A Brain-Inspired Way of Reducing the Network Complexity via Concept-Regularized Coding for Emotion Recognition

Han Lu[1], **Xiahai Zhuang**[2], **Qiang Luo**[1, 3*]

[1]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University
[2]School of Data Science, Fudan University
[3]Research Institute of Intelligent Complex Systems, Fudan University
{hlu20, zxh, qluo}@fudan.edu.cn

Abstract

The human brain can effortlessly and reliably perceive emotions, whereas existing facial emotion recognition (FER) methods suffer from drawbacks such as complex model structures, high storage requirements, and poor interpretability. Inspired by the role of emotion concepts in visual perception coding within the human brain, we propose a dual-pathway framework emulating the neural computation of emotion recognition. Specifically, these two pathways are designed to model the representation of emotion concepts in the brain and the visual perception process, respectively. For the former, we adopt a disentangled approach to extract emotion concepts from complex facial geometric attributes; for the latter, we employ an emotional confidence evaluation strategy to determine which concept is optimal for regularizing the perceptual coding. The proposed concept-regularized coding strategy endows the framework with flexibility and interpretability, as well as good performances on several benchmarking FER datasets.

Introduction

The recent proliferation of large-scale models (e.g., GPT-4) has swept across the entire deep learning community. These large-scale models have also achieved impressive performance in vision tasks (Ding et al. 2023; Wang et al. 2023). However, they suffer from issues such as prohibitively high memory and computational costs, as well as poor interpretability. For human beings, vision perception tasks can be conducted rapidly and seemingly effortlessly (Schiller 1995; Freeman and Simoncelli 2011; DiCarlo, Zoccolan, and Rust 2012). This remarkable ability implies the essential collaboration among multiple brain regions and the visual cortex during visual perception (Lumer and Rees 1999; Cela-Conde et al. 2004; Lee, Yeung, and Barense 2012). In this work, we take emotion recognition as an example to exploit the biological plausibility of the human brain to build up intelligent systems with reduced complexity and biological interpretability.

Facial emotion recognition (FER) aims to discern and interpret emotions displayed on other people's faces. In recent years, FER has received increasing interest in the deep learning community, as it holds promise for various applications, such as human-computer interaction and mental health assessment. Computer scientists have attempted to endow computers with the capability of FER by devising different algorithms, such as CNN (Li, Deng, and Du 2017a; Pons and Masip 2017; Zeng, Shan, and Chen 2018) and Transformer (Li et al. 2021a; Xue, Wang, and Guo 2021), and have made remarkable progress in this field. The current techniques have demonstrated that increasing the network size is not the only way to improve FER performance. It is natural to ask whether it is possible to design brain-inspired FER algorithms based on the neural computation of emotion recognition in the human brain.

It has long been believed that the neural representations of emotion concepts are formed and maintained in the high-level brain regions, such as the prefrontal cortex, the amygdala, and the hippocampus (Rolls 2023; Camacho et al. 2023). However, recent evidence suggests that the brain's visual pathway also encodes the emotion concepts to facilitate efficient emotion recognition (Brooks and Freeman 2018; Kragel et al. 2019; Brooks et al. 2019), which can be both fast and energy-saving. Therefore, the concept-regularized coding in the visual pathway provides a brain-inspired way of elevating the efficiency and reducing the complexity of the artificial neural network for emotion recognition.

In light of the concept-regularized coding, we proposed a brain-inspired network for emotion recognition with reduced complexity and enhanced interpretability. As depicted in Figure 1, our framework contains two pathways, named the conceptual pathway and the perceptual pathway. The former models the representation of emotional concepts in high-level brain regions. The latter employs a simple CNN to model the process of visual perception, as considerable evidence indicates that CNNs currently offer the best quantitative models of the hierarchical response patterns within the visual system (Kriegeskorte 2015; Zhuang et al. 2021; Kanwisher, Khosla, and Dobs 2023). The conceptual pathway shapes the visual emotion perception in the perceptual pathway by providing emotional concepts. To the best of our knowledge, we are the first to utilize concept-regularized coding in the brain to advance the FER algorithms. It is worth noting that, when modeling the conceptual pathway, we adopted a disentangled form to extract abstract emotional concepts (captured by emotion features in our model) from

*Corresponding author.

intricate and diverse facial attributes, thus avoiding the influence of confounding factors on the perceptual pathway. Meanwhile, an adaptive module for importance evaluation is implemented in this pathway to evaluate the emotional confidence of the emotion features. The acquisition of such confidence is crucial for associating the conceptual pathway with the perceptual pathway. One way to provide emotion concepts to the perceptual pathway is through knowledge distillation. After training the conceptual pathway, a facial emotion image is inputted into both pathways to simultaneously obtain emotion features and perceptual features. Subsequently, we can distill knowledge from emotion features to perceptual features. Although this distillation method can successfully transfer knowledge from the conceptual pathway to the perceptual pathway, it overlooks the relative stability of the human brain in representing emotion concepts. To avoid the potential disruption of visual encoding in the perceptual pathway caused by high-ambiguity emotion features, we employed a confidence evaluation strategy. If the emotional confidence of an image's emotion feature is high, we use this emotion feature to guide perceptual encoding. Otherwise, we use emotion features from other images with higher confidence in the same emotion category to guide the encoding in the perceptual pathway. The effectiveness of this approach can be observed in the experiment section.

