OUS: Bridging Scene Context and Facial Features to Overcome the Rigid Cognitive Problem

Xinji Mai[1], Haoran Wang[1], Zeng Tao[1], Junxiong Lin[1], Shaoqi Yan[2], Yan Wang[1, *], Jiawen Yu[1], Xuan Tong[1], Yating Li[1], Wenqiang Zhang[1, 3, 4, *]

[1]Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering & Technology, Fudan University
[2]School of Electrical and Electronic Engineering, Shanghai Insittute of Techonlogy
[3]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
[4]Engineering Research Center of AI & Robotics, Ministry of Education, Fudan University
xjmai23@m.fudan.edu.cn, wqzhang@fudan.edu.cn, yanwang19@fudan.edu.cn

Abstract

Dynamic Facial Expression Recognition (DFER) is crucial for affective computing but often overlooks the impact of scene context. We have identified a significant issue in current DFER tasks: human annotators typically integrate emotions from various angles, including environmental cues and body language, whereas existing DFER methods tend to consider the scene as noise that needs to be filtered out, focusing solely on facial information. We refer to this as the Rigid Cognitive Problem. The Rigid Cognitive Problem can lead to discrepancies between the cognition of annotators and models in some samples. To align more closely with the human cognitive paradigm of emotions, we propose an Overall Understanding of the Scene DFER method (OUS). OUS effectively integrates scene and facial features, combining scene-specific emotional knowledge for DFER. Extensive experiments on the two largest datasets in the DFER field, DFEW and FERV39k, demonstrate that OUS significantly outperforms existing methods. By analyzing the Rigid Cognitive Problem, OUS successfully understands the complex relationship between scene context and emotional expression, closely aligning with human emotional understanding in real-world scenarios.

Code — https://github.com/Xinji-Mai/OUS

Introduction

DFER is a key aspect of affective computing, facing the ongoing challenge of ambiguous emotion classification. During dataset construction, unclear expressions are often discarded, yet DFER methods still struggle with ambiguous classifications. Our analysis reveals that this issue stems from the cognitive gap between annotators and models, rather than annotation errors.

Most DFER approaches treat scene information as noise, retaining only facial inputs, while human annotators consider full scene context during expression evaluation (De Gelder, de Borst, and Watson 2015)(Sinke, Kret, and de Gelder 2013)(Mai et al. 2024). This discrepancy creates what we term the Rigid Cognitive Problem (RCP). For example, without context, a facial expression might suggest sadness or fear; however, scene context can clarify it as happiness or comfort, as seen in Fig. 2 (Righart and Gelder 2008)(Porter, Spencer, and Birt 2003)(Sabatinelli et al. 2011). RCP arises because human annotators instinctively combine scene and facial information, using the full context to resolve ambiguity. In contrast, DFER methods, relying solely on facial data, see performance drops due to this incomplete information, as shown in Fig. 1.

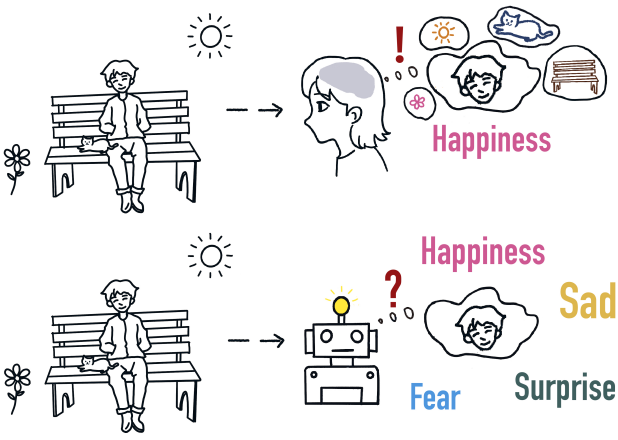

Figure 1: Human annotators instinctively combine scene information and faces when labeling emotions. DFER methods only uses face information for emotion prediction.

Psychological theories support the existence of RCP. Ekman (Ekman 1971) noted that while basic facial expressions are universally recognized, accurate interpretation requires contextual understanding. Schachter and Singer's Two-Factor Theory (Schachter and Singer 1962)(Müller et al. 2013) highlights how emotion results from physiological arousal and its cognitive interpretation within a context. These theories underline that emotion recognition, as conducted by human annotators, integrates multiple factors including facial cues, body language, and scene context (Abramson et al. 2021)(Metallinou et al. 2011). Further, environmental psychologist Roger Barker's Behavior Settings Theory (Hall 1969) shows that emotions are deeply

*Corresponding author

influenced by surroundings, where scene polarity impacts emotion perception. Positive scenes typically evoke positive emotions, and negative scenes the opposite (Duncan Jr 1969)(Wiener et al. 1972)(LaFrance and Mayo 1978). Thus, considering scene polarity in DFER is intuitive.

However, current DFER methods mainly focus on facial features, discarding scene information which conveys crucial emotional nuances, including scene polarity. Humans use scene polarity, objects, and facial cues together to infer emotions, but DFER lacks this capability, leading to incomplete emotion classification (Ziemke, Zlatev, and Frank 2007)(Beck, Cañamero, and Bard 2010).

