VERT: End-to-End Visual Emotion Recognition with Context-aware Transformer

Xinpeng Li, Shuyi Mao, Teng Wang, Jinbao Wang, Feng Zheng, *Member, IEEE*, Xiaojiang Peng, *Member, IEEE*

***Abstract*—The typical visual emotion recognition system is characterized by sequential steps: subject detection, feature extraction, and emotion classification. However, this system might suffer from accumulative errors due to disjoint optimization and deficient feature interaction between subjects and contexts. To address these limitations, this paper introduces a novel end-to-end Visual Emotion Recognition framework with context-aware Transformers (VERT). The key principle is to directly predict subject locations and emotions from the entire image. To achieve seamless optimization and fine-grained feature interaction, the VERT adopts a deformable DETR architecture with set-based prediction. Furthermore, to better capture informative contextual clues, we utilize pre-trained queries of existing DETR models as context queries. In detail, we supplement the input with the context queries and design a fusion module to combine contexts into the face query. The module leverages position distance and feature relevance between face and context queries for effective context aggregation. Without bells and whistles, extensive experiments on two popular datasets demonstrate the superiority of FETR. Particularly, we achieve 98.1% on CAER-S and 37.26% on EMOTIC. Note that our method outperforms similar-parameters alternatives by 3.39% on CAER-S and 5.91% on EMOTIC. The code and models will be released soon.**

***Index Terms*—Affective computing, visual emotion recognition, end-to-end detection transformer.**

I. Introduction

Visual Emotion Recognition (VER) is to identify the emotions of each character in an image. This technique can be widely applied in health care, driver surveillance, and various human-computer interaction systems [5], [30], [31], [42] since emotion is the foundation for daily communication [8].

VER has attracted significant attention in the computer vision community and has achieved notable progress in the deep learning era [6], [17], [29], [33], [49]. As depicted in Fig. 1, the existing visual emotion recognition system generally includes sequential steps, i.e., subject detection, feature learning, and emotion classification. Traditionally, the system focuses on the perception of facial expressions, regarding facial images as discriminative emotional responses. As shown in Fig. 1(a), an off-the-shelf detector predicts face boxes and a face encoder extracts face features and classifies them into facial expression categories. The studies mainly focus on addressing issues of label uncertainty [3], [4], [13], [25], [34], [37], [38], [43], [53], [54], micro expression [28], [46], and disentangled representation [44], [52]. Recent studies move to context-aware emotion recognition, showing contexts, such as body language, scene semantics, and interactions, are crucial for emotion analysis [12], [14], [19], [26], [27], [35], [40], [41], [51]. As illustrated in Fig. 1(b), after subject detection, the system processes multiple emotional modalities through the subject and context encoders independently and fuses features or scores lately for emotion classification.

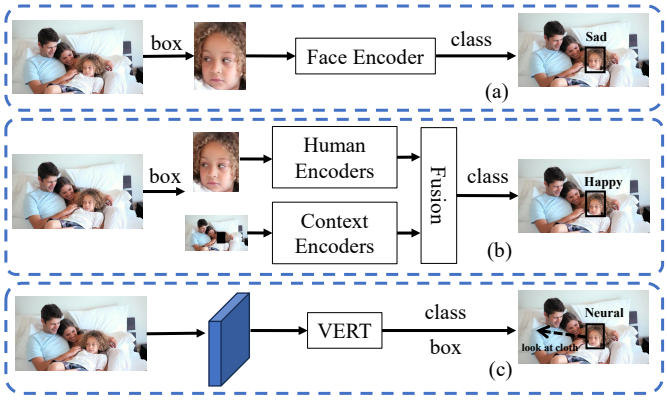

Fig. 1. Comparison on designs of visual emotion recognition framework. In the existing pipeline, human boxes are first detected and emotion classes are then inferred. The studies traditionally focused on the perception of facial expression in (a) and recently shifted to context-aware emotion recognition with subject and context streams. Instead, we argue that in (c) an end-to-end pipeline that directly predicts boxes and classes from the entire image is more favorable due to its seamless optimization and integrated feature interaction.

The current system might suffer from accumulative errors and deficient feature interaction between emotional inputs. Firstly, the networks of the detector and encoders are separate, and some handcrafted procedures like non-maximum suppression are not differentiable. Therefore, the system is disjointly optimized and thus suffers from accumulative errors. The boosted performance in Fig. 2 demonstrates the benefit of seamless optimization. Secondly, the emotion modalities are pre-processed independently with encoders and then combined together with a light fusion module. Such a late fusion paradigm inhibits interactions of low-level emotional cues between subjects and contexts. In Fig. 1, our proposal notices that the child is looking at clothes instead of the parents.