Overall, our contributions can be summarized as follows,

- We design a biologically interpretable and lightweight FER algorithm, drawing inspiration from how emotion concepts are represented in the brain and their guidance in visual perception encoding.
- In our framework, we adopt a disentangled approach to establish the abstract emotion concepts, which are independent of facial geometric features.
- We employ a confidence evaluation strategy to provide the perceptual pathway with stable and reliable emotion concepts, thereby avoiding interference from confounding factors.

Related Works

Facial Expression Recognition

Recently, an increasing number of FER methods have been proposed, driven by the advancements in deep neural networks and the availability of large-scale FER datasets. In our opinion, these deep methods designed for FER can be roughly grouped into two broad categories: CNN-based methods and Transformer-based methods.

CNN-based methods may incorporate attention mechanisms (Li et al. 2018, 2020; Wang et al. 2020c) or consider the issue of uncertainty (Wang et al. 2020a; She et al. 2021). With the attention mechanism, multiple branches can be set in the model, each taking different facial regions as input. The model then adaptively learns the attention weights of each part to capture the importance of various facial regions in making the recognition. When considering the annotation or emotion ambiguity, the models can dynamically learn the ambiguity of each sample. Subsequently, the impact of samples with higher ambiguity on the loss function can be reduced, thereby facilitating the learning of discriminative features.

Transformer-based methods typically divide the input original image or the embedded representation of the original image into different patches. By learning the correlations between each patch and other patches, discriminative features can be obtained (Li et al. 2021a; Xue, Wang, and Guo 2021). However, the increase in computational cost of Transformer-based methods does not necessarily correspond to a proportional improvement in performance. Moreover, the biological interpretability in visual processing using Transformers is far less advanced compared to CNNs, which exhibit similar computational mechanisms to the human visual cortex. Therefore, in this paper, we did not extensively explore Transformers.

Feature Regularization

Regularization techniques (Srivastava et al. 2014; DeVries and Taylor 2017) have gained widespread popularity for training deep neural networks. These techniques aim to prevent overfitting and enhance the generalization performance of the models. Knowledge distillation (Hinton, Vinyals, and Dean 2015) can be viewed as a kind of feature regularization. It involves transferring knowledge from a more complex teacher model to a simpler student model by minimizing the discrepancy between their intermediate features or predicted logits (Heo et al. 2019; Park et al. 2019; Zhao et al. 2022). This process helps the student model learn informative representations from the teacher model. However, if knowledge is distilled directly from the teacher model to the student model without any quality control, it may introduce noise to the student model. In this paper, we propose a novel feature regularization technique using emotional confidence evaluation to guide the visual encoding of facial emotions, leading to improved performance in a simple network simulating the visual cortex.

Approach

Conceptual Pathway

Establish emotion concepts from facial expressions of emotion. Inspired by the neuroscience studies (Haxby, Hoffman, and Gobbini 2000; Zhang et al. 2023) on two distinct neuroanatomical pathways of the human brain in processing changeable (e.g., facial expression) and invariant characteristics (e.g., identity, age and sex) of a face, we develop two branches that relate to emotion and non-emotion encoding respectively. Specifically, we use the emotion encoder E_{emo} to model the emotion conceptualization of emotion-related regions like the prefontal cortex, amygdala and hippocampus. Meanwhile, we use the non-emotion encoder E_{non} to capture the confounding factors of a face, such as age, gender and identity, in the other branch. Given the i-th facial image, the emotion feature extracted by E_{emo} is denoted as $\boldsymbol{f}_{emo_i} \in \boldsymbol{R}^P$, where P is the dimension of the emotion feature. Similarly, we can also obtain the non-emotion feature $\boldsymbol{f}_{non_i} \in \boldsymbol{R}^P$ by E_{non}. The emotion feature is expected to solely reflect emotion