To address RCP, we propose the Overall Understanding of the Scene (OUS) framework. OUS aims to: (1) Isolate scene polarity to guide ambiguous expression classification via polarity loss; (2) Align scene and facial information, recognizing the strong link between non-facial cues (e.g., body movements, objects) and emotions through similarity loss; (3) Design a novel fusion module for merging scene and facial features, two homogeneous and cross-scale multimodal information, utilizing the Multimodal Cross-Scale Fusion Encoder (MCFE), given the spatial rather than temporal nature of these cues; (4) Develop robust feature classification and loss functions, employing contrastive loss and flexible prompt design to manage noise from scene information.

Our contributions are:

- We identify and analyze the Rigid Cognitive Problem, introducing the OUS method that effectively integrates scene information for emotion extraction. Experiments on DFEW and FERV39k datasets demonstrate OUS's superiority over existing methods.
- We design three loss functions—similarity loss, polarity loss, and contrastive loss—to align scene and facial information, extract scene polarity, and classify emotional information. These multi-loss strategies reduce the latent space distance between scene and facial features, aiding emotion classification.
- We introduce the Multimodal Cross-Scale Fusion Encoder (MCFE), leveraging cross-scale attention to effectively integrate scene and facial features, capturing emotion-related cues.

Relate Work

DFER Methods

DFER is generally more robust than static expression recognition due to temporal correlations between frames (Saleem, Zeebaree, and Abdulrazzaq 2021)(Guo, Zhao, and Pietikäinen 2012). Recent advancements in techniques and datasets have significantly pushed this field forward (Fang et al. 2014)(Montirosso et al. 2010), offering deeper insights and paving the way for practical applications (Liu et al. 2014). Early work by Tao (Tao et al. 2023) established the foundation for high dynamic emotion extraction through frequency-based analysis of complex videos. Vision Transformer (ViT) (Dosovitskiy et al. 2020), built on the Transformer architecture (Vaswani et al. 2017), has shown remarkable potential in DFER due to its superior feature extraction and ability to capture long-range temporal dependencies. Originally designed for NLP, Transformers have been adapted to excel in DFER by processing sequential data. Moreover, methods like CLIP (Radford et al. 2021), CLIPER (Li et al. 2023), and A^3lign-DFER(Tao et al. 2024), leveraging cross-modal contrastive learning, have improved emotion recognition by integrating visual and textual data.

Figure 2: Scene polarity helps determine mood. When look at the face alone, the upper and lower expressions may be happiness, fear, sadness, etc., but by injecting the polarity of the scene, it can be judged that these are two completely different expressions (happiness and sadness).

Unlike these approaches, which primarily focus on facial information, our method emphasizes the importance of recognizing the entire scene alongside facial expressions. We argue that scene information, often discarded as noise, is a valuable context that significantly enhances emotion understanding.

Scene Information in DFER Datasets

The growth of DFER has been fueled by the development of comprehensive datasets. CK+ (Lucey et al. 2010), RAF-DB (Gera and Balasubramanian 2021), AFEW (Ekenel 2012), MMI (Valstar, Pantic et al. 2010), DFEW (Jiang et al. 2020), and FERV39k (Wang et al. 2022) are among the most notable, each contributing to advancing DFER in natural and controlled settings.

DFEW and FERV39k stand out as the largest unconstrained DFER datasets, containing extensive scene information beyond facial details. Despite this, most existing methods discard scene context when analyzing facial expressions. We challenge this approach, proposing that scene information provides essential context for accurate emotion recognition. Our method uniquely incorporates scene data to enhance emotional understanding, setting it apart from traditional DFER approaches.