We argue that an end-to-end framework should be devised for seamless optimization and fine-grained feature interaction. However, such an end-to-end framework remains unexplored. The reasons might be two-fold. 1) Conventional VER, focusing on facial expression recognition, is originally transferred from face recognition and is thus pipelined with detection and recognition [18]. 2) Learning through an early fusion of different emotion modalities is challenging since contextual emotion cues are subtle, uncertain, and ambiguous.

To address the limitations, we propose **VERT**, an end-

Corresponding Author: Xiaojiang Peng (pengxiaojiang@sztu.edu.cn)

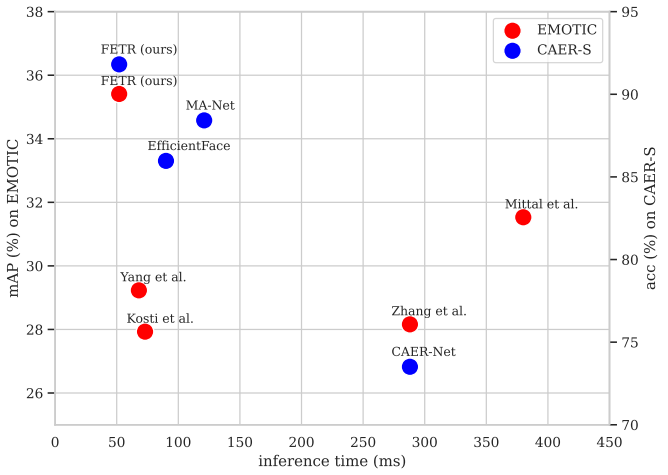

Fig. 2. Performance comparison of various methods on EMOTIC [12] and CAER-S [14]. VERT outperforms alternatives in accuracy and practical speed. The data needs to be supplemented and updated.

to-end Visual Emotion Recognition framework with context-aware Transformers. As shown in Fig. 1(c), VERT is performed on the entire image and outputs subject locations and emotions simultaneously and directly. The key elements of VERT are set-based prediction and deformable transformer architecture. Firstly, to offer continuous optimization and fine-grained interaction, we propose to adopt set-based prediction that directly predicts sets of emotion categories and subject locations. Additionally, considering emotional clues are size-changing, we suggest leveraging the deformable transformer architecture for multi-scale feature learning and interaction. As displayed in Fig. 2, experiments show that this end-to-end architecture outperforms alternative context-aware methods. Besides, VERT practically runs faster since it reduces non-maximum suppression and unifies multiple branches into one.

In addition to an end-to-end framework, it is desirable to design a module that captures informative emotional clues in the contexts. Building this module is non-trivial, which is required to determine where the valuable context areas are and distinguish useful information from disturbing ones [40]. Existing works rely on off-the-shelf detectors to find useful context elements and attentive learning to aggregate the contexts [12], [14], [26], [40], [41]. However, these methods are neither accurate for finding emotional clues nor suitable for end-to-end architecture. Instead, we resort to pre-trained context queries to find contextual clues and utilize position distance and feature relevance for attentive aggregation. To this end, the input of the decoder is supplemented by additional context queries, which adopt pre-trained queries of current DETR architectures. To combine contexts into the subject query, we design a plug-in context fusion module that takes subject and context queries as input and outputs augmented subject query. Particularly, we select a subset of context queries that have the K smallest position distance with subject queries. After that, we attentively aggregate this subset of context queries based on feature relevance between subject and context queries. Finally, we add the aggregated context to the subject query. The augmented subject query is demonstrated to be more effective.

Extensive experiments on two popular benchmarks demonstrate the effectiveness of our approach. The VERT achieves 98.1% on CAER-S [14] and 37.26% on EMOTIC [12]. On CAER-S, the VERT not only outperforms similar-parameter methods by 3.39% but also exceeds large-parameter approaches; On EMOTIC, the VERT surpasses similar-parameter methods by 5.91% and shows comparable performance with large-parameter approaches. Besides, we visualize the network output, feature map activation, and query position, along with ablation studies, showing that the VERT reasons out the valid and even subtle emotional clues in the context.

Overall, our key contributions can be summarized as:

- We embrace a new outlook of end-to-end VER framework, and demonstrate the necessity of end-to-end optimization and feature interaction between subject and context for FER.
- We present VERT, an end-to-end VER framework with contextual transformers. The keys are deformable DETR and set-based prediction, which facilitate end-to-end training.
- We propose a context-aware module that utilizes pre-trained context queries. This module resorts to pre-trained context queries to find contextual clues and leverages position distance and feature relevance for attentive aggregation.
- We exhibit significant accuracy gains over alternative methods on three popular benchmarks, showing the effectiveness of VERT. In detail, VERT exceeds similar-parameters methods on CAER-S by 3.39% and on EMOTIC by 5.91%.