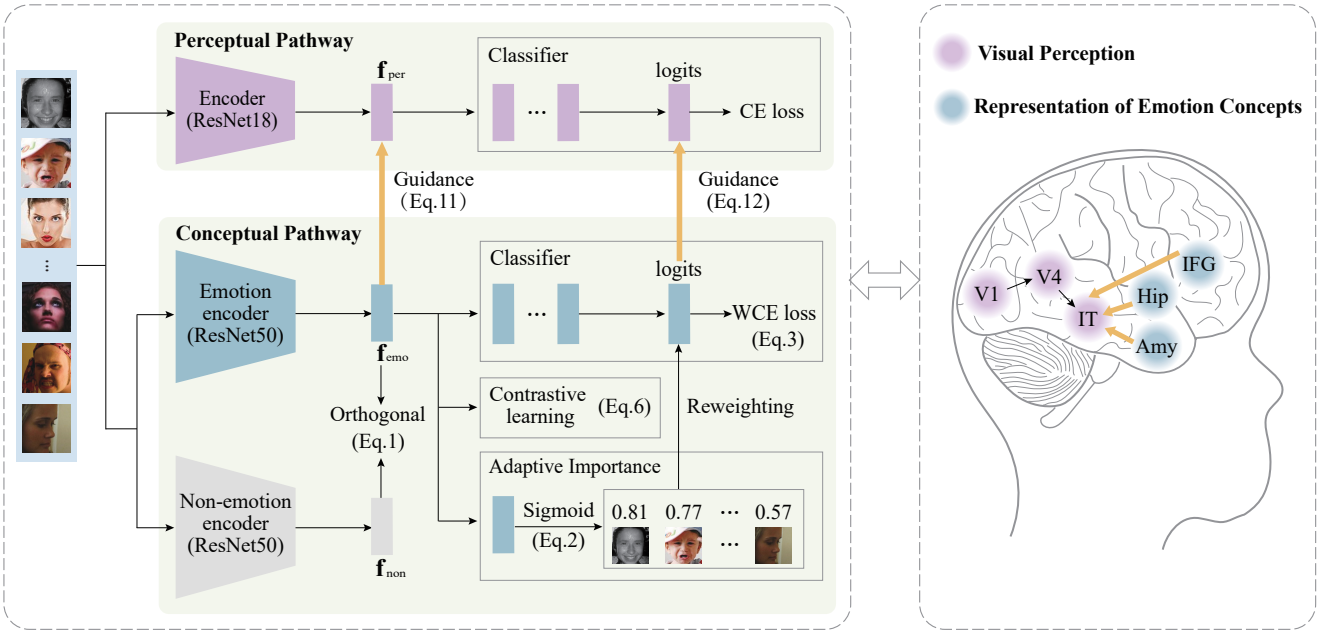

Figure 1: Overview of our proposed framework and its underlying biological model. Our brain-inspired ANN consists of three parts: 1) the conceptual pathway, which models the representation of emotion concepts in emotion-related brain regions in a disentangled manner; 2) the perceptual pathway, which perceives facial emotion information similar to that employed by the ventral visual stream (*i.e.,* from primary visual cortex (V1) to mid-level visual areas (e.g., V4) and to inferior temporal (IT) cortex); and 3) the feature regularization from the conceptual domain to the perceptual domain mimics the involvement of emotion concepts in perceiving visual emotion. Amy, amygdala; Hip, hippocampus; IFG, inferior frontal gyrus.

concepts and remain unaffected by other facial characteristics. To achieve this, we adopt a disentangled approach to facilitate the learning of emotion features. Disentanglement involves two aspects of constraints on the conceptual pathway: firstly, we employ a soft subspace orthogonality constraint (Chan et al. 2022) to encourage the separation of emotional and non-emotional features; and secondly, we append an emotion classifier after the emotion features to encourage the capture of emotion-related content. Let $\boldsymbol{F}_{emo} = [\boldsymbol{f}_{emo_1}, \boldsymbol{f}_{emo_2}, ..., \boldsymbol{f}_{emo_N}]^T$ and $\boldsymbol{F}_{non} = [\boldsymbol{f}_{non_1}, \boldsymbol{f}_{non_2}, ..., \boldsymbol{f}_{non_N}]^T$ respectively, where N is the batch size. We define a orthogonal loss $\mathcal{L}_{orth}$ as

$$\mathcal{L}_{orth} = \left\| \boldsymbol{F}_{emo}{}^T \boldsymbol{F}_{non} \right\|_F^2, \tag{1}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm.

During the training of the emotion classifier, one way is to directly employ the multi-class cross-entropy loss function. However, this approach may lead to the learning process of emotion features being influenced by samples with high emotional ambiguity. Therefore, we employ an adaptive importance module that can assess the emotional confidence score for each emotion feature. It is expected that emotion features with ambiguity may have low confidence scores, while those with high certainty have high scores. These scores are subsequently used to adjust the loss function of the emotion classifier. Specifically, this module consists of a linear fully-connected (FC) layer and a sigmoid activation function, enabling the output confidence scores to range from 0 to 1, which can be formulated as,

$$\alpha_i = \sigma(\boldsymbol{W}_{FC}{}^T \boldsymbol{f}_{emo_i}), \tag{2}$$

where α_i is the confidence score of the $\boldsymbol{f}_{emo_i}$, $\boldsymbol{W}_{FC}$ is the parameters of the FC layer, and σ is the sigmoid function.

Then, these confidence scores can be used to re-weight the logits in the emotion classifier. We adopt the Logit-Weighted Cross-Entropy loss $\mathcal{L}_{WCE}$ (Wang et al. 2020b) as the loss function for the emotion classifier.

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{\alpha_i \boldsymbol{W}_{y_i}{}^T \boldsymbol{f}_{emo_i}}}{\sum_{j=1}^{C} e^{\alpha_j \boldsymbol{W}_j{}^T \boldsymbol{f}_{emo_i}}}, \tag{3}$$

$$\mathcal{L}_{RR} = max\{0, \delta_1 - (\alpha_H - \alpha_L)\}, \tag{4}$$

$$\alpha_H = \frac{1}{M}\sum_{i=0}^{M} \alpha_i,\ \alpha_L = \frac{1}{N-M}\sum_{i=M}^{N} \alpha_i, \tag{5}$$

where W_{y_i} and W_j represent the model parameters between $\boldsymbol{f}_{emo_i}$ and the ground-truth emotion label y_i, and between $\boldsymbol{f}_{emo_i}$ and the j-th emotion label, respectively. In a batch, the

confidence scores of all samples are ranked in descending order, where M samples belong to the high-score group and N-M samples belong to the low-score group. $\mathcal{L}_{RR}$ is used to enforce that the mean of the high-score group should be larger than the mean of the low-score group by a margin δ_1. We set $\delta_1 = 0.15, M = N \times 0.7$ in this paper.