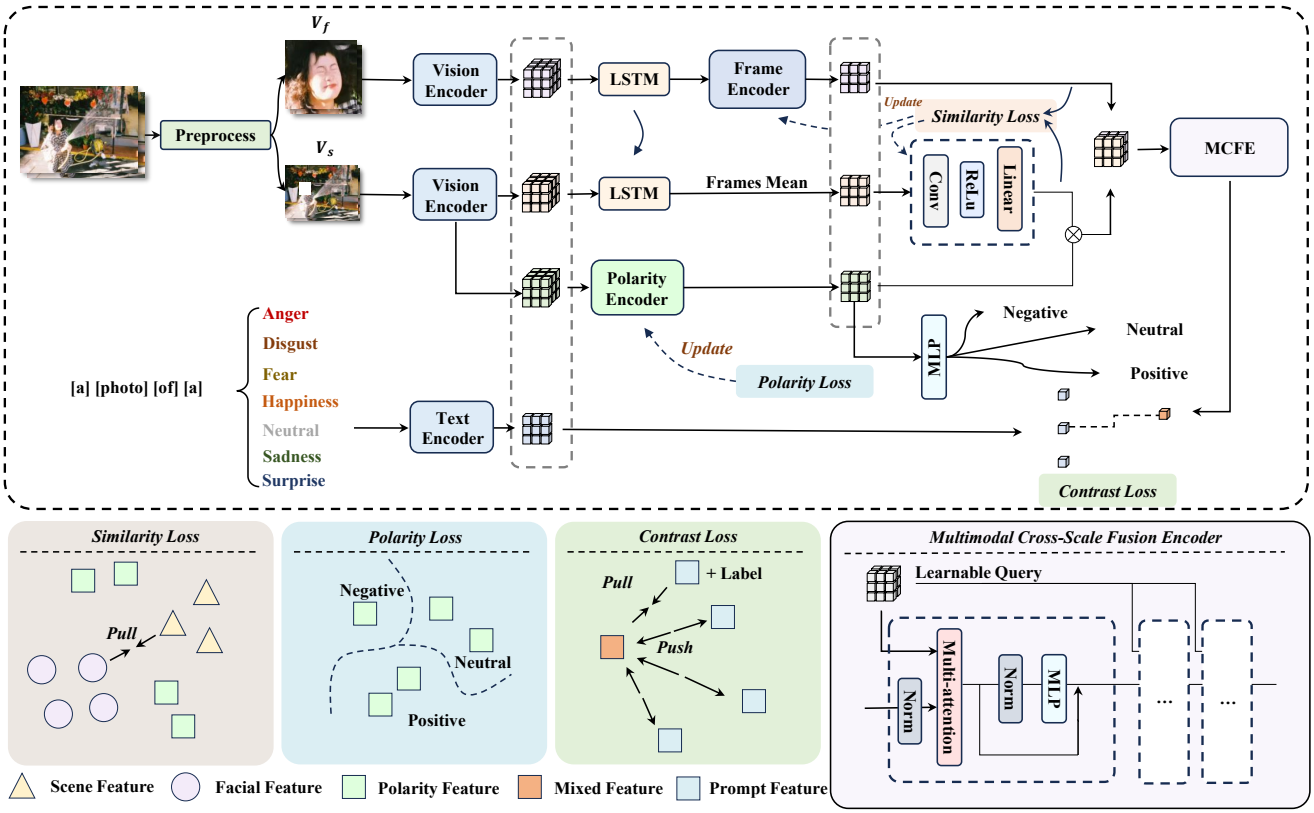

Figure 3: **OUS Overall Architecture Diagram.** OUS employs a dual-stream structure, separating images into facial and scene images. These are encoded into latent spaces V_f and V_s using a shared Vision Encoder $f_V(\cdot)$. Temporal features are fused via LSTM, with V_f processed through a Frame Encoder and V_s through Frames Mean. Similarity Loss $L_{similarity}$ aligns the latent spaces, reducing their distance. Polarity Loss extracts polarity information, which is fed into the Multimodal Cross-Scale Fusion Encoder using Learnable Queries (Q_{learn}). These queries, initialized with a Gaussian distribution, refine during training to capture emotional entities. OUS uses updatable prompts for contrastive loss with the output features.

Methodology

In this section, we explore the complexity of the proposed OUS framework. The overall structure of OUS is shown in Fig. 3. OUS mainly includes spatial encoding, temporal encoding, multi-Loss constraints and MCFE fusion composition. Vision Encoder is used for spatial feature extraction. Frames Encoder is used to extract time characteristics. Similarity Loss is used to align the hidden space, and the Contrast Loss is used to classify emotional classification. MCFE is used to integrate scene characteristics and facial features.

Overview of Training Strategy for Different Losses

Our framework primarily relies on a training strategy reinforced by three distinct loss functions to guide the optimization strategy. We posit that the introduction of scene information comprises three parts: the guidance of ambiance and texture information, the alignment with facial information in the latent space, and a loss that can flexibly match environmental entities. Specifically, color, textural, and luminance information, as well as related visual entities in the scene (such as carousels, artillery fire, etc.), can preliminarily determine the emotional polarity (positive, neutral, negative) of humans in that scene. For instance, in bright, flowery environments like amusement parks, the emotional polarity is likely positive, whereas in dark, bloody environments with ongoing artillery fire, it is more likely negative. We believe that this emotional polarity can guide the extraction of concepts related to emotions from the scene.

We use the polarity loss to optimize the polarity encoder so it can extract emotion-related polarity information from the output of the first four layers of the vision encoder to guide the cross-scale attention mechanism. The polarity loss is defined as the cross-entropy loss:

$$L_{polarity} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{1}$$

where y_i is the true polarity label and $\hat{y}_i$ is the predicted polarity.

To fuse scene features with facial features, we need to align them in the same latent space. The similarity loss is used to optimize the frame encoder, linear, and convolutional layers, helping align scene features V_s and facial features V_f

in the latent space and narrowing the relationship between scene features and emotion classification. The similarity loss is defined as the cosine similarity:

$$L_{similarity} = 1 - \frac{\sum_{i=1}^{N}(V_{f_i} \cdot V_{s_i})}{\sqrt{\sum_{i=1}^{N} V_{f_i}^2}\sqrt{\sum_{i=1}^{N} V_{s_i}^2}}, \quad (2)$$

where V_{f_i} and V_{s_i} represent the facial and scene features, respectively.

Finally, Due to the introduction of complex scene information, fixed prompts such as "a photo of" are no longer suitable for complex environmental scenes. Based on this, we use variable prompts that update during the training process. We compare the output features with the prompts using a contrastive loss function L_{cont} to optimize the overall model, defined as:

$$L_{contrast} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(V_{s_i}, V_{f_i})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(V_{s_i}, V_{f_j})/\tau)}, \quad (3)$$

where sim denotes the similarity function, τ is a temperature scaling parameter, and N is the batch size.

