Pose-Aware Facial Expression Recognition Assisted by Text Description

Although expression descriptions provide additional information about facial behaviors and pose features are beneficial to facial expression recognition (FER), neither has been fully leveraged. This paper proposes a pose-aware text-assisted method of facial expression recognition using cross-modality attention. Specifically, the proposed method consists of three components: the pose feature extractor, the text feature extractor, and the cross-modality module. The pose feature extractor cooperates with a fully connected layer so poses can be clearly discriminated and classified to represent corresponding poses. Cluster centers are taken as the final pose features to eliminate bias due to appearance and illumination. The text feature extractor is based on a transformer. Expression description texts are first passed through Intra-Exp attention to obtain expression embeddings. All expression embeddings are then concatenated and passed through Inter-Exp attention to leverage correlations among different expressions. The cross-modality module attempts to learn attention maps that distinguish the importance of different regions by taking advantage of prior knowledge about poses and expressions for the extracted facial image feature. Pose and expression classification utilize the weighted attention maps. Experiments on two benchmark data sets demonstrate that the proposed method achieves superior performance over state-of-the-art methods.

# Introduction



Facial expression is one of the most powerful, natural, and universal ways in which human beings convey their emotional states and intentions (Darwin 2015; Tian, Kanade, and Cohn 2001). Studies on facial expression analysis have been applied to social robots, medical treatments, driver fatigue surveillance, and human-computer Interaction systems (S. Li and Deng 2020). As expression recognition applications migrate to more flexible and mutable scenarios in the wild, the demand for multi-pose facial expression recognition has increased. This is the focus of our paper.

Two methods are commonly used for multi-pose FER. One intuitive approach is to enrich and enlarge the data set by generating fresh images with different poses and expressions for each original image, so that the network can learn various patterns (F. Zhang et al. 2018; F. Zhang et al. 2020; X. Zhang, Zhanga, and Xu 2021). A discriminator is used to force similarity between real and generated images. It determines whether an image is an original image (TRUE) or a generated image (FALSE). Once the discriminator is unable to determine whether a generated image is TRUE or FALSE, it is added to the original images and used to train a robust FER system. The pose and expression information directing image generation is usually offered through pairs of label or face landmarks. However, the approach is heavily dependent on the quality of generated images, and is inefficient due to the large amount of training data needed.

The other approaches learn pose-robust features through either normalization, adversary, disentanglement, or subspace methods. For normalization-based methods (Y.-H. Lai and Lai 2018; Jampour and Moin 2021; Jampour et al. 2017; F. Zhang, Xu, and Xu 2021), pose normalization is used at the image level to generate the frontal facial image or at the feature level by directly transforming to the frontal face. In other words, the original image is mapped into the feature space under the frontal face. Adversary-based methods (C. Wang, Wang, and Liang 2019) use a pose discriminator to distinguish the pose from the feature extracted by encoder. The objective is to defeat the discriminator so the coded feature hardly contains any pose information. The disentanglement method (Ruan et al. 2020) supposes that pose and expression information involved in a facial image can be segmented explicitly. This implies that expression-related features would be immune to pose information, ensuring an explicit classification boundary. For subspace-based methods (Eleftheriadis, Rudovic, and Pantic 2014; T. Zhang et al. 2016; Liu et al. 2021), expression feature spaces corresponding to different poses share a common subspace. The goal is to find the shared subspace and map the original facial image to it in order to obtain expression-related features. These methods learn pose-robust expression features. However, in real-life scenarios, pose and expression information are coupled with each other in an extremely complex nonlinear mode. Zhu et al. (Zhu and Ji 2006) proved that in 2D images, rigid facial changes due to head pose and non-rigid facial changes due to expression are nonlinearly coupled and therefore challenging to deal with using these methods. Our method takes advantage of the pose features for simultaneous pose and expression classification, so extra images don’t need to be generated.