II. Related Work

Visual Emotion Recognition. Given an in-the-wild image, the VER system performs subject detection, feature extraction, and emotion classification. 1) Some methods utilize face regions for facial expression recognition while regarding context areas as noise [6], [17], [29], [33], [49]. The studies mainly focus on addressing issues of label uncertainty [3], [4], [13], [25], [34], [37], [38], [43], [53], [54], micro expression [28], [46], and disentangled representation [44], [52]. 2) Recently, context-aware emotion recognition, which leverages multiple contexts for emotion classification, has attracted increasing attention [12], [14], [19], [26], [27], [35], [40], [41], [51]. Typically, a multiple-streams architecture followed by a fusion network encodes the face and context information independently. For the context information, [12] proposes using the whole scene image, [14] suggests masking the face of the scene image, [26] further combines inter-agent interaction, [41] adds agent-object interactions. *These methods contain two steps of detection and extraction and adopt heavy encoders and a late fusion module. Such a paradigm might suffer from accumulative errors and hinder low-level feature interaction between subjects and contexts. Differently, FETR processes the whole image in an end-to-end way, allowing seamless optimization and fine-grained feature interaction.*

End-to-End Object Detection. The end-to-end framework with vision Transformers stirs up wind in the object detection task. DETR [1] streamlines object detection into one step by a set-based loss and a transformer encoder-decoder

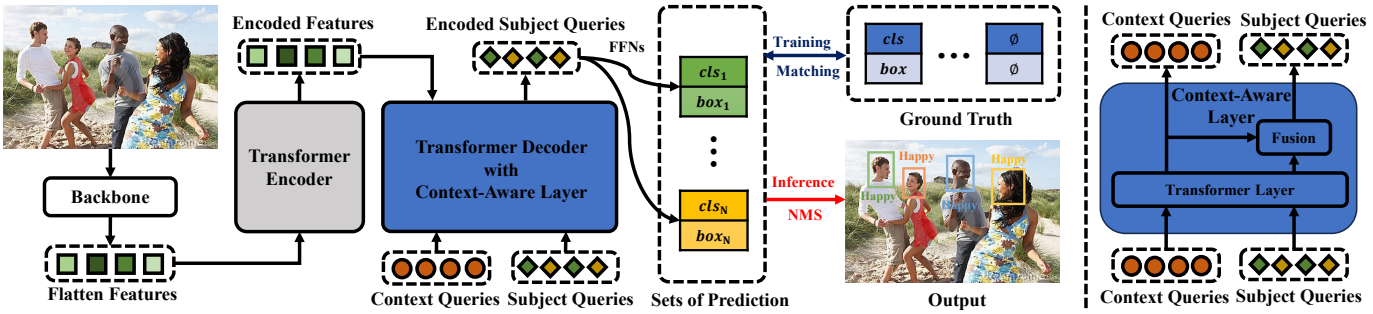

Fig. 3. Overall pipeline of VERT. Following [57], we extract multi-scale features through the backbone. Then, the encoder takes as input the multi-scale features and returns encoded features through the transformer encoder. After that, the decoder with context-aware layers is fed with the encoded features, subject queries, and context queries. The key is that subject queries are effectively fused with pre-trained context queries. Finally, individual Feed-Forward Networks (FFNs) turn the encoded subject queries into N sets of expression and box predictions. In training, we adopt Hungarian Matching [1] between predictions and padded ground truths to achieve end-to-end training. In inference, we output the high-quality positions and emotions of subjects after NMS.

architecture. The following works have attempted to eliminate the issue of slow convergence by designing architecture [7], [36], query [23], [39], [57], and bipartite matching [2], [15], [16], [47], [48]. Particularly, Deformable DETR [57] replaces Transformer attention with deformable attention that resorts to a small set of key sampling points around a reference. The DETR and its variants bring a simple but strong end-to-end architecture to other detection-related tasks. For example, MOTR [45] extends DETR to multiple-object tracking and introduces track queries to model the tracked instances in the entire video; TadTR [24] adopts DETR for temporal action detection with a temporal deformable attention module, a segment refinement mechanism, and an action regression head; HOTR [11] applies DETR to human-object interaction detection and introduces human and object pointers instead of directly regressing the bounding box. *Differently, this paper is the first work to extend DETR to facial expression recognition in the wild. The difference between object detection and facial expression recognition is that objects are clear and local but facial expressions might be ambiguous and scatter clues in the context. Therefore, this paper replaces deformable attention with context-aware attention that is able to capture emotional cues in the context.*

III. VERT Architecture

Existing methods for visual emotion recognition mainly follow several steps: subject detection, feature learning, and emotion classification. However, they might suffer from accumulative errors due to disjoint optimization and deficient feature interaction between subjects and contexts. Instead, in order to provide seamless optimization and fine-grained feature interaction, we propose a novel end-to-end framework VERT. As shown in Fig. 3, the VERT is based on set prediction and transformer encoder-decoder architecture.