Contrastive learning on emotion features. Concepts of specific emotions (anger, happiness, sadness, etc.) exhibit a high degree of dissociability, eliciting unique activation patterns in the brain (Camacho et al. 2023). Motivated by this insight, we apply the supervised contrastive learning (SupCon) (Khosla et al. 2020) to emotion features. This contrastive technique is conducive to facilitating the proximity of the same emotion in the feature space. In this work, we construct positive and negative pairs for each emotion feature. In a batch, for $\boldsymbol{f}_{emo_i}$, the positive pairs consist of emotion features belonging to the same emotion category as $\boldsymbol{f}_{emo_i}$, while the remaining emotion features form negative pairs with it. We define concept-contrastive loss for $\boldsymbol{f}_{emo_i}$ as

$$l_i = -\sum_{d \in D_i} log \frac{exp(\boldsymbol{f}_{emo_i} \cdot \boldsymbol{f}_{emo_d}/\tau)}{\sum_j exp(\boldsymbol{f}_{emo_i} \cdot \boldsymbol{f}_{emo_j}/\tau)}, \quad (6)$$

where $i, j \in \{1, 2, ..., N\}$, D_i denotes the set consisting of the index d of all positive emotion features for $\boldsymbol{f}_{emo_i}$ and τ is a temperature parameter. We set $\tau = 0.07$ in this paper.

The FER datasets collected from the Internet suffer from a significant class imbalance, with insufficient samples for negative emotions such as anger, disgust, fear and sadness. Therefore, when comparing the similarity of emotion features, we pay more attention to negative emotions.

Guided by a previous work (Li et al. 2022), we design a weight vector $\boldsymbol{V} = [w_1, w_2, ..., w_N]^T \in \boldsymbol{R}^{N \times 1}$ to re-weight the importance of each emotion feature in a batch. When the feature pertains to negative emotions, the importance coefficient is relatively large. Let E_c denotes the number of samples belonging to the c-th emotion in a batch, the importance coefficient of the i-th emotion feature can be formulated as,

$$w_i = 1 - \frac{E_c}{N}, c = 1, 2, ...C, i = 1, 2, ...N \quad (7)$$

where C represents the maximum label index. Therefore, the concept-contrastive loss $\mathcal{L}_{CC}$ is

$$\mathcal{L}_{CC} = \frac{1}{N} \sum_{i=1}^{N} V_i l_i, \quad (8)$$

where V_i is the i-th importance coefficient of $\boldsymbol{V}$.

In summary, in the training process of conceptual pathway, the whole loss function is given below,

$$\mathcal{L}_{CP} = \alpha \mathcal{L}_{orth} + \beta(\mathcal{L}_{WCE} + \mathcal{L}_{RR}) + \gamma \mathcal{L}_{CC}, \quad (9)$$

where the coefficients α, β and γ are used to balance three parts. After the conceptual pathway is well trained, the parameters of it are frozen to provide guidance for the perceptual pathway.

Perceptual Pathway

From a neuroscience perspective, the perceptual pathway models the process of visual perception and receives the guidance of emotion concepts in the conceptual pathway (Brooks et al. 2019). From a machine learning perspective, the perceptual pathway can exhibit reduced computational complexity and storage requirements compared to the conceptual pathway, benefiting from the abundant knowledge available within the conceptual pathway (Gou et al. 2021). Hence, the training process of the perceptual pathway is actually the transfer of conceptual knowledge from the conceptual pathway, which will be detailed below.

The perceptual coding is regularized by emotion concepts. During the training of the perceptual pathway, a facial emotion image $\boldsymbol{x}_i$ simultaneously entered both pathways to obtain the perceptual feature $\boldsymbol{f}_{per_i}$ and the emotion feature $\boldsymbol{f}_{emo_i}$, respectively. It is worth noting that if the confidence score of the emotional feature (computed by the adaptive importance module in the conceptual pathway) is too low, it may impede the provision of reliable conceptual knowledge to the perceptual pathway. To alleviate this situation, we adopt an emotional confidence evaluation strategy to determine which emotion concept is suitable to regularize the perceptual coding. Specifically, we set a threshold for the emotion confidence scores within each emotion label to assess the quality of the emotion feature for each sample. If α_i is greater than the threshold, we use $\boldsymbol{f}_{emo_i}$ for feature regularization. Otherwise, we randomly select and aggregate k emotion features from the high-score pool that belong to the same label as the $\boldsymbol{f}_{emo_i}$ and then use them to regularize the $\boldsymbol{f}_{per_i}$. The emotion concepts from the conceptual pathway can be formulated as

$$\boldsymbol{f}_{ec_i} = \begin{cases} \boldsymbol{f}_{emo_i}, & if\ \alpha_i > \delta_c \\ \frac{1}{K} \sum_{j=0}^{K} \boldsymbol{f}_{emo_j}, & otherwise \end{cases} \quad (10)$$

where $\boldsymbol{f}_{ec_i}$ denotes the emotion concepts for the i-th sample used to guide the perceptual pathway, δ_c is the threshold for the c-th emotion label that the i-th sample belongs to, and $\boldsymbol{f}_{emo_j}$ is the j-th emotion feature from K high-score emotion features that belong to the same label as the $\boldsymbol{f}_{emo_i}$. We independently arranged the confidence scores of samples for each emotion label in the training set in ascending order. Subsequently, we selected the 20th percentile of confidence scores as the threshold for each label by default. Additionally, we set $K = 8$ for each emotion label by default. We will discuss the impact of these two parameters in the ablation studies.