Most recent FER works focus on attention mechanisms. The attention-based approaches attempt to assign different weights to different regions to represent correlations between region and expression. The methods use a deep neural network (DNN) to automatically learn facial image features, and a homogeneous-modality attention map that yields the final features weighted by the attention map (Marrero Fernandez et al. 2019). Depending on the degree, the attention map can function on a patch level, channel level, and pixel level. For the first case (K. Wang et al. 2020), the facial image is cropped into several patches, as in manual partition based on human intuition. Regions such as eye side and mouth corner are always cropped individually. For the second case (H. Yang et al. 2021), the extracted feature contains the channel dimension. Different channels have different influences and must be handled accordingly. For the third case (J. Li et al. 2020), the feature and attention maps are directly aligned with each other through the fully connected layer, and the attention map directly acts on pixels of the feature map. The attention mechanism improves FER under the frontal face. However, performance drops significantly when there are multiple poses. In addition, attention map generation lacks explicit guidance, leading to slower convergence. The proposed method utilizes the semantic information found in expression texts to focus on the crucial facial regions through cross-modality attention. This extends the attention mechanism to multi-pose states.

We propose a pose-aware text-assisted FER method to solve the weaknesses of previous works. Our method makes use of the pose features and the expression texts. Training occurs in two stages. First, the pose feature extractor is trained for pose classification in cooperation with a fully connected layer, so the pose features can be represented as latent vectors. Next, the head pose features are fused with the expression text embeddings extracted by the text feature extractor. Inspired by (H. Yang et al. 2021), cross-modality attention is used to calculate attention maps for all pairs of head pose feature and embedded expression texts. Lastly, multi-task learning is implemented; the facial image feature weighted by the attention maps is passed through two fully connected layers for simultaneous pose and expression classification. In general, head pose features and expression texts are taken as prior knowledge and pose and expression information coexist. The pose features allow the model to automatically adapt to pose variety. Since the expression texts contain stronger semantic information, the attention maps have better Interpretability. Figure [figure1] illustrates an example of surprise. Although the facial image is not frontal, the expression text description still includes important facial regions for recognizing surprise, like the eyebrows and lips.

In summary, the main contributions of our work are as follows:

1. We propose a novel method for effective FER. The proposed method automatically adapts to pose variety instead of forcibly eliminating or segmenting out poses. We apply a multi-task learning method for simultaneous pose and expression classification.

2. We introduce cross-modality attention to generate attention maps. Expression texts contain strong semantic information that can guide the model to focus on crucial regions. Therefore, cross-modality attention is better able to explain why a facial image is judged to have a certain expression. To the best of our knowledge, expression texts derived from action unit (AU) descriptions (Ekman, Friesen, and Ellsworth 2013) have not previously been considered for FER.

3. The proposed method is evaluated on the Multi-PIE and BU-3DFE data sets. Experimental results show that our proposed method outperforms several state-of-the-art methods, verifying the effectiveness of leveraging pose awareness and expression texts to improve FER.

# Related Work

This section discusses existing generation-based, pose-robustness-based, and attention-based FER methods, as they are closely related to our proposed method.

## Generation-Based FER

Zhang et al*.* (F. Zhang et al. 2018) propose a joint pose and expression modeling method. The identity feature extracted from the encoder and the coded pose and expression are passed through the decoder to synthesize facial images with different expressions under arbitrary poses. Ultimately, the method expands the training set to tens of times of the original. Zhang et al. (F. Zhang et al. 2020) improve the above method by guiding the generator using facial landmarks rather than pose and expression code, as landmarks contain more intuitive and detailed guidance. The performance of the classifier largely depends on the quality of the training set. While the generated images greatly enrich and enlarge the training set, they are typically quite different from the original images, reducing the efficiency of the training model.

## Pose-Robustness-Based FER

Lai et al. (Y.-H. Lai and Lai 2018) propose an emotion-preserving representation learning method, which uses a generative adversarial network (GAN) to convert non-frontal face images into frontal face images while preserving the expression characteristics. The feature extracted by the generator is used for FER. Jampour et al. (Jampour and Moin 2021) propose a pose-invariant face frontalization method to learn mapping functions between frontal and non-frontal faces’ coefficients. Then spare coding is exploited to synthesize the frontalized faces. However, the quality of the generated image decreases significantly as the pose difference increases due to stretching artifacts. In addition to image level, pose normalization is used at the feature level. Jampour et al. (Jampour et al. 2017) propose a kernel-based pose-specific nonlinear mapping method to map the features extracted from the input image to corresponding features of the face with the same expression seen in a frontal view. Zhang et al. (F. Zhang, Xu, and Xu 2021) propose a pose modeling network to adaptively capture the discrepancy of facial images under different head poses in a deep representation. The discrepancy can guide the model as it converts features from non-frontal to frontal. However, most of these methods require paired images during training.

