and showed that the chirality of faces can be predicted. This is the first work that takes a step forward to explore the robust facial feature learning based on *facial chirality*. However, there are few facial expression works extend their researches based on this hypothesis. Ling *et al.* [20] is the first research on FER in explicitly considering *facial chirality* to discover discriminative facial features. In this paper, we propose a hybrid model for learning facial chirality features to enhance the robustness of FER.

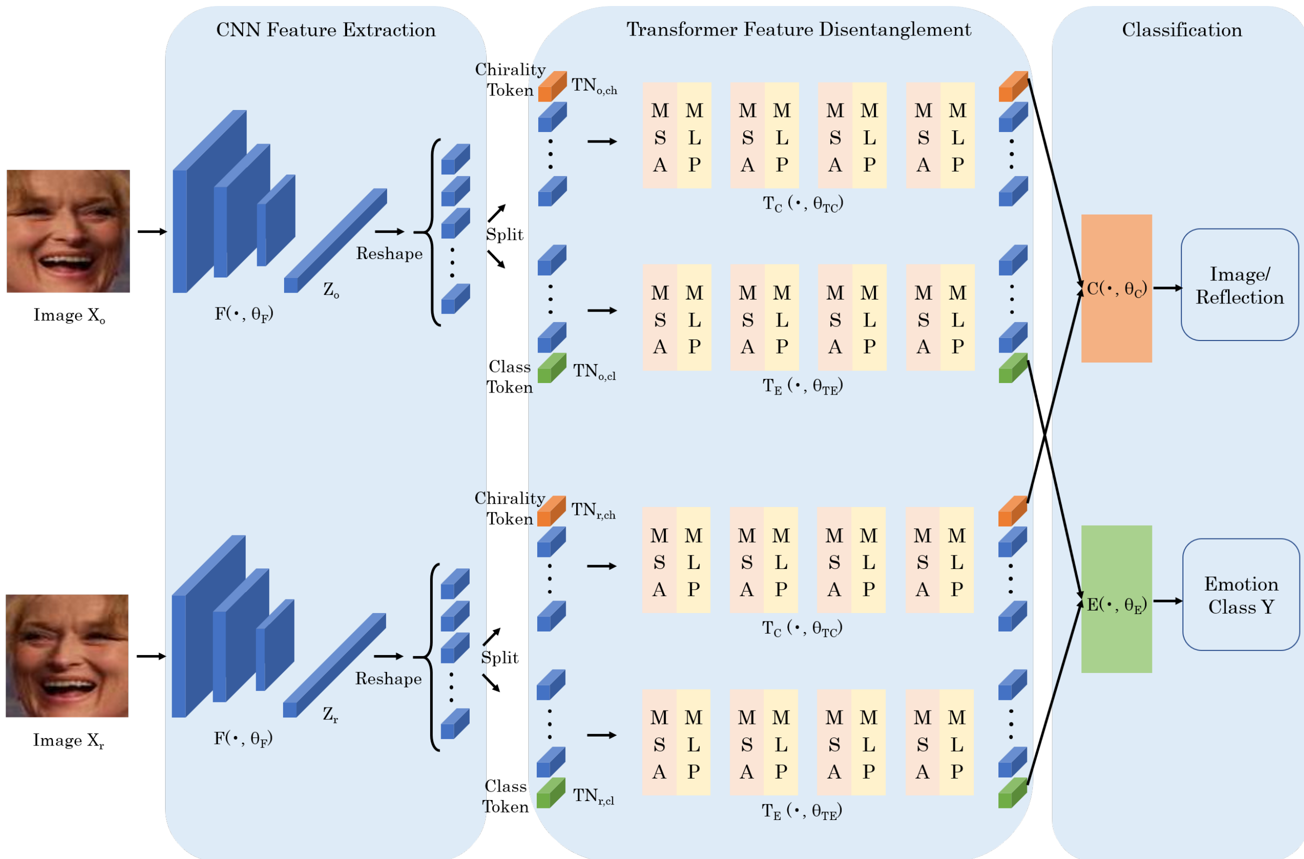


Fig. 2. Overall architecture of the proposed approach. (The figure is best viewed in color.)

III. Methodology

In this section, we introduce an efficient feature disentanglement learning approach to eliminate the bias caused by facial chirality and discover the decisive information from the noise for FER. Especially, we apply a hybrid model, which utilizes the merits of both CNN feature extraction and transformer encoders for efficient and robust feature learning. The multi-head self attention mechanism allows the class token to interact with the embedding patches extracted from a CNN feature map. Moreover, our proposed chirality token could utilize the mirror-reflection of an input image and decouple the chirality-related features for the class token so that it could learn the crucial emotion-related features for more robust FER.

A. Overview

As face alignment is typically required to align the facial landmarks and reduce the variation in face scale, a common assumption is that a more precise alignment can substantially enhance the FER performance. However, whether the performance between FER and the pre-processing landmark alignment is positively correlated remains unrevealed. On the other hand, our approach aims to directly align an image and its reflection at the emotional state level. With the simple aligned data provided in datasets, our model can boost the FER performance significantly compared to baseline with other deliberate landmark alignment approaches. More discussion and experiments are presented in Section IV.