When the loss value exceeds α, we employ a global loss, which is the sum of similarity loss, polarity loss, and contrast loss:

$$L_{global} = L_{similarity} + L_{polarity} + L_{contrast}, \quad (4)$$

and when it is less than α, we only use the contrastive loss. This design helps to reduce the potential spatial distance between the scene and human features, and helps the model classify emotions effectively through polarity guidance and contrast loss. Besides, It helps our model converge quickly and find an appropriate solution space.

Multimodal Cross-Scale Fusion Encoder

In the fusion of facial and scene features, we designed a cross-scale attention mechanism based on the attention mechanism. Unlike the block structure in transformers where attention is followed by normalization and then a feed-forward network, we first perform layer normalization on the facial and scene features used as keys and values.

$$V^{norm} = \text{LayerNorm}(V) \quad (5)$$

Following this, when inputting into the multi-head attention mechanism, we use a shared Learnable Query ($Qlearn$), initialized with a Gaussian distribution, as the Query, with facial features V_f^{norm} and scene features V_s^{norm} as the Key and Value respectively. $Qlearn$ is shared across all cross-scale attention blocks. During training, as backpropagation progresses, $Qlearn$ gradually denoises and learns to capture entities related to emotions from the facial and scene features, and filters the features that need attention from the Value matrix.

Learnable Queries are crucial for capturing the specific emotional entities within the context of the scene and facial expressions. A learnable query as it gradually denoises and learns, it pays attention to the parts related to emotion, and filters out the parts that do not need to be noticed in the process of multiplying with the value matrix, leaving only the features that need attention. These are initialized with a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ to introduce variability and flexibility in capturing diverse features. The parameters of the Gaussian distribution are chosen to reflect the expected distribution of the features:

$$Q_{learn} \sim \mathcal{N}(\mu, \sigma^2) \quad (6)$$

where μ and σ^2 are the mean and variance of the Gaussian distribution, respectively. This initialization ensures that the model starts with a diverse set of queries that can adapt to various features during training.

The output of the attention mechanism is then subjected to layer normalization:

$$O_{attn} = \text{LayerNorm}(\text{Attention}(Q_{learn}, K, V)) \quad (7)$$

Next, the normalized output is passed through a Multi-Layer Perceptron (MLP) as the feed-forward network:

$$O_{ff} = \text{MLP}(O_{attn}) \quad (8)$$

The MLP can be defined as:

$$\text{MLP}(x) = W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2 \quad (9)$$

Finally, the output of the MLP is normalized again:

$$O_{final} = \text{LayerNorm}(O_{ff}) \quad (10)$$

This design helps OUS effectively integrate facial and scene features. The layer normalization before the multi-head attention ensures that both types of features are on a similar scale, allowing the attention mechanism to more effectively learn the important features related to emotions. The use of a shared Learnable Query enables the model to focus on relevant entities and improve the emotional understanding of the scene and facial expressions.

Prompt Engineering

The application of prompt engineering in our OUS is inspired by the methodologies of the CLIP and CoOp papers. In CLIP, a series of text prompts are trained to assist the model in understanding and categorizing image content, while CoOp further learns a series of continuous vectors (i.e., learnable prompts) for zero-shot or few-shot learning on a pre-trained CLIP model. In our model, learnable prompts are used to guide the model in associating image features with emotional states. Specifically, we define a set of learnable vectors as prompts that are combined with image features to compute the final emotion classification probabilities. These prompts undergo optimization along with image features during training, capturing important semantic connections in the emotion recognition task. The prompts given to the text encoder are designed as follows:

$$Prompt = [V]_1[V]_2 \dots [V]_M + [\text{CLASS}], \quad (11)$$

Where each $[V]_m$ ($m \in \{1, \dots, M\}$) is a vector of the same dimension as the word embeddings (i.e., 512 for CLIP ViT-B/32, 768 for CLIP ViT-L/14), and M is a hyperparameter specifying the number of context tokens (Zhou et al. 2022).

Other Detail of Network Architecture

Video segments $\{V \mid V \in \mathbb{R}^{B\times T\times C\times H\times W}\}$ are initially processed by a preprocessing block $B_p(\cdot)$, resulting in the facial video sequence and the scene video sequence. Here, B, T, C, H, and W represent the batch size, number of frames, channels, height, and width of the video sequence, respectively. The preprocessing module includes a crucial component: the facial recognition module, which separates facial and environmental information within the video. The facial recognition module is set to a frozen state, meaning its weights remain unchanged during the training process, ensuring stability and consistency from video input to feature extraction.

Subsequently, we encode the facial and environmental information into latent space facial features V_f and scene features V_s using a shared Vision Encoder. The weights we use are from ViT-L/14. Each input frame of the dual-stream facial features and scene features is processed independently. Frames are divided into N patches and then flattened into a D-dimensional latent space as follows:

$$V_p = [v_p^1 E; v_p^2 E; \ldots; v_p^N E] + E_{pos}, \tag{12}$$

where v_p^i is the flattened patch vector, E is the embedding matrix, and E_{pos} is the positional embedding matrix.