Wang et al. (C. Wang, Wang, and Liang 2019) propose an adversarial feature learning method to address pose variations. Their method extracts pose-invariant expression features by fooling a pose discriminator. When the discriminator cannot determine the pose of the input image from the encoded feature, the feature does not contain much pose information.

For the method based on disentanglement, Ruan et al. (Ruan et al. 2020) propose a deep disturbance-disentangled learning method for FER. The method takes advantage of multi-task learning and adversarial transfer learning to simultaneously and explicitly disentangle multiple disturbing factors. Then the expression subnetwork adopts a multilevel attention mechanism to extract expression-specific features.

For the method based on subspace, Eleftheriadis et al. (Eleftheriadis, Rudovic, and Pantic 2014) propose a discriminative shared Gaussian process latent variable model. In the model, a discriminative manifold shared by multiple poses of a facial expression is learned prior to performing FER in the expression manifold. Zhang et al. (T. Zhang et al. 2016) propose a DNN-driven feature learning method. The scale invariant feature transform (SIFT) descriptors are extracted from a set of landmark points. Then the SIFT feature matrix is sent to a well-designed DNN model to learn optimal discriminative features for expression classification. Liu et al*.* (Liu et al. 2021) propose a dynamic multichannel metric learning network. The method aims to map the original feature into an embedding feature space so that features with the same expression form clusters, while those with different expressions are farther apart. However, the common subspace may not exist when the pose range is too large.

Even pose-robust methods have difficulty due to the complex nonlinear coupling of pose and expression in 2D images.

## Attention-Based FER

*Wang et al.* (K. Wang et al. 2020) propose a region attention network (RAN) to adaptively capture the importance of facial regions for pose variant FER. Inspired by the fact that facial expressions are primarily defined by AUs, the method crops the original image according to different regions. The proposed region-biased loss encourages high attention weights for more important regions. Yang et al. (H. Yang et al. 2021) propose to exploit semantic embedding and visual features (SEV-Net) for AU detection. This method uses channel dimension to extract the visual feature. Attention maps take advantage of semantic embedding, acting on the visual features to weigh channels based on relevance. Li et al*.* (J. Li et al. 2020) propose an end-to-end network with an attention mechanism for automatic FER. The local binary pattern (LBP) descriptors first extract image texture information and catch the small movements of the face. Then LBP features and the attention mechanism are combined to create the attention map for the visual feature. The weighted visual features are more discriminative for FER. Attention mechanisms work well for the frontal face, but when poses vary, the performance of these method drops. Additionally, the lack of explicit information guidance slows convergence, since the optimization direction of the model is arbitrary at the beginning.

# PROBLEM STATEMENT

Let denote the training set, where represents a training facial image, represents the ground truth expression label, and represents the pose label. Let denote the expression texts. The goal is to extract the facial image feature weighted by attention maps while simultaneously learning two classifiers for expression and pose recognition. Let denote the testing set. Given an unseen sample , the weighted facial image feature is extracted and the expression classifier is applied to predict its label.

# METHODOLOGY



Figure [figure2] illustrates the framework of our proposed approach. The framework consists of three components. The pose feature extractor cooperates with a fully connected layer to discriminate the input facial image’s pose. After several accurate iterations, all pose features can be obtained through the extractor. The feature for each pose is the cluster center of the features corresponding to the pose. The text feature extractor is made up of an Intra-EXP encoder and an Inter-EXP encoder. The Intra-EXP encoder transforms expression description text into embeddings using a self-attention mechanism. Then the Inter-EXP encoder extracts more cognitive expression embeddings by exploring the correlation among different expressions. The cross-modality module first extracts the facial image feature through DNN. It combines pairs of head pose features and expression text features to calculate attention maps through cross-modality attention. Specifically, each pair is first added to each of the others, and then cross-modality attention calculates normalized cosines for the facial image feature and all pairs as attention maps. The facial image features weighted by the attention maps are used for simultaneous pose and expression classification.