Given an image X_o with facial expression label Y, its mirror-reflection X_r should share the same label Y because facial expressions are invariant to horizontal flipping. However, facial chirality would result in differences between the feature maps of the image pair since the face images are usually not horizontally symmetric. Here we argue that only the shared features of the image pair are the crucial clues for FER. Taking advantage of that, our model takes an image and its mirror-reflection as input and aims to decouple the identical emotion-related features from the non-identical chirality-related features for more robust expression classification. The proposed approach contains CNN feature extraction, transformer feature disentanglement, and two classification heads. Firstly, the input image pair X_o and X_r would be fed into the CNN feature extraction backbone $F(\cdot, \theta_F)$ to get the latent embeddings Z_o and Z_r. Then, they are both reshaped and separated into two sets of patches that are emotion-related and chirality-related, respectively. Afterward, the emotion-related sets are appended with the class tokens and processed by the class transformer encoders $T_E(\cdot, \theta_{T_E})$, while the chirality-related sets are appended with the proposed chirality token and fed into the chirality transformer encoders $T_C(\cdot, \theta_{T_C})$. In both of the transformer blocks, the class token and the chirality token would interact with the corresponding embedding patches through self attention and capture the important relations in between. To discover the distinguishing features of facial expressions, the class tokens of the image pair aim to learn the identical emotion-related features. On the other hand, we rely on the proposed chirality tokens to disentangle the non-identical features of the image pair from the original embeddings since they are considered as noise and would harm the robustness of the FER. Finally, the emotion classifier $E(\cdot, \theta_E)$ takes the learned class tokens as input to predict the emotion of the image pair and the chirality tokens would be fed into the chirality classifier to distinguish the image from its reflection. The overall architecture of our approach is shown in Fig 2 and more details are explained as follows.

B. CNN Feature Extraction

In previous works [31], experiments have shown that using raw image patches as transformer input takes a large amount of training data and time to converge as it has less inductive bias on small datasets than CNN. Therefore, instead of using raw image patches, we first apply a CNN feature extractor $F(\cdot, \theta_F)$ to capture both the global and local spatial information efficiently and generate the high-level latent embeddings $Z = F(X; \theta_F)$ of the image. The latent embeddings are then reshaped by an 1×1 convolution layer and sliced along the channel dimension into small patches, which can be further separated into two sets since we suppose the CNN extracted features should consist of two branches that are emotion-specific and chirality-specific, respectively. The features of the image pair are processed by a share-weights CNN model and separated in the same way for further feature disentanglement.

C. Transformer Feature Disentanglement

In addition to decoupling through feature separation, we further employ transformer encoders to accomplish efficient feature disentanglement. A transformer encoder is composed of Multi-head Self Attention layers (MSA) and Multi Layer Perceptron (MLP). It takes the input as a series of N patch tokens. Typically, a learnable class token would be appended to the series of input tokens for classification and standard learnable 1D position embeddings are added to inject position information.

After feature separation, the chirality-related patches would contain the perturbation noise information in the input pair due to facial chirality and thus should be wiped out. On the other hand, the emotion-related patches the input pair share are considered as the key to FER and thus can be utilized for effective classification. The emotion-related patches are appended with the class token to capture the decisive information for emotion classification. Since the reflection and the image share the same emotional state, the optimization goal of the transformer is to make the class token of the image and its reflection as similar as possible. On the other hand, we propose a new token, the chirality token, to interact with the chirality-related patches and enhance the robustness of the class token. The chirality token aims to learn the features the input pair is different from each other, identify the emotion-invariant features, and further decouple such noise for the class token to only learn from crucial features. Therefore, the target objective of a chirality token is to discriminate the reflection from the original image.

Next, emotion-related patches with the class token would be processed by the emotion transformer encoder $T_E(\cdot, \theta_{T_E})$. The MSA layer of the transformer allows the class token to interact with the emotion-related embedding patches and learn the features that the input pair are identical for further emotion classification. Likewise, the chirality token interacts with other patches through self attention mechanism similarly in the chirality transformer encoder $T_C(\cdot, \theta_{T_C})$ and aims to discover the chirality-related features that are non-identical for the input pair. As the output of the two transformer encoders, the class token and the chirality token can thus be viewed as the two disentangled features. During training, the model is encouraged to learn the expression-related information and the chirality-related information separately through the two tokens. The optimization goal is to minimize the distance between the class tokens and maximize the distance between chirality tokens of an image pair:

$$\underset{\theta_F, \theta_{Tch}, \theta_{Tcl}}{\arg\min} \quad \|(TN_{o,cl} - TN_{r,cl})\| - \|(TN_{o,ch} - TN_{r,ch})\|. \quad (1)$$

Specifically, since the input patches are initially extracted using a CNN backbone instead of raw images and the features are separated along the channel dimension, both the class token and the chirality token can be learned from high-level feature maps with global information instead of partial raw images patches after feature separation. Moreover, the proposed feature disentanglement approach allows the class token to identify the crucial features more efficiently as the emotion-invariant features are simultaneously learned and eliminated with the addition chirality token. In our experiment, we notice that using the chirality token can achieve better performance with a smaller model size compared to other state-of-the-art FER models based on transformer, which manifests the effectiveness and efficiency of our proposed model.

D. Reflection and Expression Classification

After being processed, the two tokens are then fed into two different branches for classification. The expression classifier $E(\cdot, \theta_E)$ takes the class tokens that contain the shared information of the image pair as input and predicts the corresponding expression. On the other hand, for more robust chirality-related feature learning, another classifier $C(\cdot, \theta_C)$ is applied to distinguish the reflection images from the original ones, enforcing the model to separate the subspace where the image and its reflection are far from each other.

E. Loss Function

The learning objective of the proposed model is to separate the latent features of an emotional image into emotion-related set and chirality-related set. For the latent feature space to be properly separated, we utilize four objective functions to guide our feature disentanglement model during training.

For the expression-related subspace, since FER can be considered as a classification problem, a cross-entropy loss L_{CE} between the ground truth expression label Y, and the two class tokens processed by the emotion transformer encoder,

$T_E(TN_{o,cl};\theta_{T_E})$ and $T_E(TN_{r,cl};\theta_{T_E})$, are included for the classification:

$$\begin{aligned} L_{CE} =& CE(E(T_E(TN_{o,cl};\theta_{T_E});\theta_E),Y)+ \\ & CE(E(T_E(TN_{r,cl};\theta_{T_E});\theta_E),Y), \end{aligned} \tag{2}$$

.