Set Prediction. To obtain end-to-end training, we adopt set-level prediction and wrap the prediction into a set. Each set includes 1) the coordinate of the subject bounding box and 2) the emotion class. Let us denote the i-th set of ground truths as $y_i = (cls_i, box_i)$, where cls_i is the target emotion label and $box_i \in [0, 1]^4$ is a vector that defines ground truth box center coordinates and its height and width relative to the image size.

As shown in Fig. 3, we transform a fixed number of N subject queries (diamond-shaped) into N sets of predictions, where N is typically larger than the number of subjects in an image. To the best of our knowledge, VERT is the first framework that adopts a direct set-level prediction for visual emotion recognition, benefiting from end-to-end training.

Overall Architecture. We adopt deformable DETR [57] as the base architecture, considering emotional clues are size-changing. As shown in Fig. 3, the VERT includes a backbone, a transformer encoder, and a transformer decoder. Given an image x with height H and width W, the backbone extracts multi-scale feature maps $\{x^l\}_{l=1}^4$, where l indicates the feature level and $x^l \in \mathbb{R}^{C \times H_l \times W_l}$. Before being input, the multi-scale feature maps are added with position encoding and level embeddings. Then, the transformer encoder takes as input the flattened multi-scale features and returns encoded features after 6 transformer layers. After that, the transformer decoder updates the N subject queries and 300 context queries, which are processed by 6 context-aware transformer layers. Finally, Feed-Forward Networks (FFNs) turn the updated subject queries into N sets of class and box predictions.

The key difference from the base transformer layer is the context-aware layer (described in section IV). As shown in the right part of Fig. 3, the context-aware layer consists of a base transformer layer and a fusion module. The context queries and subject queries are concatenated and separated before and after the base transformer layer. Particularly, the subject query is fused with the context queries before being output.

Training and Inference. In training, since the prediction number is larger than the actual number of subjects in an image, we first pad the ground truths with $\emptyset$ to have the same size. Similar to [1], the bipartite matching then computes one-to-one pairs between the predictions $\hat{y}$ and the padded ground truths y. The formulation is:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\text{match}} \left(y_i, \hat{y}_{\sigma(i)} \right), \tag{1}$$

where $\hat{\sigma}$ is the optimal assignment, $\sigma \in \mathfrak{S}_N$ is a permutation of N elements, $\mathcal{L}_{\text{match}} \left(y_i, \hat{y}_{\sigma(i)} \right)$ indicates a pair-wise matching cost between ground truth and a prediction with index $\sigma(i)$.

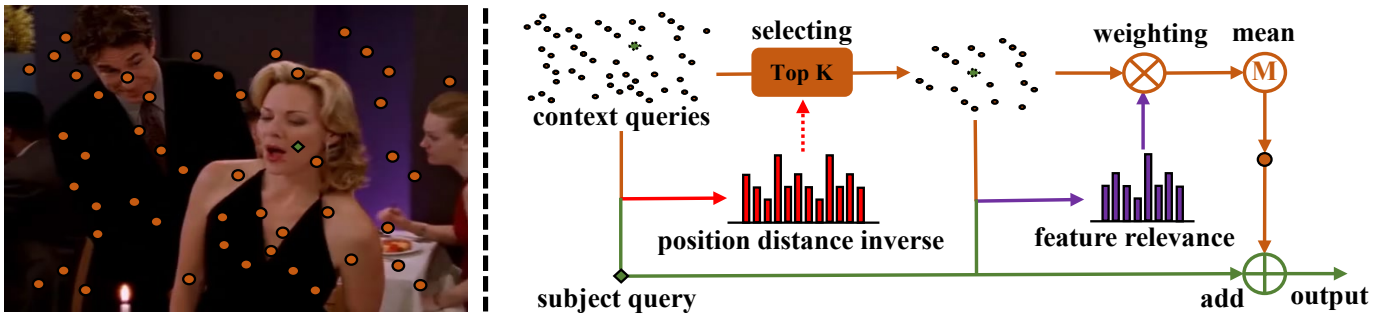

Fig. 4. Illustration of query positions and context fusion module. The object with dotted lines is imaginary for better understanding. The left figure shows the subject and context queries that attend to different areas. The right part describes how the subject query is fused with context queries. Firstly, the position distance inverse between the subject query and context queries is calculated. We select K context queries with top distance inverse. Then, we compute the feature relevance between the subject query and selected context queries and re-weight the context queries with the feature relevance. Finally, the subject query is added with the mean of re-weighted context queries as the final output.

$\mathcal{L}_{\text{match}}$ contains classification loss $\mathcal{L}_{\text{cls}}$ and box loss $\mathcal{L}_{\text{box}}$:

$$\mathcal{L}_{\text{match}} = \theta_{cls}\mathcal{L}_{\text{cls}}(y_i^{cls}, \hat{y}_{\sigma(i)}^{cls}) + \theta_{box}\mathcal{L}_{\text{box}}(y_i^{box}, \hat{y}_{\sigma(i)}^{box}), \quad (2)$$

where $\theta_{cls}, \theta_{box} \in \mathbb{R}$ are hyperparameters. The matching result is computed by the Hungarian algorithm [1] efficiently.