## Pose Feature Extractor



It is simple to determine an input facial image pose if samples of every pose are available. However, precise descriptions of facial images are difficult to create. Pose features are reflected at the pixel level and lack semantic information. Due to the powerful representation of DNNs, pose features can be extracted after full training. Figure [figure3] shows the pose features mapped from a high dimension by t-distributed stochastic neighbor embedding (t-SNE). Features with the same pose tend to form clusters, and those with different poses are far apart.

To maximally eliminate bias due to appearance and illumination, the cluster centers (marked as pentagrams with different colors) are calculated to represent the final pose features.

## Facial Expression Texts



Facial expressions can be described through text. These expression texts depict the actions of various facial regions, such as brows, eyes, nose, and mouth.

However, although we are able to recognize facial expressions (typically through our visual nervous system), images are rarely accompanied by expression texts, although that might make the predicted expression label more convincing. Inspired by action units (AUs), which are more refined discrete emotion descriptors, we summarize expression texts by observing facial expressions and merging corresponding AU descriptions. For example, surprise means AU1, AU2, AU5, and AU26 are activated; happiness means AU6 and AU12 are activated; and fear means AU1, AU2, AU4, AU5, AU7, AU20, and AU26 are activated. Table [table1] shows common expression texts appearing in almost all facial expression data sets.

## Text Feature Extractor

**Input Embeddings.** Prior to the preliminary extraction of expression embeddings via the Intra-EXP attention mechanism, expression texts where m denotes the number of expression texts, need to be converted into equal-length numerical sequences. First, ‘CLS’ and ‘SEP’ are added to the beginning and end of all texts respectively. Then the texts are tokenized to numerical sequences according to words coded in the vocabulary. Later the sequences are padded with zero to the same length, recorded as . Since the relative positions among individual tokens in a sequence are crucial for semantic information, position code for is added. After embedding, the text can be converted to as follows:

**Intra-EXP Attention** After conversion from the original text, text embeddings rarely contain correlations among individual tokens. Transformer models (Vaswani et al. 2017) can learn any existing correlations. Therefore, each text embedding is encoded using Intra-EXP attention, a multi-layer transformer encoder. Each layer of the Intra-EXP attention is the same as the vanilla transformer encoder layer. Let be the input layer; the encoder feature at the (l+1)-th layer can be obtained from the l-th layer, which is formulated as follows:

where is the multi-headed self-attention module that attaches each token to the other tokens with appropriate weights, is the layer-norm function to ensure the stability of the feature distribution, and is the feed-forward sub-layer consisting of two fully connected layers and a ReLU activation function, formulated as follows:

Like the other methods, the first part of the encoder feature at the output layer is taken as the preliminary embedding of the corresponding expression text.

**Inter-EXP Attention** Preliminary embeddings for all expression texts have been obtained through the Intra-EXP attention, but correlations among different expressions can still be leveraged. For example, when either surprise or anger appears, the upper eyelid (AU5) is activated. The Inter-EXP attention is introduced to learn these correlations. It is structured like the Intra-EXP attention, but regards each preliminary embedding as a word. The embeddings are connected to form a new sequence. Final expression embeddings are obtained through the Inter-EXP attention and recorded as .

## Cross-Modality Module

In this section, the facial image features, the pose features, and the expression text embeddings are used to calculate the attention maps, and the facial image feature weighted by the attention maps is utilized for pose and expression classification.

Specifically, the facial image feature is obtained from the input image through DNN. Then, pairs of pose feature and expression text embedding are added with each other. Therefore, each fused feature represents a pattern that contains both pose and expression information. Let be the expression text embeddings and be the pose features; fused features are obtained as follows:

.

To take advantage of the pose features and expression text embeddings, the facial image feature and all fused features are used to calculate cosine matrices, which are converted to attention maps through a ReLU activation function and Norm function. Let DNN be E and the facial image be x; the attention map (AM) is formulated as follows:

where and c, w, and h denote channel, width, and height of the facial image feature respectively. , c denotes the length of the fused feature. Therefore, is formulated as:

where indicates the sum operation along with the channel dimension and means matrix multiplication, so .

To ensure the stability of the feature distribution, a Norm function is formulated as follows:

Lastly, the facial image features weighted by the attention maps are merged together as follows:

The final feature is sent to the pose classifier and expression classifier for simultaneous pose and expression classification. The total loss function is defined as:

where and represent the pose and expression classifiers. The ground truths of pose and expression are represented by and . represents the cross-entropy function.