The temporal (T) and batch (B) dimensions are merged into one dimension to form the facial features V_f and the scene features V_s, such that $V \in \mathbb{R}^{B\cdot T\times N\times D}$ and $V \in \mathbb{R}^{B\cdot T\times N\times F}$, with F denoting the feature dimension.

The Vision Encoder remains frozen throughout. Next, we input the facial features V_f and scene features V_s into an LSTM for early feature fusion. The LSTM processes the input features as follows:

$$h = \sigma(W_h \cdot [h_{\text{prev}}, x] + b_h), \tag{13}$$

where h represents the hidden state, h_{prev} is the previous hidden state, x is the input feature vector (either V_f or V_s), and σ is the activation function.

The facial temporal features are then processed to be V_{ft} by a trainable Frames Encoder $f_f(\cdot)$, designed to capture the dynamic nature of facial expressions over time. We first re-transform the facial features back into tensor shape $B \times T \times F$ and then input them into the Frames Encoder to be V_{st}. Scene information, having less temporal variation, is extracted by computing the frames means:

$$V_{st} = \frac{1}{T}\sum_{t=1}^{T} V_s(t), \tag{14}$$

providing a stable representation of the scene.

We use convolutional layers and fully connected layers to align the latent spaces of facial V_{ft} and scene features V_{st}, reducing the distance between them. Similarity Loss L_{sim} is used to update the Frame Encoder, convolutional layers, and fully connected layers.

Additionally, the features output from the initial attention blocks during encoding contain information related to color, texture, and overall ambiance of the environment, which are abstractly linked to emotions. We use a Polarity Encoder to encode these features to be V_{pol} and Polarity Loss L_{pol} to update the Polarity Encoder.

The output polarity features V_{pol} are concatenated with the processed facial V_{ft} and scene features V_{st} and fed into the cross-scale attention mechanism. The cross-scale attention mechanism uses Learnable Queries Q_{learn} as the Query, with facial and scene features as the Key and Value. The Learnable Queries Q_{learn} are initially randomly initialized to a Gaussian distribution and shared across all cross-scale attention blocks. During training, as backpropagation updates proceed, the Learnable Queries Q_{learn} gradually denoise and learn to capture entities related to emotions in the V_{pol}, V_{st}, and V_{ft}. The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q_{learn}K^T}{\sqrt{d_k}}\right)V, \tag{15}$$

where Q, K, and V represent the Query, Key, and Value matrices, respectively, and d_k is the scaling factor.

Finally, we believe that OUS introduces substantial scene information, making fixed prompt methods unsuitable. Instead, we use updatable prompts to compute contrastive loss $L_{contrast}$ with the output features.

Experiment

We aim to meticulously evaluate the performance of OUS in DFER tasks, conducted across two mainstream datasets in DFER: FERV39k and DFEW, which cover a wide range of real-world scenarios.

Training Details

OUS was trained in a computing environment with 4 NVIDIA GeForce RTX 3090 GPUs and an Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz. The training utilized the Adam optimizer. With an initial learning rate of 0.002 and a batch size of 16, the model was trained for 60 epochs. The learning rate was reduced to a third of its value whenever the loss on the validation set didn't decrease for five consecutive epochs, and training was considered converged when the rate fell below 1e-7. The model was deemed overfitting if the training accuracy exceeded 80%. The final model saved was the one with the lowest loss on the validation set.

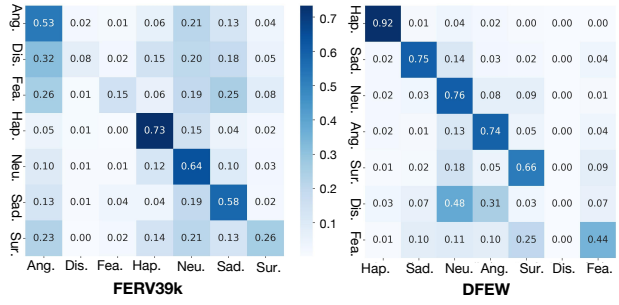

Figure 4: Confusion Matrices on DFEW and FERV39k. The tags are happiness, sadness, surprise, anger, neutral, disgust and fear.

Method	Publication	Happy	Sad	Neutral	Angry	Surprise	Disgust	Fear	UAR	WAR
VGG13+LSTM	/	76.89	37.65	58.04	60.7	43.70	0.00	19.73	42.39	53.70
C3D	CVPR'15	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
ResNet18+LSTM	/	83.56	61.56	68.27	65.29	51.26	0.00	29.34	51.32	63.85
ResNet18+GRU	/	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02
I3D-RGB	CVPR'17	78.61	44.19	56.69	55.87	45.88	2.07	20.51	43.4	54.27
P3D	ICCV'17	74.85	43.40	54.18	60.42	50.99	0.69	23.28	43.97	54.47
R(2+1)D18	CVPR'18	79.67	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22
3D R18+Center Loss	/	78.49	44.30	54.89	58.40	52.35	0.69	25.28	44.91	55.48
3D Resnet18	CVPR'18	76.32	50.21	64.18	62.85	47.52	0.00	24.56	46.52	58.27
EC-STFL	MM'20	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
Former-DFER	MM'21	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
NR-DFERNet	arXiv'22	88.47	64.84	70.03	75.09	61.60	0.00	19.43	54.21	68.19
GCA+IAL	C&C23	87.95	67.21	70.10	76.06	62.22	0.00	26.44	55.71	69.24
SW-FSCL	AAAI'23	88.35	68.52	70.98	**78.17**	64.25	1.42	28.66	57.25	70.81
LSGTNet	Appl Soft Comput'24	90.67	71.70	70.48	76.71	**65.01**	14.48	**40.24**	61.33	72.34
OUS (Ours)	/	**94.40**	**83.23**	**71.03**	77.33	60.98	**31.01**	34.12	**64.33**	**74.02**