The logits from the conceptual pathway also contain abundant emotional information due to our constraints on emotion features. Therefore, we design the similarity loss $\mathcal{L}_S$ and distillation loss (Li et al. 2021b) $\mathcal{L}_D$ to encourage the perceptual pathway to extract rich emotional information from the conceptual pathway. We calculate the mean square error between $\boldsymbol{f}_{ec}$ and $\boldsymbol{f}_{per}$, and calculate the Kullback–Leibler (KL) divergence of the distribution of logits between the perceptual pathway and the conceptual pathway as follows,

Method	RAF-DB	AffectNet	Pre-train	FED-RO	#Params	Run time (R)
VGG16 (Simonyan and Zisserman 2014)	85.16	58.21	ImageNet	63.49	138M	$\times$12
ResNet18 (He et al. 2016)	86.08	59.15	ImageNet	65.32	11M	$\times$1
gACNN (Li et al. 2018)	85.07	58.78	R & A	66.50	224M	$>$12
SPWFA-SE (Li et al. 2020)	86.31	59.23	R & A	67.25	21M	$>$2
RAN (Wang et al. 2020c)	86.90	59.50	MS-Celeb-1M	67.98	11M	$\times$6
SCN (Wang et al. 2020a)	87.03	60.23	R & A	68.24	11M	$\times$1
DMUE (She et al. 2021)	89.42	63.11	-	-	$>$25M	$\times$2
Ours (percepPath)	86.28	59.52	R & A	66.75	11M	$\times$1
Ours (concepPath)	90.33	63.97	R & A	**73.00**	24M	$\times$2
Ours (CRPN)	89.71	63.06	R & A	**71.00**	11M	$\times$1

Table 1: Comparison with State-of-the-art CNNs Methods on RAF-DB, AffectNet and FED-RO (%). Abbreviations: percepPath – perceptual pathway; concepPath – conceptual pathway; CRPN – concept-regularized perceptual network; #Params – number of parameters; Run time (R) – the ratio of the computing time of each method to that of Ours (CRPN) for inferring one image on average (5 milliseconds).

$$\mathcal{L}_S = \frac{1}{N}\sum_{i=1}(\boldsymbol{f}_{ec_i} - \boldsymbol{f}_{per_i})^2, \tag{11}$$

$$\mathcal{L}_D = \frac{1}{Q}\sum_{i=1}(\boldsymbol{V}_{emo_i})[log(\boldsymbol{V}_{emo_i}) - log(\boldsymbol{V}_{per_i})], \tag{12}$$

where $\boldsymbol{V}_{emo_i}$ and $\boldsymbol{V}_{per_i}$ are the logits of the i-th sample in the conceptual pathway and the perceptual pathway, respectively, and Q is the number of samples in a batch with confidence scores higher than the thresholds.

In summary, in the training process of the perceptual pathway, the whole loss function is defined as

$$\mathcal{L}_{PP} = \lambda_1\mathcal{L}_{CE} + \lambda_2\mathcal{L}_S + \lambda_3\mathcal{L}_D, \tag{13}$$

where $\mathcal{L}_{CE}$ is the multi-class cross-entropy loss in the perceptual pathway.

Experimental Setup

Datasets

We evaluate the proposed framework on three in-the-wild facial emotion data sets (**RAF-DB** (Li, Deng, and Du 2017b), **AffectNet** (Mollahosseini, Hasani, and Mahoor 2017), **FED-RO** (Li et al. 2018)) and one in-the-lab data set (**IMAGEN face** task (Grosbras and Paus 2006)). In our experiment, we use six basic emotions (happiness, anger, sadness, fear, disgust, and surprise) and a neutral emotion in three in-the-wild data sets, along with three emotions (happiness, anger, and neutral) in the IMAGEN face data set. We adopt the overall sample accuracy as a performance metric in each data set.

RAF-DB The RAF-DB data set contains 30,000 facial images annotated with basic or compound expressions by 40 trained human coders. Consistent with the most previous work, we use 12,271 images as training data and 3,068 images as test data in this work.

AffectNet The AffectNet data set is by far the largest data set that provides both categorical and Valence-Arousal annotations. There are 283,901 images as training data and 3,500 images as test data in this work.

FED-RO The FED-RO dataset is the first FER dataset in the presence of real occlusions in the wild, and each image in FED-RO was carefully labeled by three people. There are 400 images in total. We employ FED-RO to assess the generalization performance of our framework in this work.

IMAGEN face The face task paradigm is used to elicit strong activation in the facial emotion processing systems. In this task, participants passively watched 18-second blocks of a face video where up to six actors displayed emotions such as anger, neutrality, or happiness. These three emotions each consist of four face videos. We converted these videos into a dataset of facial emotion at 30 frames per second. We use the model trained on R & A to obtain the predicted probabilities for each frame in every emotion video from the IMAGEN face dataset. Subsequently, we select the top 10 frames with the highest predicted probabilities from each actor's emotional face within each emotion video. This serves as the IMAGEN test data, consisting of 240 angry faces, 120 happy faces, and 230 neutral faces. The remaining frames will be utilized as the IMAGEN training data.

Data Processing and Encoders

In our experiments, we use Retinaface (Deng et al. 2020) to detect and resize all facial emotions to the size of $224 \times 224 \times 3$. ResNet-50/18(He et al. 2016) is used for two pathways. Specifically, we use the ResNet-50 pre-trained on VGGface2 as the backbone for the non-emotion encoder and fix its parameters. The emotion encoder has the same structure as the non-emotion encoder but with trainable parameters. For the percept encoder, we use ResNet-18 pre-trained on Ms-Celeb-1M as the backbone. We remove the last classifiers for all these ResNet models and project the embeddings into a 256-dimension feature, respectively.