Our method is able to automatically adapt to pose variety due to the introduction of pose features and expression texts. Expression texts containing strong semantic information direct the model to focus on the regions causing certain expressions in a cross-modality mode.

# EXPERIMENTS

To demonstrate the effectiveness of the proposed method, we conduct experiments on two public pose-related facial expression databases: Multi-PIE (Gross et al. 2010), the public multi-pose facial expression data set; and BU-3DFE (Yin et al. 2006), the 3D facial expression data set. The details are as follows.

## Experimental Conditions

Multi-PIE contains 755,370 images from 337 subjects under 15 viewpoints. The facial images in the data set are annotated with six expressions: disgust (DI), neutral (NE), scream (SC), smile (SM), squint (SQ), or surprise (SU). In our experiments, we follow the two settings used in Wang et al. (C. Wang, Wang, and Liang 2019). The first setting uses 7,095 images of 129 subjects under five viewpoints (,, and ). Subjects are randomly divided into a training set with 103 subjects and a testing set of 26 subjects. For the second setting, we use 6,174 images of 147 subjects under seven viewpoints (,,,,, , and ). Subjects are randomly divided into a training set of 118 subjects and a testing set of 29 subjects.

The BU-3DFE is a 3D-model data set containing 100 subjects including 56 females and 44 males. The facial images synthesized from 3D models are annotated with seven expressions: anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), surprise (SU), or neutral (NE). Each of the six prototypical expressions (excluding NE) includes four levels of intensity. In our experiments, we use the six prototypical expressions and follow three settings (C. Wang, Wang, and Liang 2019; F. Zhang et al. 2020). The first setting uses 12,000 images under five viewpoints (, , , , and ). The second setting uses 16,000 images under seven viewpoints (, , , and ). The third setting uses 21,000 images under 35 viewpoints, including seven pan angles (, , , and ) and five tile angles (, , and ). In this case, only the fourth intensity level is used. The subjects are randomly divided into a training set with 80 subjects and a testing set with 20 subjects, so there is subject independence.

## Implementation Details

First, face parts are recognized using OpenCV. Then images are cropped and resized to pixels for the Multi-PIE database and pixels for the BU-3DFE database. All networks are implemented by PyTorch. The pose feature extractor contains Resnet50’s layers, apart from the fourth residual layer and the last fully connected layer. The Intra-Exp attention and the Inter-Exp attention adopt the multi-layer transformer. We set the number of layers to twelve, the hidden size to 768, and the number of heads to twelve. The Intra-Exp encoder is initialized with a pre-trained parameter and frozen during training. The parameter of the Inter-Exp encoder is initialized randomly and updated based on the backpropagation of loss. The length of sentence embedding is 512 and the size of the final expression embedding is 768, which is then projected to 1024 through a linear layer. The facial image feature extractor in the cross-modality module has the same structure as the pose feature extractor, which is used as the baseline and followed by an expression classifier. Prior to the classification layer, the weighted features are summed together and flattened to a vector. For all experiments, the batch size is set to 32 and the learning rate is set to 5e-5. Classification accuracy score is used as the performance metric.

## Experimental Results and Analyses

First, the effectiveness of each component of the proposed method is examined. As shown in Table [table2], we analyze four different combinations: only baseline (B), baseline cooperating with pose features (BP), baseline cooperating with text features (BT), and baseline cooperating with pose and text features (BPT). From the table, we can draw the following conclusions:

BP and BT achieve better results than baseline, which just conducts FER using Resnet50. Specifically, BP outperforms the baseline by 1.56% on the Multi-PIE data set with pan angles and by 6.00% on the BU-3DFE data set with pan angles. These results indicate that simultaneously learning the expression feature and pose information prevents the model from regarding facial deformation due to pose as ingredients of the expression. Similarly, BT outperforms the baseline by 1.70% on the Multi-PIE data set with pan angles and by 5.34% on the BU-3DFE data set with pan angles. These results indicate that expression texts are capable of directing the model to focus on those crucial regions.

In addition, BPT outperforms BP and BT in all cases except on the BU-3DFE data set with pan angles. BPT outperforms BP by 0.99% and BT by 0.41% on the Multi-PIE data set with pan angles, and by 1.91% and 1.95% on the BU-3DFE data set with pan angles. These results suggest that the use of prior pose features and the guidance of expression texts can improve the accuracy of FER.