Getting the optimal assignment $\hat{\sigma}$, the training loss $\mathcal{L}$ is:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{\text{cls}}(y^{cls}, \hat{y}_{\hat{\sigma}}^{cls}) + \lambda_{box}\mathcal{L}_{\text{box}}(y^{box}, \hat{y}_{\hat{\sigma}}^{box}), \quad (3)$$

where $\lambda_{cls}, \lambda_{box} \in \mathbb{R}$ are hyperparameters. For matching and training, $\mathcal{L}_{\text{cls}}$ is set as the cross-entropy loss for multi-class task or focal loss [21] for multi-label task, $\mathcal{L}_{\text{box}}$ is set as the l_1 loss and generalized IoU loss [32].

In inference, we first set the score of each prediction as the mean of the class output logit, and then we use NMS to remove duplicate predictions. For the multi-label task, the emotions of a subject are calculated by a threshold t:

$$o = \arg_i \hat{y}_i > t, \quad (4)$$

where o is the index list of the emotion class. For the multi-class task, the emotion of a subject is calculated by:

$$o = \arg\max_i \{\hat{y}_i\}, \quad (5)$$

where o is the index of the emotion class.

IV. Context Fusion Module

Many studies show adding more emotional contexts boosts performance, which motivates us to add contextual clues to subject queries. However, locating context information and suppressing the noise is challenging. Existing works rely on off-the-shelf detectors to find useful context elements and attentive learning to fuse the contexts. These methods are neither accurate for finding emotional clues nor suitable for end-to-end architecture. To solve this issue, we propose to adopt pre-trained context queries of current DETR models as contextual clues. Since the context query might be useful or harmful for the emotion recognition of the subject, we design a context fusion module that utilizes position distance and feature relevance to attentively fuse contexts.

Context Queries. As shown in Fig. 3, in order to introduce context information, we supplement the subject queries with 300 pre-trained context queries of current DETR models. As shown in the left part of Fig. 4, the subject query (diamond-shaped) mainly attends to the subject area and the context queries (circle-shaped) are densely scattered throughout the picture. Since the context queries are pre-trained from object detection tasks, they can locate and encode contexts.

Fusion Module. As shown in Fig. 3, to combine contexts into the subject query, we design a context fusion module plugged into context-aware layers. The fusion module takes subject and context queries as input and outputs the augmented subject query, as shown in the right part of Fig. 4.

Firstly, we choose a subset of useful context queries and remove irrelevant ones based on position distance, assuming that the context queries are more relevant and useful if they are closer to the subject query. We calculate the position distance of the subject query and context queries. Let vector f and C_i represent the position of the subject query and the i-th point in context query set C, where i is the index of the point in set C. The distance d_i is calculated as:

$$d_i = \|f - C_i\|, \quad (6)$$

where d is the distance vector that contains the distance value between the subject query and context queries. Then, we select the top 100 context queries that are close to the subject query. The process can be formulated as:

$$D = \arg\max_{C_i \in C} \{1/d_i\} \quad \text{s.t.} \quad |D| = 100, \quad (7)$$

where D is a subset of context queries. The selected query set is later used for fusion with the subject query.

Secondly, we aim to fuse the subject query with selected 100 context queries based on feature relevance. Since the emotion in the context might be more uncertain than in the subject area [41], we resort to the relevance of subject features and context features to re-weight information. We first adopt the inner product to compute the relevance between features of the subject and context queries, followed by a softmax function for normalization. The relevance values are used to re-weight and aggregate all the context queries. Finally, we add the subject query with the aggregated context queries. The above operation can be formulated as:

$$o = softmax(a \left[D_1^{\mathrm{T}}, \quad \cdots \quad , D_n^{\mathrm{T}}\right]) \begin{bmatrix} D_1, \\ \vdots \\ D_n \end{bmatrix} + a, \quad (8)$$

where $a \in \mathbb{R}^{(1,d)}$ is the feature of the input subject query, $D_k \in \mathbb{R}^{(1,d)}$ is the k-th feature of context queries, and $o \in \mathbb{R}^{(1,d)}$ is the feature of the output subject query.

V. Experiments

A. Implements

For CAER-S [14] and EMOTIC [12], the number of queries N is 1. We initialize the weights of VERT and context queries with Deformable DETR [57] pre-trained on COCO [22]. The batch size is 32. θ_{box}, λ_{box}, θ_{cls}, and λ_{cls} are 5, 5, 2, and 5. We perform experiments on 8 GPUs of the NVIDIA Tesla A6000. Other architecture settings, training strategies, and pre-processing are the same with [57]. For evaluation, we compare the ground truth with the output subject that has the maximum overlap of the bounding box with the ground truth.