Training Setting

We conduct all experiments with the PyTorch toolbox and four NVIDIA GeForce RTX 3090 GPUs. During training,

$\mathcal{L}_{orth}$	$\mathcal{L}_{CC}$	$\mathcal{L}_{RR}+\mathcal{L}_{WCE}$	RAF-DB	AffectNet
×	×	×	87.06	60.01
✓	×	×	87.52	61.33
×	✓	×	87.64	61.26
×	×	✓	88.02	61.97
✓	✓	×	88.78	63.08
✓	×	✓	89.16	63.23
×	✓	✓	88.25	62.89
✓	✓	✓	90.33	63.97

Table 2: Accuracy (%) comparsion of the different components in the conceptual pathway. When we do not use the combination of $\mathcal{L}_{RR}+\mathcal{L}_{WCE}$, we use the general multi-class cross-entropy loss as a replacement.

Method	RAF-DB	AffectNet
perceptPath	86.28	59.52
Knowledge distillation	88.36	61.63
CRPN	89.71	63.06

Table 3: Evaluation of the influence of emotion concepts on the accuracy (%) of the perceptual pathway.

the batch size is 64. We use Adam with weight decay of 1e-5 and initial learning rate of 1e-4. The learning rate is updated by a cosine function with a period of 5. The training ends at epoch 100 and 40 for the conceptual and perceptual pathways, respectively. The α, β and γ are set as 0.1, 1 and 0.01 in the conceptual pathway, while the λ_1, λ_2 and λ_3 are set as 2, 1 and 1 in the perceptual pathway, as this combination can achieve the highest performance in this work. The source code of this paper is available at https://github.com/hanluyt/emotion-conceptual-knowledge.

Results and Analysis

Comparison With State-of-the-Art Methods

We construct the concept-regularized perceptual pathway of our network (*i.e.,* CRPN) using a CNN, therefore, we compare its performance with several state-of-the-art CNN methods for emotion recognition. As listed in Table 1, our methods are among the best on both the RAF-DB and the AffectNet data sets. When we combine the RAF-DB and the AffectNet data sets (*i.e.,* R & A) for training, our proposed CRPN achieves the best performance when tested using the independent FED-RO data set, but is also among the lightest (*i.e.,* 11M parameters) and fastest networks (*i.e.,* 5 ms per image for inference).

Ablation Studies

Evaluation of three components in the conceptual pathway. To assess the effect of each components, we design an ablation study to investigate $\mathcal{L}_{orth}$, $\mathcal{L}_{CC}$ and $\mathcal{L}_{RR}+\mathcal{L}_{WCE}$ on RAF-DB and AffectNet. We show the experimental results in Table 2. Several observations can be concluded in the following. First, when adding only one module into the baseline (1st row), the use of $\mathcal{L}_{RR}+\mathcal{L}_{WCE}$ (4th row) resulted in the highest improvement, indicating the significance of considering emotional confidence in FER. Secondly, when comparing rows 5-7, we can observe that disentanglement leads to better performance, indicating the contribution of disentanglement in extracting discriminative features for FER.

The i-th percentile	K	RAF-DB	AffectNet
15	4	88.40	61.77
15	6	88.44	61.95
15	8	88.89	62.04
15	10	89.11	62.19
20	4	88.78	62.14
20	6	89.26	62.65
20	8	89.71	63.06
20	10	89.74	63.02

Table 4: Ablation studies for the parameters of different values in the emotional confidence evaluation strategy.

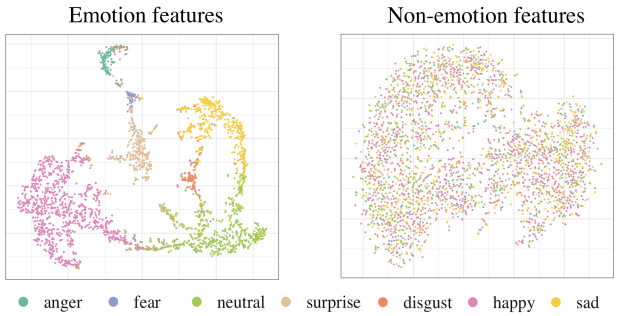

Figure 2: Visualization of the disentangled features in the conceptual pathway on RAF-DB.

Evaluation of the effect of emotion concepts in the perceptual pathway. We show the experimental results in Table 3. To evaluate the impact of emotion concepts on visual perception encoding, we compared three scenarios for the perceptual pathway: 1) training the perceptual pathway independently (1st row); 2) using the emotion features from the conceptual pathway to regularize the perceptual features in the perceptual pathway (2nd row); and 3) employing the emotional confidence evaluation strategy to obtain stable and reliable emotion concepts for guiding the encoding of the perceptual pathway (3rd row). Firstly, when comparing the 1st row with the 2nd and 3rd rows, we can observe that the transfer of emotional information from the conceptual pathway leads to better performance of the perceptual pathway. Secondly, when comparing the 2nd row with the 3rd row, we can observe the effectiveness of the emotional confidence evaluation strategy.

