

# VERT: End-to-End Visual Emotion Recognition with Subject-Context Transformer

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

**Abstract**—The typical visual emotion recognition system involves sequential steps: subject detection, feature extraction, and emotion classification. However, this system may suffer from accumulative errors due to disjoint optimization and limited feature interaction between subjects and contexts. To address these limitations, this paper introduces a novel end-to-end Visual Emotion Recognition framework with subject-context Transformers (VERT), allowing for seamless optimization and fine-grained feature interaction. The VERT directly processes raw images into set predictions that identify subjects’ locations and emotions simultaneously. Given the dynamic nature of emotion cues, the VERT adopts a deformable DETR architecture. Furthermore, to better capture informative contextual clues, we propose the subject-context transformer that encompasses context decomposition and fusion. Specifically, we separate the input queries into the subject and context queries and initialize the context queries with pre-trained queries of existing DETR models. We leverage spatial and semantic relations between face and context queries to facilitate query aggregation. Without bells and whistles, the VERT demonstrates its superiority on two popular datasets. Particularly, we achieve 98.1% on CAER-S and 37.81% on EMOTIC. Note that our method outperforms alternatives with similar parameter numbers by 3.39% on CAER-S and 6.46% 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) entails the precise identification of emotions portrayed by individuals in images. Its potential applications span across healthcare, driver surveillance, and diverse human-computer interaction systems [7], [36], [37], [50], reflecting the fundamental role of emotions in daily communication [10]. VER has gained substantial attention within the multimedia and computer vision community, particularly deep learning era [8], [23], [35], [40], [58].

Traditionally, the VER system comprises a sequence of key steps, including subject detection, feature learning, and emotion classification. These steps predominantly revolve around the interpretation of facial expressions, perceiving facial images as expressive indicators. A standard off-the-shelf detector is employed to predict facial regions, followed by a dedicated face encoder that extracts facial features, subsequently classifying them into distinct emotional categories, as depicted in Fig.1(a). Research in this multi-stage paradigm has primarily addressed challenges such as label uncertainty [5], [6], [19], [31], [41], [44], [45], [51], [62], [63], micro expression [34], [55], and disentangled representation [53], [61].

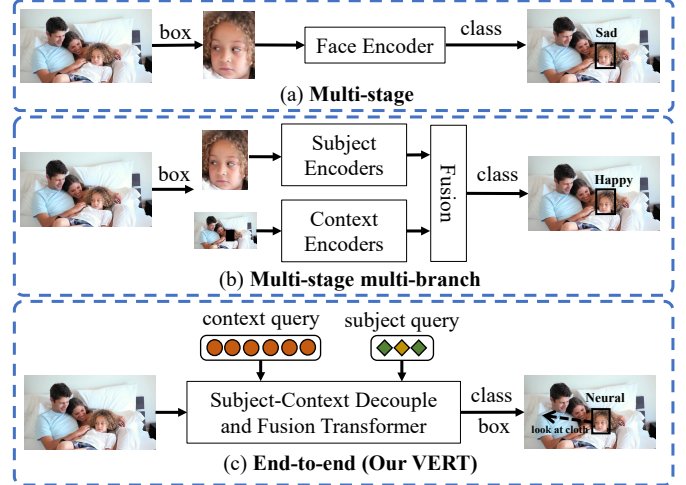


Fig. 1. Comparison of our VERT architecture (c) with the existing multi-step (a) and multi-stage multi-branch (b) paradigms. (a) In the multi-step pipeline, subject boxes are first detected and emotion classes are then inferred. The studies traditionally focused on the perception of facial expressions. (b) Within the multi-stage multi-branch paradigm, subject and context are detected, then encoded by different streams and fused together. These architectures mainly come from context-aware emotion recognition. (c) Our VERT is an end-to-end framework with subject-context decouple and fusion transformers to directly predict boxes and classes. We argue that an end-to-end pipeline is more favorable due to its seamless optimization and fine-grained feature interaction.

However, recent advances in VER pay increasing attention to context-aware emotion recognition, recognizing the significance of contextual elements like body language, scene semantics, and interactions in the analysis of emotions. The system first identifies subjects and contexts within the image and then independently processes them through subject and context encoders. The resulting features or scores are subsequently integrated for the final emotion classification, as illustrated in Fig. 1(b). In this multi-stage multi-branch paradigm, research mainly focuses on issues of emotion clue acquirement and combination [18], [20], [25], [32], [33], [42], [48], [49], [60].