B. Datasets

The CAER-S dataset [14] consists of 70K images, which are randomly split into training (70%), validation (10%), and testing (20%) sets. The annotation contains the bounding box of the face and multi-class emotion. The emotion includes 7 categories: *Surprise, Fear, Disgust, Happiness, Sadness, Anger, Neutral*. The standard metric is the overall accuracy [14].

The EMOTIC dataset [12] contains a total number of 23,571 images and 34,320 annotated agents, which are randomly split into training (70%), validation (10%), and testing (20%) sets. The images are annotated with the bounding box of the body and head and multi-label emotion. The emotion includes 26 emotion categories: *Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning*. We report mean AP (mAP) of all classes, following [12].

C. Quantitative and Qualitative Results

Table I and Table II show the performance of different approaches on CAER-S and on EMOTIC. For a fair comparison, we separate the methods into similar-parameter measures and large-parameter ones, comparing their numbers of parameters with our proposal. The right column shows the detailed networks of the methods for parameter calculation. The subscripts of EmotiCon [26] and HECO [41] refer to specific context modalities according to their papers.

On CAER-S, the VERT not only outperforms similar-parameter methods by 3.39% but also exceeds large-parameter approaches. On EMOTIC, the VERT surpasses similar-parameter methods by 5.91% and shows comparable performance with large-parameter approaches. Fig. 5 illustrates the output of VERT, demonstrating its promising ability.

Methods	Acc (%)	Param. (M)	Details
CAER-Net-S [14]	73.51	22	12-layers CNN
GNN-CNN [50]	77.21	23	VGG16
EfficientFace [56]	85.87	25	MobileNet28, ResNet18
EMOT-Net [12]	74.51	32	ResNet18 x 2
SIB-Net [20]	74.56	33	ResNet18 x 3
VERT (ours)	**91.81**	39	ResNet50
GRERN [9]	81.31	45	ResNet101
RRLA [19]	84.82	48	ResNet50, RCNN50
MA-Net [55]	88.42	52	Multi-Scale ResNet18
EmotiCon [26]	88.65	181	OpenPose, RobustTP, Megadepth
VRD [10]	90.49	380	{VGG19, ResNet50, FRCNN50} $\times 2$
CCIM [40]	91.17	223	OpenPose, RobustTP, Megadepth, ResNet101

Table I. Performance of different approaches on the CAER-S [14].

Methods	mAP (%)	Param. (M)	Details
VERT (ours)	**37.26**	**39**	ResNet50
EMOT-Net [12]	27.93	84	YOLO, ResNet18*2
CAER-Net [12]	20.84*	84	YOLO, 12-layers CNN
EmotiCon$_{(1)}$ [26]	31.35	85	OpenPose, 15-layers CNN
EmotiCon$_{(1,3)}$ [26]	35.28	108	OpenPose, 15-layers CNN, Medadepth
HECO$_{(1)}$ [41]	22.25	135	YOLO, Alphapose, ResNet18
EmotiCon$_{(1,2)}$ [26]	32.03	139	OpenPose, RobustTP, ResNet18
HECO$_{(1,2)}$ [41]	36.18	146	YOLO, Alphapose, ResNet18 $\times 2$
HECO$_{(1,2,3)}$ [41]	34.93	173	YOLO, Alphapose, ResNet50, ResNet18 $\times 2$
EmotiCon$_{(1,2,3)}$ [26]	32.03	183	OpenPose, RobustTP, ResNet18, Megadepth
GCN-CNN [50]	28.16	189	YOLO, VGG16, 6-layers GCN
HECO$_{(1,2,3,4)}$ [41]	37.76	219	YOLO, Alphapose, {ResNet18, ResNet50} $\times 2$, Faster RCNN50
CCIM [40]	39.13	261	YOLO, Alphapose, {ResNet18, ResNet50} $\times 2$, Faster RCNN50, ResNet101

Table II. Performance of existing methods on the EMOTIC [12].

D. Why the end-to-end framework boosts performance?

We further conduct an evaluation for images with subject numbers and visualize the feature map to explore why the end-to-end framework boosts performance in this section. We choose EMO-Net [12] as the typical context-aware method with late fusion. For a fair comparison, we re-implement EMO-Net and adopt ResNet50 as the backbone to have similar parameters with the VERT (denoted as EMO-Net-R50).

Table III shows the performance of images with different subject numbers on EMOTIC. We also report the result when the input context is masked [14] (denoted as EMO-Net-R50-M). When the subject number in an image increases, it means the context becomes increasingly complicated. As we can see, when the subject number increases, VERT keeps a stable performance while EMO-Net-R50 deteriorates. It can be explained that the late fusion in existing context-aware networks fails to deal with complicated contexts.