Given our goal of FER, the learned class token of the image pair should be identical since they have the same expression label. In contrast, the chirality token is designed to capture the information that the image pair are different from each other, so the chirality token of the image pair should be far distanced in the latent space to distinguish the reflection from the input image. Similar to contrastive learning [41], here we take the two class tokens of the input pair as the positive sample and the chirality tokens as the negative sample to obtain the contrastive loss:

$$\begin{aligned} L_{CON} =& d(T_E(TN_{o,cl};\theta_{T_E}),T_E(TN_{r,cl};\theta_{T_E}))+ \\ & max(0, m - d(T_C(TN_{o,ch};\theta_{T_C}),T_C(TN_{r,ch};\theta_{T_C}))), \end{aligned} \tag{3}$$

where m is the distance margin and d represents the Euclidean distance of the two tokens in the latent space. Training with the contrastive loss, the chirality tokens are optimized to be pulled away from each other to learn the emotion-invariant features while the class tokens are guided to be close to each other as they aim to extract chirality-invariant features that are identical for both image and its reflection.

Additionally, to ensure that chirality tokens can eliminate the noise for the class tokens rather than causing information loss, a binary cross entropy loss L_{BCE} on both $T_C(TN_{o,ch};\theta_{T_C})$ and $T_C(TN_{r,ch};\theta_{T_C})$ are used during training to constraint the binary classification learning of images and their reflections. The combined BCE loss can be written as:

$$\begin{aligned} L_{BCE} =& BCE(C(T_C(TN_{o,ch};\theta_{T_C});\theta_c),Y_o) \\ & +BCE(C(T_C(TN_{r,ch};\theta_{T_C});\theta_c),Y_r), \end{aligned} \tag{4}$$

and Y_o, Y_r are the labels for original images and their reflections.

The overall objective can be written as:

$$\min_{\theta_{en},\theta_c,\theta_e} L_{CE} + \lambda_1 L_{CON} + \lambda_2 L_{BCE}, \tag{5}$$

where the λ_1 and λ_2 are the weights for different objective terms.

IV. Experiment

A. Datasets

To evaluate the robustness and effectiveness of the proposed method, we focus on large-scale, in-the-wild problems since controlled datasets are limited and may not be directly applicable to real-world scenarios. Experiments are conducted on two in-the-wild datasets, along with subsets of occlusion and pose variation. **RAF-DB** contains 12,271 training samples and 3,068 testing samples in the single-label subset, including seven classes of basic emotions (surprise, fear, disgust, happiness, sadness, anger and neutral). **Affectnet** is one of the largest existing facial expression dataset. Since the labels of its testing data are not released, we follow the settings in [42] to use 280000 images for training and 3500 validation images (500 images per category) with seven emotion categories as in RAF-DB for testing.

TABLE I
ACCURACY COMPARISON OF EACH COMPONENT BEFORE THE EXPRESSION CLASSIFIER IN OUR MODEL.

Method	RAF-DB	AffectNet
CNN	0.8792	0.6137
CNN-FS	0.8801	0.6326
CNN-Trans	0.8921	0.6508
CNN-Trans-FS	0.8810	0.6417
CNN-FS-Trans	**0.9120**	**0.6640**

In addition, we examine our model based on the subsets of the RAF-DB and Affectnet collected in [43] to compare its performance to that of other FER approaches under the perturbation of real-world noise, such as occlusions and pose variations. The manually selected images from the validation set of AffectNet and the test set of RAF-DB form the subsets of Occlusion-AffectNet, Pose-45-AffectNet, Occlusion-RAF-DB, Pose-30-RAF-DB and Pose-45-RAF-DB. Specifically, the Pose-45-Affectnet and Pose-45-RAF-DB contain faces with an angle larger than $45°$ while the Pose-30-RAF-DB contains faces with an angle larger than $30°$.

B. Implementation Details

In our experiments, an IR-50 model [44] pretrained on ms-celeb-1M dataset [45] is used as the CNN feature extractor. For feature disentanglement, both the chirlaity and emotion transformer encoders are composed of 4 identical layers with 8 attention heads. The classifiers are implemented using stacked fully-connected layers for both chirality and expression, respectively. For RAF-DB, we train our model for 30 epochs using SGD optimizer with a learning rate of 1e-3 initially and then decayed by a factor of 10 at epoch 15. For Affectnet, due to the large amount of samples, we train our model for 3 epochs using SGD optimizer with a learning rate of 1e-3 initially and then decayed by a factor of 10 at epoch 2.

C. Component Analysis

We conduct experiments to analyze the feature learning effectiveness of each component before the classifier in our model in Table I. CNN means using CNN feature extractor only. CNN-Trans means applying a transformer encoder to process the CNN output without feature separation and the proposed chirality token. CNN-Trans-FS stands for performing feature separation on the class token after processed by the transformer without the chirality token for feature disentanglement while CNN-FS-Trans is the proposed method.

From the table some observation can be found: **(1)** Using single CNN feature extractor with feature separation for feature disentanglement also helps the model learn the useful features for better emotion classification. This result proves that the feature learning strategy of decoupling emotion-related features from chirality-related features can combined with other feature extraction methods and enhance model