The current system might suffer from accumulative errors and deficient feature interaction between subjects and contexts. The issues stem from the disjoint nature of the multi-stage pipeline, including the separate detector and encoders and non-differentiable handcrafted procedures like non-maximum suppression (NMS). Therefore, the system operates disjointly, resulting in accumulated errors. Another aspect of concern pertains to the multi-branch strategy that independently pre-processes multiple inputs and then combines them with a dedicated late fusion module. Such a paradigm inhibits interactions of low-level emotional cues between subjects and contexts.

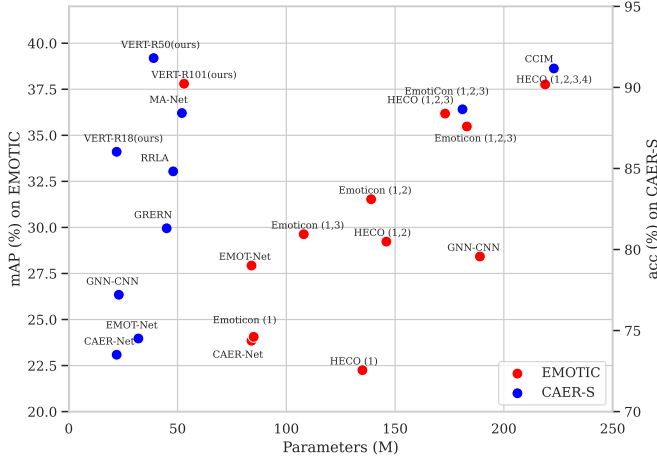


Fig. 2. Performance comparison of various methods on EMOTIC [18] and CAER-S [20]. VERT outperforms alternatives with fewer parameters.

We advocate for the development of an end-to-end framework for seamless optimization and fine-grained feature interaction. Surprisingly, such an end-to-end framework remains relatively unexplored, possibly due to two primary reasons. Firstly, conventional VER, focusing on facial expression recognition, inherits its structure from face recognition and is thus pipelined with separate detection and recognition stages [24]. Secondly, the early fusion of subjects and contexts is challenging, because contextual emotion cues are too subtle, uncertain, and ambiguous to capture from raw pixels.

In response to these challenges, we introduce VERT (Visual Emotion Recognition with Subject-Context Transformers), an end-to-end framework for Visual Emotion Recognition. Illustrated in Fig.1(c), VERT operates directly on the entire image, providing simultaneous and direct outputs of subject bounding boxes and emotion classes. This innovative approach leverages transformers to process both subjects and contexts from low-level pixels to high-level features, enabling fine-grained interactions. Furthermore, VERT streamlines the detection process by eliminating non-maximum suppression and consolidating multiple branches into a unified framework, allowing for seamless optimization. Additionally, considering emotional clues are size-changing, we propose the use of deformable transformer architecture for multi-scale feature learning and interaction. As depicted in Fig.2, experimental results demonstrate that this end-to-end architecture surpasses alternative methods, while utilizing fewer parameters.

In addition to achieving an end-to-end framework, it is desirable to design a module that effectively captures contextual emotional cues. Building this module is non-trivial, which is required to determine where the valuable context areas are and distinguish useful information from disturbing ones [48]. Existing approaches rely on off-the-shelf detectors to locate pertinent context elements and employ attentive learning to aggregate them [18], [20], [32], [48], [49]. In this work, we introduce the Subject-Context Transformer, a novel component that encompasses context decomposition and fusion. As depicted in Fig. 1(c), we partition the input queries into subject and context queries. To initiate context queries

with accurate and abundant context information, we employ pre-trained queries from existing DETR models. Subsequently, we exploit spatial and semantic relations between subject and context queries to facilitate query aggregation. Specifically, we identify a subset of context queries with the smallest position distance from subject queries and attentively aggregate this subset based on the feature relevance between subject and context queries. Ultimately, we merge the aggregated context query with the subject query. This decomposition and fusion strategy has proven to be highly effective.

Extensive experiments conducted on two widely-recognized benchmarks validate the efficacy of our approach. The VERT framework attains impressive results, achieving 98.1% on the CAER-S dataset [20] and 37.81% on EMOTIC [18]. Notably, on CAER-S, VERT not only surpasses similar-parameter methods by a substantial margin of 3.39% but also outperforms large-parameter approaches. In the case of EMOTIC, VERT outpaces similar-parameter methods by a notable margin of 6.46% while maintaining competitive performance with its large-parameter counterparts. Furthermore, we provide valuable insights by visualizing network output, feature map activation, query positions, and conducting ablation studies. These findings underscore VERT’s ability to discern meaningful and nuanced emotional cues within the context.