Table 1: **Overall Model Performance Comparison on the DFEW dataset.** Bold represents the optimal result, and the underscore represents the suboptimal result.

Method	DFEW		FERV39k	
	UAR	WAR	UAR	WAR
C3D	42.74	53.54	22.68	31.69
P3D	43.97	54.47	23.20	33.39
I3D-RGB	43.40	54.27	30.17	38.78
3D ResNet18	46.52	58.27	26.67	37.57
R(2+1)D18	42.79	53.22	31.55	41.28
ResNet18-LSTM	51.32	63.85	30.92	42.95
ResNet18-ViT	55.76	67.56	38.35	48.43
EC-STFL	45.35	56.51	-	-
Former-DFER	53.69	65.70	37.20	46.85
NR-DFERNet	54.21	68.19	33.99	45.97
DPCNet	57.11	66.32	-	-
EST	53.94	65.85	-	-
LOGO-Former	54.21	66.98	38.22	48.13
IAL	55.71	69.24	35.82	48.54
CLIPER	57.56	70.84	41.23	51.34
M3DFEL	56.10	69.25	35.94	47.67
AEN	56.66	69.37	38.18	47.88
DFER-CLIP	59.61	71.25	41.27	51.65
EmoCLIP	58.04	62.12	31.41	36.18
LSGTNet	61.33	72.34	41.30	51.31
OUS (Ours)	**64.33**	**74.02**	**42.43**	**53.30**

Table 2: **Results of OUS on the two datasets DFEW and FERV39k.** Bold represents the optimal result, and the underscore represents the suboptimal result.

Performance Evaluation

OUS's performance was evaluated on FERV39k and DFEW datasets using weighted accuracy (WAR) and unweighted accuracy (UAR) as primary metrics. As shown in Tables 1 and 2, OUS outperforms existing methods, achieving a 1.68% improvement on DFEW and a 1.65% increase on FERV39k over the current SOTA.

Fig. 4 illustrates that categories like Happiness, Sad, Neutral, and Disgust achieve the highest accuracy, while Disgust and Fear remain the most challenging due to the long-tail distribution in DFER datasets. These categories also exhibit the most significant performance gains, supporting our hypothesis that incorporating contextual scene information enhances recognition. Particularly, Happiness, Sad, Neutral, and Disgust, often confused in isolated facial recognition, show marked accuracy improvement when scene context is considered, confirming the Rigid Cognitive Problem and the effectiveness of OUS.

We conducted a comprehensive comparison with the most advanced DFER methods over the past decade, including C3D (Tran et al. 2015), I3D-RGB (Carreira and Zisserman 2017), P3D (Qiu, Yao, and Mei 2017), and various 3D ResNet18 configurations (Hara, Kataoka, and Satoh 2018). ResNet-based models, such as ResNet18 with LSTM (He et al. 2016), GRU, and ViT (Dosovitskiy et al. 2020), were also evaluated. Additionally, we reviewed CLIP-based methods like CLIPER (Li et al. 2023), DFER-CLIP (Zhao and Patras 2023), and EmoCLIP (Foteinopoulou and Patras 2023), alongside others like Former-DFER (Jiang et al. 2020), LOGO-Former (Ma, Sun, and Li 2023), and NR-DFERNet (Li et al. 2022). Our results demonstrate OUS's superior robustness and effectiveness across FERV39k and DFEW datasets, making it a leading approach in dynamic facial expression recognition.

Ablation Study

Module ablation. We conducted ablation studies to assess the impact of multi-loss constraints and MCFE on model performance. We compared three settings: using only contrastive loss, applying multi-loss constraints (similarity, polarity, and contrastive losses), and replacing MCFE with average pooling.

The result, presented in Table 3, shows that removing multi-loss constraints significantly decreases both WAR and UAR, indicating that the interaction of these losses is crucial for model effectiveness. Specifically, using only contrastive loss led to slower convergence and reduced performance, highlighting the role of multi-loss in creating a more convergent solution space. Replacing MCFE with average pooling also resulted in a marked drop in accuracy, demonstrating MCFE's essential role in capturing emotional context from facial and scene features. The absence of MCFE caused sig-