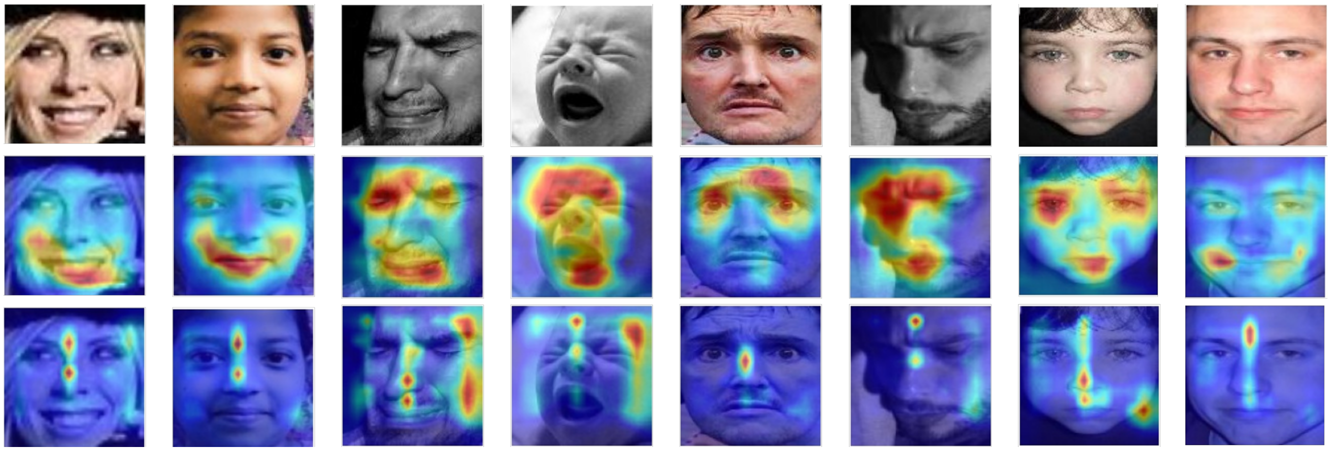


Fig. 3. The rollout attention maps of different emotions. The first row is the original image. Heatmaps of the emotion-variant features learned in the class token are in the second row while emotion-invariant features learned in the chirality token are presented in the third row.

TABLE II
ABLATION STUDY ON DIFFERENT LOSS TERMS OF OUR METHOD.

L_C	L_{CON}	L_{BCE}	Accuracy	
			RAF-DB	Affectnet
V	-	-	0.8967	0.6434
V	V	-	0.9038	0.6527
V	-	V	0.9035	0.6532
V	V	V	**0.9120**	**0.6640**

performances generally. **(2)** To eliminate the chirality bias, one of the possible approaches is to perform feature separation on the class token after patches are processed by the transformer encoder and optimize the model to learn the features separately through the disentanglement components of loss. Although the transformer encoder's self attention layers allow the class token to interact with every patch through learned attention weights, we argue that it is incompatible for the learned class token to obtain both emotion-related and chirality-related features for disentanglement with the same set of attention weights because the two features are mutually exclusive. Therefore, the model could not learn them separately, resulting in performance drop as shown in Table I. In contrast, in our approach we perform feature separation before the features are processed by the transformer encoder. With the proposed chirality token, the chirality-related features can be learned independently and be eliminated from the emotion-related features. Therefore, the decisive parts of the features for facial expression can interact with each other and more efficient feature learning are perform with a smaller model.

On the other hand, to examine the effectiveness of each term in the loss function, we compare the individual performance of the single crossentropy loss, the combination of the cross-entropy loss with the contrastive loss, the combination of the cross-entropy loss with the binary-cross-entropy loss, and the full objective functions. Table II shows the ablation study results on both RAF-DB and AffectNet datasets. With each additional loss term the overall accuracy is improved and the model perform best when all the loss terms are considered.

D. Attention and Visualization

To investigate the effectiveness of our approach, we show the feature heatmaps of representations learned from the class token and the chirality token by examining the attention rollout [46] of the tokens. Fig 3 illustrates the visualization results. The second row contains the attention maps of the class tokens learned from different emotions and the third row contains the emotion-invariant features learned in the chirality token. Firstly, the features are successfully disentangled as the two heatmaps of the same image are nearly non-overlapped. Moreover, we can observe that in the second row, image regions that are semantically relevant for emotion classification are obtained effectively in the class token globally. For example, the big smiles on the faces in the first two columns are accurately captured to recognize the face with happiness, while the model attends to the frowning eyebrows and the stretching mouths of the sad faces in the third and fourth columns. Even for the last two columns where the facial images are with no obvious expressions, the class token could still recognize the neutral expression with attention on the eye and the mouth regions. Meanwhile, the chirality tokens aim to decouple the emotion-irrelevant features for the class token. As shown in the third row of the figure, it pays more attention to the nose region and the edge of the images as they provide the least emotion information and could contain disturbing noise, harming the robustness of the expression recognition.

We further use t-sne [47] to visualize the features extracted by the baseline method (without the chirality token and feature disentanglement), and emotion-related features learned in the class token of our proposed method in Fig 4. We can observe that the features learned in our proposed method are more distinguishable as the intra-class similarity of each emotion and the inter-class differences are increased compared to the baseline method. The overall accuracy is then improved for 2%, indicating that our approaches can acquire more effective features for expression recognition.