Our work makes several significant contributions:

- We introduce a fresh perspective on the end-to-end Visual Emotion Recognition (VER) framework, emphasizing the critical need for end-to-end optimization and nuanced feature interaction between subjects and contexts.
- We present the VERT framework, an end-to-end VER architecture empowered with subject-context transformers. By processing subjects and contexts together at various levels, from pixel-level details to high-level features, this design enables end-to-end training and fine-grained interaction.
- We propose the subject-context transformer, featuring context decomposition and fusion. This innovative approach involves decoupling input queries into subject and context queries, followed by a fusion process that leverages spatial-semantic relations. This decomposition and fusion strategy excels in capturing contextual emotional cues.
- Our empirical findings underscore the superiority of VERT over alternative methods on two prominent benchmarks. Specifically, VERT outperforms similar-parameter approaches by 3.39% on CAER-S and 6.46% on EMOTIC.

## II. RELATED WORK

**Visual Emotion Recognition.** The VER task can be broadly categorized into two main approaches. 1) *Face-Centric VER*. Many existing methods focus on facial expression recognition by utilizing face regions while treating contextual areas as noise, as observed in various studies [8], [23], [35], [40], [58]. The pipeline is featured as multi-stage, with subject face detection, feature extraction, and emotion classification. These studies primarily address challenges associated with label uncertainty [5], [6], [19], [31], [41], [44], [45], [51], [62], [63], micro expressions [34], [55], and disentangled representations [53], [61]. 2) *Context-Aware Emotion Recog-*

*dition.* In recent times, the research has paid increasing attention to context-aware emotion recognition, which emphasizes the use of multiple contexts for more robust emotion classification [18], [20], [25], [32], [33], [42], [48], [49], [60]. In addition to multi-stage components, the pipeline includes multi-branch characteristics. Typically, a multiple-stream architecture, followed by a fusion network, is employed to independently encode the face and context information. Various approaches are explored: [18] suggests using the whole scene image, [20] recommends masking the face in the scene image, [32] extends this approach by incorporating inter-agent interactions, and [49] adds agent-object interactions. *Despite the effectiveness of these methods, they all share a common multi-stage process involving detection and feature extraction, coupled with the multi-branch components of heavy encoders and late fusion modules. Such paradigms can introduce accumulative errors and hinder low-level feature interactions between subjects and contexts. In contrast, the VERT adopts an end-to-end processing approach, enabling seamless optimization and facilitating fine-grained subject and context interaction across the entire image.*

**End-to-End Object Detection.** The end-to-end framework with vision Transformers stirs up wind in the object detection task. DETR [2] streamlines object detection into one step by a set-based loss and a transformer encoder-decoder architecture. The following works have attempted to eliminate the issue of slow convergence by designing architecture [9], [43], query [29], [46], [67], and bipartite matching [4], [21], [22], [56], [57]. Particularly, Deformable DETR [67] replaces Transformer attention with deformable attention that resorts to a small set of key sampling points around a reference point. The original DETR framework, along with its various adaptations, has not only brought forth a simple yet powerful end-to-end architecture for object detection but has also been extended to other related tasks, including multiple-object tracking [54], action detection [30], human-object interaction [16], [17], person search [1], and instance segmentation [15], [47]. *In the context of VER, we propose the adaptation and modification: first, since we suggest seamless optimization and fine-grained subject and context interaction, we advocate for adopting DETR's end-to-end image processing; second, since objects exhibit distinct and localized whereas visual emotion might be ambiguous and scattered in the context, we introduce the subject-context transformer to replace the base transformer.*

### III. VERT ARCHITECTURE

Current approaches in visual emotion recognition traditionally involve a sequential process of subject detection, feature extraction, and emotion classification. Nevertheless, this conventional pipeline may encounter challenges related to cumulative errors stemming from disjoint optimization and insufficient feature interaction between subjects and contexts. In contrast, we introduce VERT, a pioneering end-to-end framework. As depicted in Fig. 3, VERT processes the entire image directly, delivering concurrent results for subject bounding boxes and emotion classes, facilitating seamless optimization and intricate feature interaction.

**Overall Architecture.** We adopt deformable DETR [67] as the base architecture because it can adapt to the size-changing nature of emotional clues. As shown in Fig. 3, our VERT system consists of three main components: a backbone, an encoder, and a decoder with subject-context transformers. Given an image, we extract multi-scale features through the backbone and add them with position encoding and level embeddings. The encoder subsequently handles the flattened multi-scale features, iteratively processing them through 6 transformer layers to yield encoded features. After that, the decoder updates both the  $N$  subject queries (used to identify subjects' emotions) and 300 context queries (used to understand the environment's emotions) across 6 subject-context transformer layers. Finally, Feed-Forward Networks (FFNs) generate the  $N$  sets of class and box predictions for the subjects from  $N$  updated subject queries.