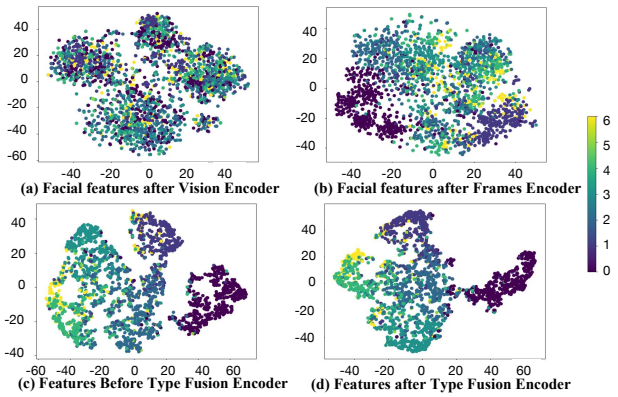

Figure 5: **Changes in Feature Clustering During the Inference Process.**The figure illustrates the global layout visualization of the feature space at different stages of the OUS model. The color legend 0 to 6 represent happiness, sadness, neutral, anger, surprise, disgust, and fear, respectively.

nificant reductions in UAR and WAR, underscoring its importance in extracting emotion-related connections.

Strategy	FERV39k			DFEW		
	UAR	WAR	Epoch	UAR	WAR	Epoch
L_{ce}	36.65	45.31	15	55.12	67.82	10
L_{global}	38.48	49.50	**6**	59.12	71.20	5
L_{ce} + MCFE	40.88	51.78	18	60.81	71.58	11
L_{global} + MCFE	**42.43**	**53.30**	**6**	**64.33**	**74.02**	**5**

Table 3: **Ablation results on FERV39k and DFEW,** L_{global} **and** L_{ce} **represents the 3-loss constraints and Cross entropy.** Bold represents the optimal result.

Hyperparametric ablation. We conducted ablation studies on prompt length and the combination of MCFE block numbers with prompt length to validate our hyperparameter choices. Results in Table 4 show that increasing prompt length positively impacts performance, with a length of 64 yielding the best accuracy and robustness.It also confirms that the optimal configuration is 12 MCFE blocks and a prompt length of 64, which effectively captures facial and scene interactions, enhancing emotion recognition accuracy.

Discussion

We compared OUS with other SOTA methods on the DFEW dataset for seven-class classification (Table 1). Our method, summarized in Table 2, consistently outperforms baselines, especially in recognizing happiness, sadness, and fear, validating the importance of scene context. OUS achieves significant improvements over SOTA: 3.73% in Happy, 11.47% in Sad, 0.55% in Neutral, and 16.53% in Disgust classification, leading by 1.68% on DFEW and 1.65% on FERV39k datasets.

As shown in Table 3 and 4, longer prompts correlate with higher accuracy. Models without MCFE and multi-loss constraints perform worst, while our approach improves performance by 7.99% and reduces convergence from 18 to 6 epochs. Fig. 5 illustrates that features are significantly clustered after MCFE, confirming the effectiveness of our multi-loss strategy and MCFE in enhancing OUS's performance.

Index	Setting		Evaluation	
	MCFE	Length	UAR	WAR
1	8	16	0.5692	0.7060
2	8	32	0.5990	0.7290
3	8	64	0.6097	0.7368
4	12	16	0.5842	0.7188
5	12	32	0.6324	0.7239
6 (Ours)	12	64	**0.6433**	**0.7402**
7	16	16	0.6015	0.7350
8	16	32	0.5997	0.7359
9	16	64	0.6162	0.7402

Table 4: **Results of hyperparametric ablation with different experimental settings on DFEW Dataset.** Bold represents the optimal result.

Conclusion

In this paper, we identified and analyzed the Rigid Cognitive Problem, a prevalent issue and methodological bias in DFER tasks. To address this problem, we designed a dynamic facial expression recognition method called Overall Understanding of the Scene (OUS), which effectively aligns and integrates scene information to extract emotional knowledge from complex features. The method primarily utilizes multiple loss constraints to reduce the latent space distance between scene and human features and efficiently classify emotions. Additionally, MCFE with a cross-scale attention mechanism and learnable queries effectively integrates scene and facial information. Extensive experiments conducted on the DFEW and FERV39k datasets demonstrated significant improvements over existing methods, validating the robustness and effectiveness of our approach. The superior performance of OUS illustrates the importance of utilizing scene information in the DFER task, and we hope our work can inspire other researchers.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No.62406075), National Key Research and Development Program of China (2023YFC3604802)

References

Abramson, L.; Petranker, R.; Marom, I.; and Aviezer, H. 2021. Social interaction context shapes emotion recognition through body language, not facial expressions. *Emotion*, 21(3): 557.

Beck, A.; Cañamero, L.; and Bard, K. A. 2010. Towards an affect space for robots to display emotional body language. In *19th International symposium in robot and human interactive communication*, 464–469. IEEE.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

De Gelder, B.; de Borst, A. W.; and Watson, R. 2015. The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2): 149–158.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Duncan Jr, S. 1969. Nonverbal communication. *Psychological bulletin*, 72(2): 118.

Ekenel, H. K. 2012. Benchmarking Facial Image Analysis Technologies (BeFIT). In *2012 3rd International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 15–15. IEEE.

Ekman, P. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Fang, H.; Mac Parthaláin, N.; Aubrey, A. J.; Tam, G. K.; Borgo, R.; Rosin, P. L.; Grant, P. W.; Marshall, D.; and Chen, M. 2014. Facial expression recognition in dynamic sequences: An integrated approach. *Pattern Recognition*, 47(3): 1271–1281.