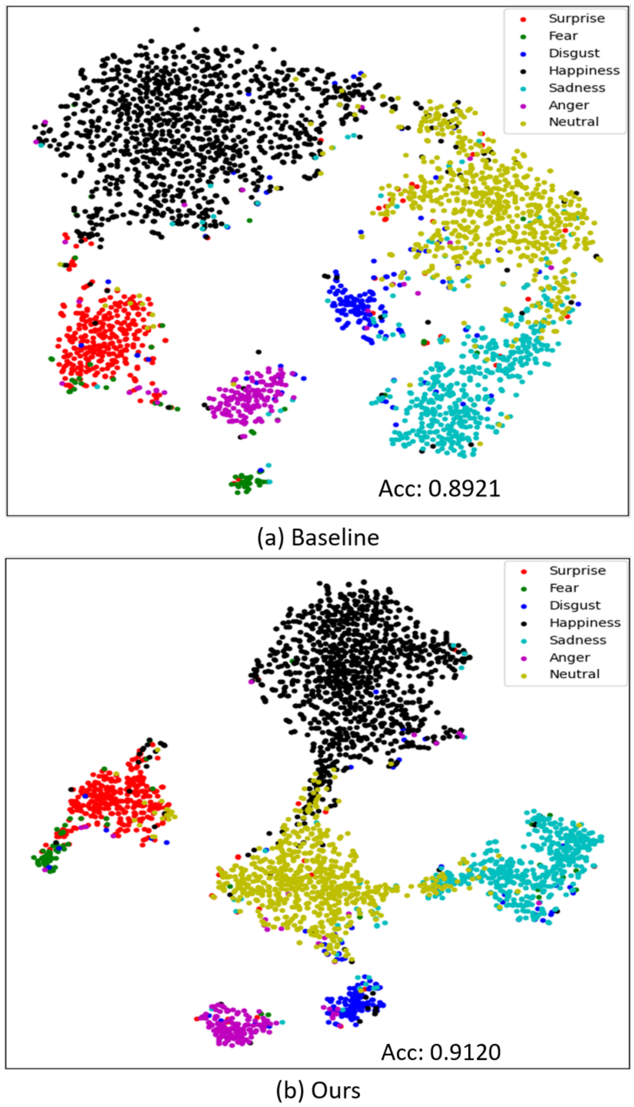


Fig. 4. Visualization of the expression features using t-SNE. Features are extracted from the RAF-DB database. The figure is best viewed in color.

E. Discussion on Robustness

Robustness to mirror-reflection image We conduct experiments to discuss the impact of facial chirality resulting in errors in existing FER approaches. To examine the robustness of several latest FER models towards the bias caused by facial chirality, we train the models[1] including EfficientFace [26], DACL [48], and TransFER [42] on the training set of both RAF-DB and AffectNet but test the models with the mirror-reflection of the testing images. It is noteworthy that as a common training strategy, random horizontal flipping is performed when training these models to generate more diverse data samples as a step of data augmentation. Moreover, since the transformer model applied on FER are not widely explored currently, we further conduct experiments on two of the latest proposed transformer architecture, ResT [49] and SwinT [32],

[1]For EfficientFace and DACL, we use the codes provided by the authors for training. For TransFER, we reimplement the model using parameters provided in the original paper.

TABLE III
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART APPROACHES ON THE MIRROR-REFLECTION IMAGES. RAF-DB-F AND AFFECTNET-F MEANS THE MODEL IS TESTED ON THE MIRROR-REFLECTION OF IMAGES IN THE ORIGINAL DATASETS.

Method	RAF-DB	RAF-DB-f	AffectNet	AffectNet-f
EfficientFace [26]	0.8835	0.8442	0.6370	0.6040
DACL [21]	0.8778	0.8507	0.6520	0.6152
TransFER [29]	0.9091	0.8979	0.6623	0.6474
ResT [49]	0.8677	0.8673	0.6011	0.5997
SwinT [32]	0.8980	0.8918	0.5953	0.5917
Ours	**0.9120**	**0.9120**	**0.6640**	**0.6640**

that have achieved great success on several classification tasks to further evaluate the robustness of our model to chirality bias with the propose chirality token.

The results in Table III illustrate that for the existing FER models, the performances will decrease up to 4% in terms of accuracy when evaluating on the mirror-reflection of the testing images. It indicates that performing flip augmentation during training can only increase the diversity of the training data instead of eliminating the chirality bias for the model. The fact that an image and its reflection share the same expression is overlooked by existing models as they could not capture the decisive emotion features of both original images their reflections for all the testing samples. Meanwhile, we observe that transformer architectures are more robust to images with horizontal flipping as ResT and SwinT only suffer from a slight performance drop. This could be attributed to the fact that, since the transformer-based models process image patches independently and learn the token representation, the noise induced by face asymmetric may be lessened. Our model, on the other hand, takes both image and its reflection as input and so the performance would not be affected by the chirality bias. In addition to robustness, the effectiveness of our model is proved as we achieve the highest accuracy whether the testing images are flipped or not.

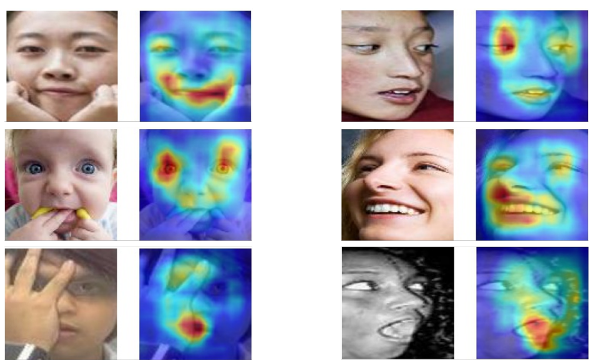
Robustness to pose and occlusion Since common noise exists in real-world scenarios, the robustness to occlusion and pose variation are very important when evaluating the quality of expression features extraction. We then perform experiments on datasets with real-world occlusion and pose variation. In addition to existing FER models focusing on robust feature learning, we also conduct experiments on the baseline model and other transformer-based classification approaches [32], [49] for comparison. Table IV and Table V are the comparison results in terms of accuracy. The performance of our proposed method is superior to compared approaches on the five subsets, proving the strength of our model against different real-world noises. Specifically, for occlusion-RAF-DB and occlusion-AffectNet, our model can achieve around 5% of accuracy increased compared to the existing FER approaches. Even for extreme pose variation with an angle larger than $45°$, our model can still hold over 90% accuracy on RAF-DB and over 60% on AffectNet, which outperforms existing approaches by a large margin. The experiment results illustrate that through feature disentanglement, our model can learn to eliminate the emotion-invariant features with the help of the chirality token

TABLE IV
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART APPROACHES ON THE RAF-DB OCCLUSION AND POSE VARIATION DATASETS.

Method	Occlusion	Pose-30	Pose-45
RAN [24]	0.8272	0.8604	0.8520
EfficientFace [26]	0.8324	0.8813	0.8692
TransFER [29]	0.8762	0.9054	0.8889
Baseline	0.8774	0.8934	0.8827
ResT [49]	0.8256	0.8701	0.8728
SwinT [32]	0.8433	0.8933	0.8961
Ours	**0.8816**	**0.9150**	**0.9086**

TABLE V
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART APPROACHES ON THE AFFECTNET OCCLUSION AND POSE VARIATION DATASETS.