We analyze our method’s ability to adapt to different poses using the FER accuracy on the Multi-PIE data set with pan angles and the BU-3DFE data set with pan angles, since these poses vary widely. In Tables [table4] and [table5], the rightmost column indicates the average accuracy for different poses, the bottom row indicates the average accuracy for different expressions, and the lower right corner cell indicates the average overall accuracy.

For the Multi-PIE data set, three expressions (surprise, smile, and scream) achieve higher recognition accuracy due to distinct facial muscle formations. Neutral also achieves high accuracy, likely because it is dissimilar to other expressions. The model is most accurate when the angle is , although it is still highly accurate as the angle approaches . This indicates that the method is extremely pose robust. The confusion matrix on the data set is shown in Figure [figure4-a], and shows that the main error classification comes from confusion between disgust and squint. Both of these expressions behave similarly around the eyes; 9.36% of squint samples are misclassified as disgust and 6.90% of disgust samples are misclassified as squint. The expression texts also show strong semantic correlations between these two expressions.



Surprise and happiness achieve the higher recognition accuracy on the BU-3DFE data set. Fear is the most difficult expression to recognize, with 69.75% accuracy. As on the Multi-PIE data set, the accuracy among various poses is similar and the maximum value is obtained at . The confusion matrix on the data set is shown in Figure [figure4-b]. Here, the main error classification comes from confusion between sadness and anger. Of the sadness samples, 19.00% are misclassified as anger; 8.00% of anger samples are misclassified as sadness.





## Comparison to Related Works

Our method is compared to current state-of-the-art methods as summarized in Table [table3]. These methods can be divided into four categories: generation-based methods, normalization-based methods, adversary-based methods, and subspace-based methods. We don’t compare our method to disentanglement-based and attention-based methods, as the authors do not provide results on the aforementioned data sets and the codes are not available.

For generation-based methods, the proposed method outperforms Zhang et al. 2018 (F. Zhang et al. 2018) and Zhang et al*.* 2020 (F. Zhang et al. 2020). There is a certain deviation in the latent space between the generative and original images, which is likely fitted by the model. However, generation-based methods rely on large numbers of images for training, which decreases the learning efficiency. Our method uses expression texts to focus on the crucial regions. It also uses the pose features as prior knowledge to adapt to pose variety.

The proposed method is superior to normalization-based methods by Jampour et al. 2017 (Jampour et al. 2017), Lai et al. 2018 (Y.-H. Lai and Lai 2018), and Zhang et al*.* 2021 (F. Zhang, Xu, and Xu 2021). These methods attempt to either reconstruct a corresponding frontal facial image through GAN or transform the feature to the state of the frontal face. They require paired images, and the deformation of poses in 2D images yields a poor fit at the image and feature levels. Our method introduces pose features. Although the final feature contains both pose and expression information, distortion of expression information caused by forcibly changing pose information is avoided.

The proposed method also obtains superior results compared to adversary-based method Wang et al. 2019 (C. Wang, Wang, and Liang 2019). This method aims to extract pose-invariant features. This is challenging since in 2D images, pose and expression couple with each other in a complex nonlinear manner.

Furthermore, the proposed method achieves better performance than subspace-based methods such as Eleftheriadis et al. 2014 (Eleftheriadis, Rudovic, and Pantic 2014), Zhang et al. 2016 (T. Zhang et al. 2016), and Liu et al. 2021 (Liu et al. 2021). These methods first extract hand-crafted features and then map them to a common subspace. Our method employs an end-to-end training approach, ensuring that the model eventually reaches a consistent optimum output.



# CONCLUSION

This paper proposes a facial expression recognition method that leverages pose features and expression texts for pose and expression classification. Specifically, the training is divided into two steps. First, the pose feature extractor obtains the pose features as prior knowledge. Next, attention maps are generated using the pose and expression text features obtained by the text feature extractor. The facial image features weighted by the attention maps are utilized for pose and classification. Rather than using homogeneous-modality attention, we adopt cross-modality attention to make full use of semantic information from expression texts. Instead of eliminating pose influence, our method preserves pose information and recognizes pose and expression simultaneously.