Evaluation of the parameters in the emotional confidence evaluation strategy. We assess the perceptual pathway's performance by varying parameters in the emotional confidence evaluation. Confidence thresholds at the 15th or 20th percentile and K values of 4, 6, 8, and 10 are considered. We show the experimental results in Table 4. Firstly, increasing the confidence threshold from the 15th percentile to the 20th percentile enhances the performance of the perceptual path-

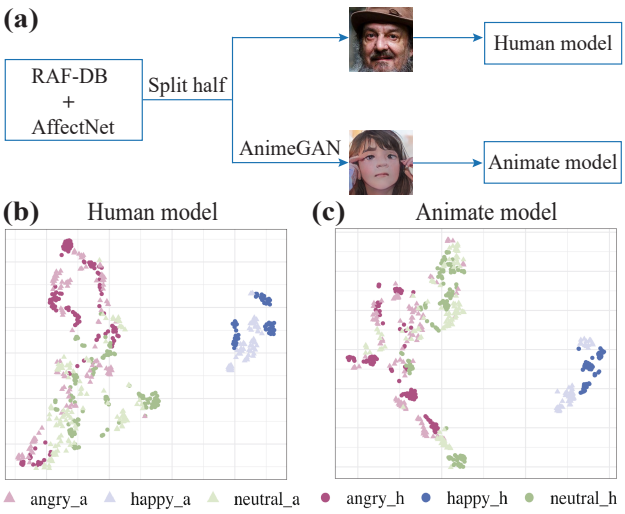

Figure 3: Exploring emotional concepts across faces with different styles. (a) The flowchart depicting the separate training process of our framework using human faces and animated faces. (b) The distribution of the emotion features obtained from IMAGEN's human faces and animated faces in the Human model. (c) The distribution of the emotion features obtained from IMAGEN's human faces and animated faces in the Animate model.

way, highlighting the significance of avoiding highly ambiguous emotion features. Secondly, increasing the value of K effectively improves the performance of the perceptual pathway, indicating that aggregating more high-confidence emotional features leads to more stable and reliable emotional concepts. However, the performance improvement becomes marginal when increasing K from 8 to 10, while demanding additional computational cost and time. Therefore, we choose the combination of the 20th percentile threshold and K=8.

Visualization

Emotion features vs. Non-emotion features. We demonstrate the effectiveness of the disentangled approach in our framework using the RAF-DB dataset as an example. We adopt t-SNE (Van der Maaten and Hinton 2008) to visualize the feature distribution of the RAF-DB test data in the conceptual pathway. As shown in Figure 2, emotion features of the same emotion label are clustered together, with clear boundaries between different emotion labels. While non-emotion features cannot distinguish emotions as expected, this indicates that the non-emotion features do not contain emotional information.

Emotion concepts are not affected by facial geometric attributes. We randomly divide the R & A data into two halves. One half is directly used to train our framework, resulting in the Human model. The other half is first transformed into animated faces using AnimeGAN (Chen, Liu, and Chen 2020) and then used to train our framework, resulting in the Animate model (Figure 3a). We also randomly split the IMAGEN test data into two halves, with one half transformed into animated faces using AnimeGAN. Then, we adopt t-SNE to observe the emotion feature distributions of the human faces and animated faces from the IMAGEN test data in both the Human model and the Animate model (Figures 3b, 3c). We observe that the emotion features from the same emotion across faces with different styles can be effectively aggregated together, indicating that our framework is capable of extracting abstract emotional concepts.

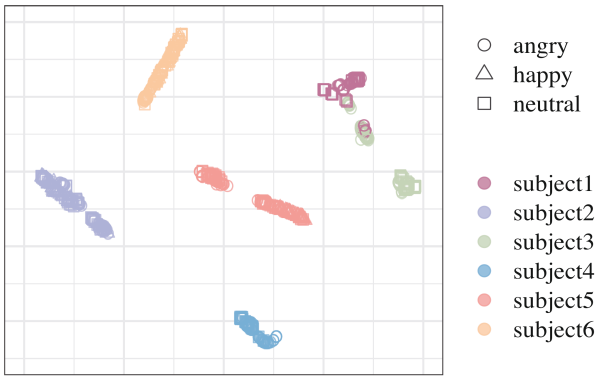

Figure 4: The distribution of non-emotion features for the same actor with different emotions in the IMAGEN test data, indicating that identity information is preserved.

Identity can be extracted from the non-emotion features. We append an identity recognition classifier after the non-emotion features and fine-tune the conceptual pathway using the IMAGEN training data. This allows the non-emotion features to better reflect identity features. Subsequently, we utilize the IMAGEN test data to observe the t-SNE distribution of non-emotion features for the same actor displaying different emotions. As shown in Figure 4, the fine-tuned model achieves a clear boundary between different actors with a large blank space. Moreover, different emotions of the same actor are clustered together. This indicates that the fine-tuned non-emotion features can better capture identity information, independent of emotional information.

Conclusion

Inspired by the neuroscience studies on the important role of emotion concepts in visual emotion perception, we proposed a novel brain-inspired FER model that is more lightweight and offers better interpretability compared to traditional FER algorithms. We considered a disentangled design to extract abstract emotion concepts that are independent of facial geometric attributes. Furthermore, we employed an emotional confidence evaluation strategy to select suitable and reliable emotion concepts for assisting visual encoding. The experiments validate the performance, showing the effectiveness and generality of our proposed framework.

Acknowledgments

This study was partially supported by grants from the National Key Research and Development Program of China (No. 2023YFE0109700), the National Natural Science Foundation of China (No.s: 82272079, 62372115 and 61971142), the Program of Shanghai Academic Research Leader (No. 23XD1423400), the Shanghai Municipal Science and Technology Major Project (No.s: 2018SHZDZX01 and 2021SHZDZX0103), and the University of Sydney – Fudan University BISA Flagship Research Program.