Method	Occlusion	Pose-45
RAN [24]	0.5900	0.5386
EfficientFace [26]	0.5912	0.5563
TransFER [29]	0.5709	0.5382
Baseline	0.6240	0.6093
ResT [49]	0.6081	0.5597
SwinT [32]	0.5997	0.5417
Ours	**0.6453**	**0.6261**



(a) Occlusion (b) Pose Variation

Fig. 5. Rollout visualization of occlusion and pose variation.

and the learned emotion-related features can thus be more robust to other distortions.

To investigate how our model focus on emotion features against occlusion and pose variation, the attention rollout visualization of our model in images with hand occlusion and head pose variation are presented in Fig 5. Particularly in the third row of the occlusion example, where half of the faces are covered by a hand, our model can still capture the uncovered mouth region as crucial expression features and recognize the surprise emotion under severe occlusion. Our model is shown to be able to classify emotions effectively under complex and challenging scenarios by attending to discriminative regions.

F. Rethinking the impact of face alignment on FER

In the real-world scenario, the variation in face scale would also cause asymmetry and induces noises that are irrelevant to facial expressions. A standard pre-processing pipeline of current FER frameworks often includes locating a face and several facial landmarks, then aligning the face image. Traditionally, face alignment using the coordinates of localized landmarks can substantially enhance the FER performance since it reduces the in-plane rotation [50]. However, the effectiveness of a deliberate face alignment approach contributing to facial expression recognition remains unknown.

TABLE VI
PERFORMANCE COMPARISON AMONG DIFFERENT ALIGNMENT APPROACHES WITH CHIRALITY MODULE.

Method	RAF-DB	AffectNet
Baseline+Dlib	0.8035	0.6237
Baseline+OpenCV	0.7542	0.6057
Baseline+MTCNN	0.8299	0.6309
Baseline+RetinaFace	0.8227	0.6254
Baseline+chirality (Ours)	**0.9120**	**0.6640**

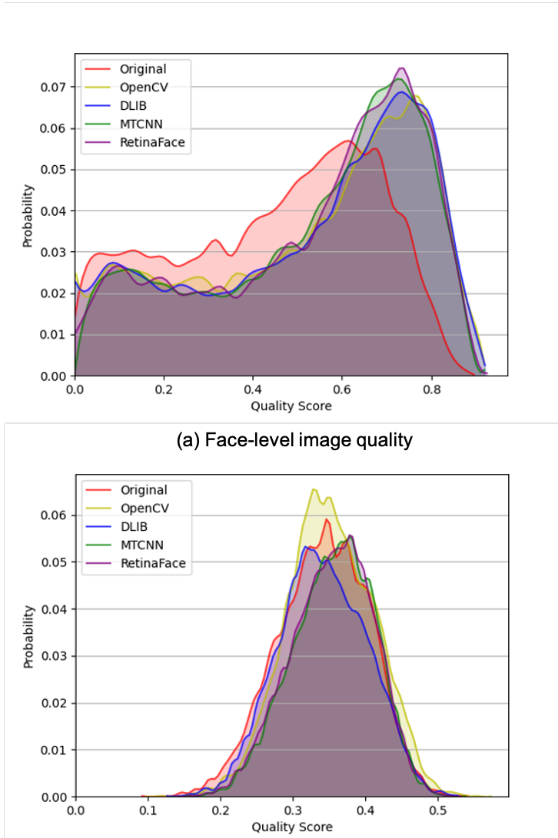
On the other hand, we argue that asymmetry noises induced by face-scale variation can be better mitigated by our proposed chirality model compared to deliberate face alignment strategies. In this section, we conduct experiments to investigate the relation between deliberate face alignment approaches as well as our approach and their resulting improvements on FER.

Specifically, we have four alignment strategies using the Deepface library[2], *i.e.*, DLIB, OpenCV, MTCNN, RetinaFace, which are commonly used in FER works. In contrast, we only use the simple aligned data provided by RAF-DB and AffectNet. Since different align algorithms may result in different image quality, which affects the accuracy of expression recognition, in order to eliminate the influence of this factor, we further computed the face-level quality [51] and image-level quality [52], as shown in Fig.6. The image-level quality curve among different align algorithms and original images show little difference. And the original images have lower quality than other four alignment images from the perspective of face-level. However, our proposed chirality module can still outperform other settings in RAF-DB and AffectNet datasets as shown in Tab.VI. The reason behind this is that instead of aligning face images based on landmarks, the FER model should focus more on facial chirality. In other words, without a deliberate alignment step, FER can recognize distinctive features of facial expressions due to our proposed chirality module.

G. Comparison with the State of the Art

The overall comparison of our model with state-of-the-art approaches in terms of model size and accuracy are shown in Table VII. The baseline model is a CNN feature extractor followed by a transformer encoder without the chirality token and feature disentanglement. Some observations can be found in this table: (1) For both RAF-DB and AffectNet, our model can achieve the best accuracy of 0.9120 and 0.6640, proving the effectiveness of the proposed method under different real-world scenarios. Especially, our proposed model can disentangle the features and capture the emotion-related feature patches. In addition, the multi-head self attention layers in the transformer encoder introduce two sets of attention weights for different feature patches contributing to the class and the chirality token respectively. The class token and the chirality

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



(a) Face-level image quality

(b) Image-level image quality

Fig. 6. Comparison of image quality between different alignment strategies on RAF-DB.

TABLE VII
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART APPROACHES ON THE RAF-DB AND AFFECTNET DATASET. † MEANS THE RESULTS ARE OBTAINED FROM THE CONFUSION MATRIX PROVIDED IN THE ORIGINAL PAPER.