Fig. 6 exhibits the feature maps of EMO-Net-R50 and VERT. We choose the feature that is outputted from the last layer of ResNet50. As we can see, EMO-Net-R50 tends to focus on several large chunks while VERT stresses areas with detailed outlines. It can be explained that VERT deals with contexts more accurately than existing late-fusion methods, thus yielding better performance.

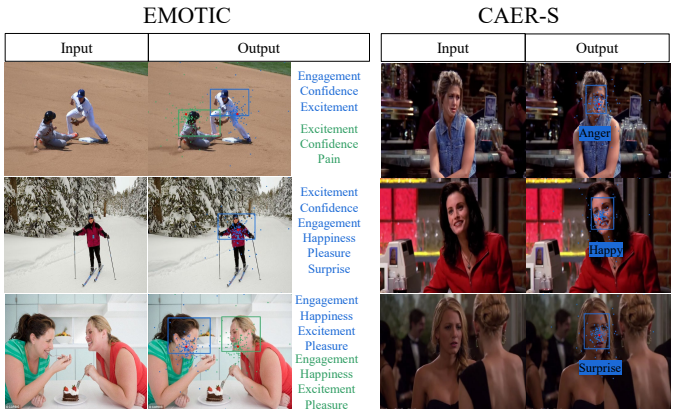

Fig. 5. The output visualization on EMOTIC [12] and CAER-S [14].

Face #	1	2	3	4	>=5
Image #	2444	938	234	37	29
EMO-Net-R50	22.34	20.50	19.62	18.77	18.06
EMO-Net-R50-M	22.54	20.96	19.65	19.20	19.50
VERT(Ours)	36.91	35.20	31.20	40.96	35.97

Table III. The performance of images with different subject numbers.

E. Ablation Study

Query number. We conduct experiments to see the effect of query number setting. As Table IV shows, VERT achieves the best result on EMOTIC and CAER-S when the query number is 1. It can be explained that most images of these two datasets have only one face. We further conduct experiments on EMOTIC-4, which is a subset of EMOTIC and only consists of 4 faces in an image. The VERT achieves the best result when the query number is 4. Thus, the query number setting is related to the average face number of a dataset.

Fig. 7 visualizes the positions of queries when the number changes. As we can see, the queries attend to different areas. This whole section needs to be supplemented and updated.

The coefficient of classification. Since emotion recognition focuses on classification accuracy while object detection focuses on box accuracy, it is crucial to tune the coefficient of classification for better performance. We conduct experiments of the hyper-parameters on EMOTIC [12] and show results in Table V. The performance is the best when the coefficient of matching is 2 and the coefficient of loss is 5.

Fig. 6. The visualization of feature maps from late and fine-grained fusion.

Set Number	1	2	5	10	20
EMOTIC (mAP %)	**36.55**	36.20	36.01	35.11	34.63
CAER-S (Acc %)	**91.78**	91.57	91.39	91.57	91.47
EMOTIC-4 (mAP %)	-	-	**40.96**	-	-

Table IV. The query number study on EMOTIC [12] and CAER-S [14].

Fig. 7. The query position visualization of different query numbers.

θ_{cls}	5	10	15	2	2	2
λ_{cls}	2	2	2	5	10	15
mAP (%)	35.41	35.89	35.41	**36.01**	34.99	34.64
θ_{cls}	2	5	10	15	1	1
λ_{cls}	2	5	10	15	10	8
mAP (%)	35.94	35.00	35.05	34.75	34.70	35.45

Table V. The performance of different coefficients of classification.

The components of VERT. The VERT consists of a base transformer, context queries, and the plug-in fusion module. We test different components of the VERT to demonstrate the effectiveness of the proposal. We conduct experiments on EMOTIC [12] and show results in Table VI. The baseline refers to the base transformer of the VERT, context means supplementing context queries without fusion, fusion implies adding the mean of the context queries with the face query, and selecting and weighting indicate the distance selection and relevance weighting in the fusion module. As we can see, the VERT exceeds the baseline by 0.71%.

The number of the selected context queries. We perform experiments of different K on EMOTIC [12] and show the results in Table VII. As we can see, the performance is the best when K is 100. It can be explained that not all contexts are useful for emotion recognition.

The strategy of re-weighting contexts. We conduct experiments of different re-weighting strategies on EMOTIC and display the results in Table VIII. Relevance, mean, distance and attention refer to aggregating context queries with feature relevance, equal values, position distance inverse, and attentive

baseline	context	fusion	selecting	weighting	mAP (%)
✓	×	×	×	×	36.55
✓	✓	×	×	×	36.21
✓	✓	✓	×	×	36.73
✓	✓	✓	×	✓	36.85
✓	✓	✓	✓	✓	**37.26**

Table VI. The performance of different components combination.

K	25	50	100	150	200	250	300
mAP %	37.01	37.25	**37.26**	37.01	36.94	36.91	36.85

Table VII. The result of different numbers of the selected context queries.