References

Brooks, J. A.; Chikazoe, J.; Sadato, N.; and Freeman, J. B. 2019. The neural representation of facial-emotion categories reflects conceptual structure. *Proceedings of the National Academy of Sciences*, 116(32): 15861–15870.

Brooks, J. A.; and Freeman, J. B. 2018. Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature human behaviour*, 2(8): 581–591.

Camacho, M. C.; Nielsen, A. N.; Balser, D.; Furtado, E.; Steinberger, D. C.; Fruchtman, L.; Culver, J. P.; Sylvester, C. M.; and Barch, D. M. 2023. Large-scale encoding of emotion concepts becomes increasingly similar between individuals from childhood to adolescence. *Nature Neuroscience*, 1–11.

Cela-Conde, C. J.; Marty, G.; Maestú, F.; Ortiz, T.; Munar, E.; Fernández, A.; Roca, M.; Rosselló, J.; and Quesney, F. 2004. Activation of the prefrontal cortex in the human visual aesthetic perception. *Proceedings of the National Academy of Sciences*, 101(16): 6321–6325.

Chan, K. H. R.; Yu, Y.; You, C.; Qi, H.; Wright, J.; and Ma, Y. 2022. ReduNet: A white-box deep network from the principle of maximizing rate reduction. *The Journal of Machine Learning Research*, 23(1): 4907–5009.

Chen, J.; Liu, G.; and Chen, X. 2020. AnimeGAN: a novel lightweight GAN for photo animation. In *International symposium on intelligence computation and applications*, 242–256. Springer.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

DiCarlo, J. J.; Zoccolan, D.; and Rust, N. C. 2012. How does the brain solve visual object recognition? *Neuron*, 73(3): 415–434.

Ding, N.; Tang, Y.; Fu, Z.; Xu, C.; Han, K.; and Wang, Y. 2023. GPT4Image: Can Large Pre-trained Models Help Vision Models on Perception Tasks? arXiv:2306.00693.

Freeman, J.; and Simoncelli, E. P. 2011. Metamers of the ventral stream. *Nature neuroscience*, 14(9): 1195–1201.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129: 1789–1819.

Grosbras, M.-H.; and Paus, T. 2006. Brain networks involved in viewing angry hands or faces. *Cerebral cortex*, 16(8): 1087–1096.

Haxby, J. V.; Hoffman, E. A.; and Gobbini, M. I. 2000. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6): 223–233.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1921–1930.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kanwisher, N.; Khosla, M.; and Dobs, K. 2023. Using artificial neural networks to ask 'why'questions of minds and brains. *Trends in Neurosciences*, 46(3): 240–254.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Kragel, P. A.; Reddan, M. C.; LaBar, K. S.; and Wager, T. D. 2019. Emotion schemas are embedded in the human visual system. *Science advances*, 5(7): eaaw4358.

Kriegeskorte, N. 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1: 417–446.

Lee, A. C.; Yeung, L.-K.; and Barense, M. D. 2012. The hippocampus and visual perception. *Frontiers in human neuroscience*, 6: 91.

Li, H.; Sui, M.; Zhao, F.; Zha, Z.; and Wu, F. 2021a. MVT: mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*.

Li, H.; Wang, N.; Ding, X.; Yang, X.; and Gao, X. 2021b. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30: 2016–2028.

Li, H.; Wang, N.; Yang, X.; and Gao, X. 2022. CRS-CONT: a well-trained general encoder for facial expression analysis. *IEEE Transactions on Image Processing*, 31: 4637–4650.

Li, S.; Deng, W.; and Du, J. 2017a. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2852–2861.

Li, S.; Deng, W.; and Du, J. 2017b. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2852–2861.

Li, Y.; Lu, G.; Li, J.; Zhang, Z.; and Zhang, D. 2020. Facial expression recognition in the wild using multi-level features and attention mechanisms. *IEEE Transactions on Affective Computing*.

Li, Y.; Zeng, J.; Shan, S.; and Chen, X. 2018. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, 28(5): 2439–2450.

Lumer, E. D.; and Rees, G. 1999. Covariation of activity in visual and prefrontal cortex associated with subjective visual perception. *Proceedings of the National Academy of Sciences*, 96(4): 1669–1673.

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

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3967–3976.

Pons, G.; and Masip, D. 2017. Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Transactions on Affective Computing*, 9(3): 343–350.

Rolls, E. T. 2023. Emotion, motivation, decision-making, the orbitofrontal cortex, anterior cingulate cortex, and the amygdala. *Brain Structure and Function*, 1–57.

Schiller, P. H. 1995. Effect of lesions in visual cortical area V4 on the recognition of transformed objects. *Nature*, 376(6538): 342–344.

She, J.; Hu, Y.; Shi, H.; Wang, J.; Shen, Q.; and Mei, T. 2021. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6248–6257.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020a. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6897–6906.

Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020b. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6897–6906.

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

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; and Dai, J. 2023. VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. arXiv:2305.11175.

Xue, F.; Wang, Q.; and Guo, G. 2021. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3601–3610.

Zeng, J.; Shan, S.; and Chen, X. 2018. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, 222–237.

Zhang, H.; Ding, X.; Liu, N.; Nolan, R.; Ungerleider, L. G.; and Japee, S. 2023. Equivalent processing of facial expression and identity by macaque visual system and task-optimized neural network. *NeuroImage*, 273: 120067.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.

Zhuang, C.; Yan, S.; Nayebi, A.; Schrimpf, M.; Frank, M. C.; DiCarlo, J. J.; and Yamins, D. L. 2021. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3): e2014196118.