Method	# Parameters	RAF-DB	AffectNet
RAN [24]	11.2M	0.8690	0.5297†
SCN [22]	11.9M	0.8703	-
ICME Chirality [20]	11.3M	0.8735	0.6189
DDA Loss [48]	11.2M	-	0.6234
EfficientFace [26]	1.3M	0.8836	0.6370†
DACL [21]	104.1 M	0.8778	0.6520
FDRL [25]	14.4 M	0.8947	-
DMUE [23]	69.9 M	0.8942	-
TransFER [29]	76.7 M	0.9091	0.6623
Baseline	44.7 M	0.8921	0.6508
ResT [49]	30.3 M	0.8980	0.6011
SwinT-B [32]	87.8 M	0.8677	0.5953
Ours	46.2M	**0.9120**	**0.6640**

token can then be learned independently and dynamically from different feature patches, improving the model performance in terms of accuracy. (2) Compared to TransFER, the other transformer model with accuracy over 90% on RAF-DB, our model can obtain better accuracy with a much smaller number of model parameters with the proposed chirality token. Similarly, for Affectnet, existing models with accuracy over 65% are all with large model sizes. Our model possesses almost half of the model size of DACL and TransFER while the accuracy is the highest among all. These results show the contribution of the proposed chirality token on effective feature learning as better accuracy can be achieved with a smaller model[3]. Overall, our proposed model can outperform state of the art approaches in terms of accuracy with a reasonable model size compared to other approaches with comparable performances.

V. CONCLUSION

In this paper, we aim to capture the facial expressions in complex or ambiguous scenarios especially considering facial chirality. Our proposed feature disentangling approach can be robust mirror-reflection as well as real-world occlusion and head pose variation. Our model can learn to be more aware of the eye and the mouth region, where most expressions occur, even for non-frontal faces, and further enhance the overall recognition accuracy. As a result, the proposed method outperforms state of the art approaches on RAF-DB and AffectNet datasets and further proves the robustness of the learned features and the efficiency of our model training. We hope that our work will attract new researchers to discover the potential of utilizing facial chirality.

ACKNOWLEDGMENT

This work was supported in part by Ministry of Science and Technology of Taiwan under the grant numbers: MOST-109-2223-E-009-002-MY3, MOST-110-2218-E-A49-018, MOST-111-2634-F-007-002 and MOST-109-2221-E-009-114-MY3.

REFERENCES

[1] H. Zhang, W. Su, J. Yu, and Z. Wang, "Weakly supervised local-global relation network for facial expression recognition," in *IJCAI*, 2020.
[2] L. Lu, L. Tavabi, and M. Soleymani, "Self-supervised learning for facial action unit recognition through temporal consistency," in *BMVC*, 2020.
[3] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *ACM Multimedia*, 2020, pp. 2871–2880.
[4] L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng, "MER-GCN: Micro expression recognition based on relation modeling with graph convolutional network," in *IEEE MIPR*, 2020, pp. 1–1.
[5] Z. Li, S. Han, A. S. Khan, J. Cai, Z. Meng, J. O'Reilly, and Y. Tong, "Pooling map adaptation in convolutional neural network for facial expression recognition," in *IEEE ICME*, 2019.
[6] G. Wen, T. Chang, H. Li, and L. Jiang, "Dynamic objectives learning for facial expression recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2914–2925, 2020.
[7] F. Zhang, M. Xu, and C. Xu, "Weakly-supervised facial expression recognition in the wild with noisy data," *IEEE Transactions on Multimedia*, 2021.
[8] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 211–220, 2018.
[9] S. Lin, M. Bai, F. Liu, L. Shen, and Y. Zhou, "Orthogonalization-guided feature fusion network for multimodal 2d+ 3d facial expression recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1581–1591, 2020.
[10] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang, and H. Yu, "Realistic facial expression reconstruction for vr hmd users," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 730–743, 2019.

[3]Note that although EfficientFace seems to with an extremely small model, an additional CNN backbone as label distribution generator is required during training, making the overall model parameters higher than reported in the table.

[11] H.-H. Lu, S.-E. Weng, Y.-F. Yen, H.-H. Shuai, and W.-H. Cheng, "Face-based voice conversion: Learning the voice behind a face," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 496–505.

[12] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "An overview of facial micro-expression analysis: Data, methodology and challenge," *IEEE Transactions on Affective Computing*, 2022.

[13] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–41, 2021.

[14] Y. Zhang, W. Dong, B.-G. Hu, and Q. Ji, "Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation," in *IEEE CVPR*, 2018.

[15] S. Wu, C. Rupprecht, and A. Vedaldi, "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild," in *IEEE CVPR*, 2020.

[16] S. Huang, W.-H. Cheng, and R. Cheng, "Single patch based 3d high-fidelity mask face anti-spoofing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 842–845.

[17] Z. Hu, H. Xie, L. Yu, X. Gao, Z. Shang, and Y. Zhang, "Dynamic-aware federated learning for face forgery video detection," *ACM Transactions on Intelligent Systems and Technology*, 2022.

[18] J. Li, H. Xie, L. Yu, X. Gao, and Y. Zhang, "Discriminative feature mining based on frequency information and metric learning for face forgery detection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[19] W. R. Powell and J. A. Schirillo, "Asymmetrical facial expressions in portraits and hemispheric laterality: A literature review," *Laterality*, vol. 14, no. 6, pp. 545–572, 2009.

[20] L. Lo, H. X. Xie, H.-H. Shuai, and W.-H. Cheng, "Facial chirality: Using self-face reflection to learn discriminative features for facial expression recognition," in *IEEE ICME*, 2021, pp. 1–6.

[21] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *IEEE WACV*, 2021, pp. 2402–2411.

[22] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *IEEE CVPR*, June 2020.

[23] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *IEEE CVPR*, 2021, pp. 6248–6257.

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

[25] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *IEEE CVPR*, 2021, pp. 7660–7669.