Strategy	relevance	mean	distance	attention
mAP %	**37.26**	36.73	36.48	36.93

Table VIII. The result of different strategies of re-weighting contexts.

weights. We can see the performance reaches its best when the re-weighting strategy is based on the feature relevance.

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[2] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*, 2022.

[3] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13984–13993, 2020.

[4] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021.

[5] Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, and Thibaut Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008.

[6] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *TPAMI*, 38(8):1548–1568, 2016.

[7] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2988–2997, 2021.

[8] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[9] Qinquan Gao, Hanxin Zeng, Gen Li, and Tong Tong. Graph reasoning-based emotion recognition network. *IEEE Access*, 9:6488–6497, 2021.

[10] Manh-Hung Hoang, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Context-aware emotion recognition based on visual relationship detection. *IEEE Access*, 9:90465–90474, 2021.

[11] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021.

[12] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *TPAMI*, 42(11):2755–2766, 2019.

[13] Nhat Le, Khanh Nguyen, Quang Tran, Erman Tjiputra, Bac Le, and Anh Nguyen. Uncertainty-aware label distribution learning for facial expression recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6088–6097, 2023.

[14] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *CVPR*, pages 10143–10152, 2019.

[15] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022.

[16] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023.

[17] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[18] Shan Li, Weihong Deng, and JunPing Du. 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*, pages 2852–2861, 2017.

[19] Weixin Li, Xuan Dong, and Yunhong Wang. Human emotion recognition with relational region-level analysis. *IEEE Transactions on Affective Computing*, 2021.

[20] Xinpeng Li, Xiaojiang Peng, and Changxing Ding. Sequential interactive biased network for context-aware emotion recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2021.

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[23] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

[24] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *arXiv preprint arXiv:2106.10271*, 2021.

[25] Tohar Lukov, Na Zhao, Gim Hee Lee, and Ser-Nam Lim. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In *European Conference on Computer Vision*, pages 648–665. Springer, 2022.

[26] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *CVPR*, pages 14234–14243, 2020.

[27] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2mm: Affective analysis of multimedia content using emotion causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5661–5671, 2021.

[28] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1492, 2023.

[29] Fatemeh Noroozi, Dorota Kaminska, Ciprian Corneanu, Tomasz Sapinski, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 2018.

[30] Giovanni Pioggia, Roberta Igliozzi, Marcello Ferro, Arti Ahluwalia, Filippo Muratori, and Danilo De Rossi. An android for enhancing social skills and emotion recognition in people with autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(4):507–515, 2005.

[31] Fuji Ren and Changqin Quan. Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. *Information Technology and Management*, 13(4):321–332, 2012.

[32] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[33] Philipp V Rouast, Marc Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 2019.

[34] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. 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*, pages 6248–6257, 2021.

[35] Dhruv Srivastava, Aditya Kumar Singh, and Makarand Tapaswi. How you feelin'? learning emotions and mental states in movie scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2528, 2023.

[36] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3611–3620, 2021.

[37] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020.

[38] Weijie Wang, Nicu Sebe, and Bruno Lepri. Rethinking the learning paradigm for facial expression recognition. *arXiv preprint arXiv:2209.15402*, 2022.

[39] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022.

[40] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015, 2023.

[41] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. Emotion recognition for multiple context awareness. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 144–162. Springer, 2022.

[42] Dongri Yang, Abeer Alsadoon, PW Chandana Prasad, Ashutosh Kumar Singh, and Amr Elchouemi. An emotion recognition model based on facial recognition in virtual learning environment. *Procedia Computer Science*, 125:2–10, 2018.

[43] Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, and Xinbo Gao. A circular-structured representation for visual emotion distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4237–4246, 2021.

[44] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20291–20300, 2022.

[45] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 659–675. Springer, 2022.

[46] Zhijun Zhai, Jianhui Zhao, Chengjiang Long, Wenju Xu, Shuangjiang He, and Huijuan Zhao. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22086–22095, 2023.

[47] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Jiaxing Huang, Kaiwen Cui, Shijian Lu, and Eric P Xing. Semantic-aligned matching for enhanced detr convergence and multi-scale feature fusion. *arXiv preprint arXiv:2207.14172*, 2022.

[48] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[49] Ligang Zhang, Brijesh Verma, Dian Tjondronegoro, and Vinod Chandran. Facial expression analysis under partial occlusion: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–49, 2018.

[50] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE, 2019.

[51] Sitao Zhang, Yimu Pan, and James Z Wang. Learning emotion representations from verbal and nonverbal communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18993–19004, 2023.

[52] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a facial expression embedding disentangled from identity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6759–6768, 2021.

[53] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021.

[54] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision*, pages 418–434. Springer, 2022.

[55] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021.

[56] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3510–3519, 2021.

[57] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.