Foteinopoulou, N. M.; and Patras, I. 2023. EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition. *arXiv preprint arXiv:2310.16640*.

Gera, D.; and Balasubramanian, S. 2021. Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. *Pattern Recognition Letters*, 145: 58–66.

Guo, Y.; Zhao, G.; and Pietikäinen, M. 2012. Dynamic facial expression recognition using longitudinal facial expression atlases. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*, 631–644. Springer.

Hall, E. T. 1969. Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6546–6555.

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.

Jiang, X.; Zong, Y.; Zheng, W.; Tang, C.; Xia, W.; Lu, C.; and Liu, J. 2020. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, 2881–2889.

LaFrance, M.; and Mayo, C. 1978. Cultural aspects of nonverbal communication. *International Journal of Intercultural Relations*, 2(1): 71–89.

Li, H.; Niu, H.; Zhu, Z.; and Zhao, F. 2023. CLIPER: A Unified Vision-Language Framework for In-the-Wild Facial Expression Recognition. *arXiv preprint arXiv:2303.00193*.

Li, H.; Sui, M.; Zhu, Z.; et al. 2022. NR-DFERNet: Noise-Robust Network for Dynamic Facial Expression Recognition. *arXiv preprint arXiv:2206.04975*.

Liu, M.; Shan, S.; Wang, R.; and Chen, X. 2014. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1749–1756.

Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; and Matthews, I. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, 94–101. IEEE.

Ma, F.; Sun, B.; and Li, S. 2023. Logo-Former: Local-Global Spatio-Temporal Transformer for Dynamic Facial Expression Recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Mai, X.; Lin, J.; Wang, H.; Tao, Z.; Wang, Y.; Yan, S.; Tong, X.; Yu, J.; Wang, B.; Zhou, Z.; et al. 2024. All rivers run into the sea: Unified modality brain-inspired emotional central mechanism. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 632–641.

Metallinou, A.; Katsamanis, A.; Wang, Y.; and Narayanan, S. 2011. Tracking changes in continuous emotion states using body language and prosodic cues. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2288–2291. IEEE.

Montirosso, R.; Peverelli, M.; Frigerio, E.; Crespi, M.; and Borgatti, R. 2010. The development of dynamic facial expression recognition at different intensities in 4-to 18-year-olds. *Social Development*, 19(1): 71–92.

Müller, C.; Cienki, A.; Fricke, E.; Ladewig, S.; McNeill, D.; and Tessendorf, S. 2013. *Body-Language-Communication. Volume 1*. Walter de Gruyter.

Porter, S.; Spencer, L.; and Birt, A. R. 2003. Blinded by emotion? Effect of the emotionality of a scene on susceptibility to false memories. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 35(3): 165.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, 5533–5541.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Righart, R.; and Gelder, B. d. 2008. Recognition of facial expressions is influenced by emotional scene gist. *Cognitive, Affective, & Behavioral Neuroscience*, 8(3): 264–272.

Sabatinelli, D.; Fortune, E. E.; Li, Q.; Siddiqui, A.; Krafft, C.; Oliver, W. T.; Beck, S.; and Jeffries, J. 2011. Emotional perception: meta-analyses of face and natural scene processing. *Neuroimage*, 54(3): 2524–2533.

Saleem, S. M.; Zeebaree, S. R.; and Abdulrazzaq, M. B. 2021. Real-life dynamic facial expression recognition: a review. In *Journal of Physics: Conference Series*, volume 1963, 012010. IOP Publishing.

Schachter, S.; and Singer, J. 1962. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5): 379.

Sinke, C. B.; Kret, M. E.; and de Gelder, B. 2013. Body language: Embodied perception of emotion. In *Measurement With Persons*, 349–366. Psychology Press.

Tao, Z.; Wang, Y.; Chen, Z.; Wang, B.; Yan, S.; Jiang, K.; Gao, S.; and Zhang, W. 2023. Freq-HD: An Interpretable Frequency-based High-Dynamics Affective Clip Selection Method for in-the-Wild Facial Expression Recognition in Videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 843–852. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Tao, Z.; Wang, Y.; Lin, J.; Wang, H.; Mai, X.; Yu, J.; Tong, X.; Zhou, Z.; Yan, S.; Zhao, Q.; et al. 2024. A3lign-DFER: Pioneering Comprehensive Dynamic Affective Alignment for Dynamic Facial Expression Recognition with CLIP. *arXiv preprint arXiv:2403.04294*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Valstar, M.; Pantic, M.; et al. 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, volume 10, 65. Paris, France.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Sun, Y.; Huang, Y.; Liu, Z.; Gao, S.; Zhang, W.; Ge, W.; and Zhang, W. 2022. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20922–20931.

Wiener, M.; Devoe, S.; Rubinow, S.; and Geller, J. 1972. Nonverbal behavior and nonverbal communication. *Psychological review*, 79(3): 185.

Zhao, Z.; and Patras, I. 2023. Prompting Visual-Language Models for Dynamic Facial Expression Recognition. *arXiv preprint arXiv:2308.13382*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Ziemke, T.; Zlatev, J.; and Frank, R. M. 2007. *Body, language, and mind*, volume 35. Walter de Gruyter.