[26] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *AAAI*, vol. 35, no. 4, 2021, pp. 3510–3519.

[27] J. Sanchez-Riera, K.-L. Hua, Y.-S. Hsiao, T. Lim, S. C. Hidayati, and W.-H. Cheng, "A comparative study of data fusion for rgb-d based visual recognition," *Pattern Recognition Letters*, vol. 73, pp. 1–6, 2016.

[28] J. Yang, J. Li, L. Li, X. Wang, and X. Gao, "A circular-structured representation for visual emotion distribution learning," in *IEEE CVPR*, 2021, pp. 4237–4246.

[29] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *IEEE ICCV*, 2021, pp. 3601–3610.

[30] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, 2021.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE ICCV*, 2021, pp. 10 012–10 022.

[33] C.-W. Yang, T. H. Phung, H.-H. Shuai, and W.-H. Cheng, "Mask or non-mask? robust face mask detector via triplet-consistency representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 1s, pp. 1–20, 2022.

[34] C.-Y. Chen, L. Lo, P.-J. Huang, H.-H. Shuai, and W.-H. Cheng, "Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 809–13 818.

[35] K.-Y. Lung, C.-R. Chang, S.-E. Weng, H.-S. Lin, H.-H. Shuai, and W.-H. Cheng, "ROSNet: Robust one-stage network for ct lesion detection," *Pattern Recognition Letters*, vol. 144, pp. 82–88, 2021.

[36] P. Barros, N. Churamani, and A. Sciutti, "The FaceChannel: A light-weight deep neural network for facial expression recognition." in *IEEE FG*, 2020, pp. 652–656.

[37] E. Saber and A. M. Tekalp, "Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions," *Pattern Recognition Letters*, vol. 19, no. 8, pp. 669–680, 1998.

[38] X. Li, G. Hu, J. Zhu, W. Zuo, M. Wang, and L. Zhang, "Learning symmetry consistent deep cnns for face completion," *IEEE Transactions on Image Processing*, vol. 29, pp. 7641–7655, 2020.

[39] F. S. Abousaleh, T. Lim, W.-H. Cheng, N.-H. Yu, M. A. Hossain, and M. F. Alhamid, "A novel comparative deep learning framework for facial age estimation," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 1, p. 47, 2016.

[40] Z. Lin, J. Sun, A. Davis, and N. Snavely, "Visual chirality," in *IEEE CVPR*, 2020, pp. 12 295–12 303.

[41] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE CVPR*, vol. 1, 2005, pp. 539–546.

[42] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *IEEE ICCV*, October 2021, pp. 3601–3610.

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

[44] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE CVPR*, 2019, pp. 4690–4699.

[45] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*. Springer, 2016, pp. 87–102.

[46] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *ACL*, 2020.

[47] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[48] A. H. Farzaneh and X. Qi, "Discriminant distribution-agnostic loss for facial expression recognition in the wild," in *IEEE CVPR Workshops*, 2020, pp. 406–407.

[49] Q. Zhang and Y.-B. Yang, "Rest: An efficient transformer for visual recognition," in *NeurIPS*, vol. 34, 2021.

[50] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[51] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness," in *IEEE CVPR*, 2020, pp. 5650–5659.

[52] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE CVPR*, June 2020.

Ling Lo received the B.S. degree from Department of Electronics Engineering, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan, R.O.C., in 2019, and now she is pursuing a Ph.D degree in Institute of Electronics, NYCU. Her current research interests include deep learning and computer vision. Recently her work focuses specifically on facial and micro-expression recognition.

Hong-Xia Xie received the B.S. degree in Internet of Things from the Zhengzhou University of Aeronautics in 2016 and received the M.S. degree in communication and information systems, Fujian Normal University, China, in 2019. She is now pursuing a Ph.D. degree in Institute of of Electronics, National Yang Ming Chiao Tung Univresity, Taiwan. Her research interests include emotion recognition and deep learning.

Hong-Han Shuai received the B.S. degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, R.O.C., in 2007, the M.S. degree in computer science from NTU in 2009, and the Ph.D. degree from Graduate Institute of Communication Engineering, NTU, in 2015. He is now an Associate Professor in NCTU. His research interests are in the area of multimedia processing, machine learning, social network analysis, and data mining. His works have appeared in top-tier conferences such as MM, CVPR, AAAI, KDD, WWW, ICDM, CIKM and VLDB, and top-tier journals such as TKDE, TMM, TNNLS and JIOT. Moreover, he has served as the PC member for international conferences including MM, AAAI, IJCAI, WWW, and the invited reviewer for journals including TKDE, TMM, JVCI and JIOT.

Wen-Huang Cheng is Distinguished Professor with the Institute of Electronics, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan. He is also Jointly Appointed Professor with the Artificial Intelligence and Data Science Program, National Chung Hsing University (NCHU), Taichung, Taiwan. Before joining NYCU, he led the Multimedia Computing Research Group at the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan, from 2010 to 2018. His current research interests include multimedia, artificial intelligence, computer vision, and machine learning. He has actively participated in international events and played important leading roles in prestigious journals and conferences and professional organizations, like Associate Editor for IEEE Transactions on Multimedia, General co-chair for IEEE ICME (2022) and ACM ICMR (2021), Chair for IEEE MSA technical committee, governing board member for IAPR. He has received numerous research and service awards, including the Best Paper Award of 2021 IEEE ICME, the Outstanding Associate Editor Award of IEEE Transactions on Multimedia, the 2018 MSRA Collaborative Research Award, the 2017 Ta-Yu Wu Memorial Award from Taiwan's Ministry of Science and Technology (the highest national research honor for young Taiwanese researchers under age 42), the Top 10% Paper Award from the 2015 IEEE MMSP, and the K. T. Li Young Researcher Award from the ACM Taipei/Taiwan Chapter in 2014. He is IEEE Distinguished Lecturer, ACM Distinguished Member, IET Fellow and BCS Fellow.