To improve our system's ability to capture contextual cues, we adopt context decomposition and fusion strategy and propose the subject-context transformer. The strategy separates the input queries into subject and context queries, allowing the easier capture of accurate and sufficient contextual information. Additionally, the subject-context transformer contains two parts: a base transformer layer and spatial-semantic relational aggregation. It ensures that both the context and subject queries are well-processed, and their information is effectively combined to improve the representation of the subject queries.

**Set Prediction.** To achieve end-to-end training, we adopt a set-level prediction approach, encapsulating predictions within sets. Each set includes 1) the coordinate of the subject bounding box and 2) the corresponding emotion class. For clarity, we denote the  $i$ -th set of ground truths as  $y_i = (cls_i, box_i)$ , where  $cls_i$  represents the target emotion label and  $box_i \in [0, 1]^4$  is a vector that specifies the center coordinate and its height and width of the ground truth box 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.

**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 ensure a consistent size. Following [2], we employ the bipartite matching that computes one-to-one associations between the predictions  $\hat{y}$  and the padded ground truths  $y$ . The formulation is:

$$\hat{\sigma} = \arg \min_{\sigma \in \mathfrak{S}_N} \sum_i^N \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}), \quad (1)$$

where  $\hat{\sigma}$  represents the optimal assignment,  $\sigma \in \mathfrak{S}_N$  denotes a permutation of  $N$  elements,  $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$  indicates a pairwise matching cost between ground truth and a prediction with index  $\sigma(i)$ .  $\mathcal{L}_{\text{match}}$  encompasses a classification loss  $\mathcal{L}_{\text{cls}}$  and a box loss  $\mathcal{L}_{\text{box}}$ , expressed as:

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

where  $\theta_{\text{cls}}, \theta_{\text{box}} \in \mathbb{R}$  are hyperparameters. We efficiently compute the matching results using the Hungarian algorithm [2].

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

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

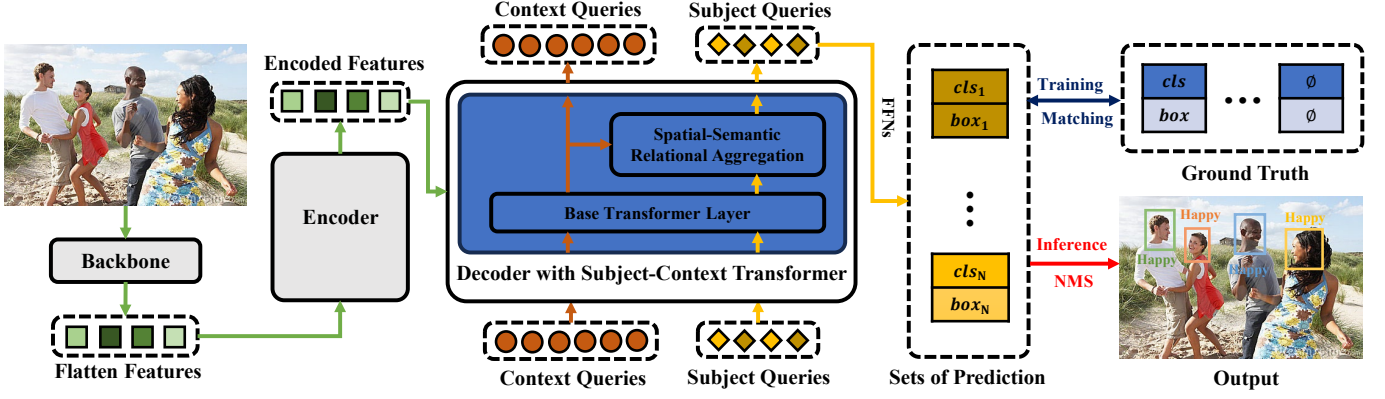


Fig. 3. Overall pipeline of VERT. Following [67], we extract multi-scale features through the backbone. Then, the encoder takes as input the multi-scale features and returns encoded features through the transformer encoder. After that, the decoder with subject-context layers is fed with the encoded features, subject queries, and context queries. The subject queries integrate the context queries through spatial-semantic relational aggregation to capture comprehensive context information. Finally, individual Feed-Forward Networks (FFNs) transform the encoded subject queries into  $N$  sets of emotion class and subject box predictions. In training, we adopt Hungarian Matching [2] 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 applying NMS.

where  $\lambda_{cls}, \lambda_{box} \in \mathbb{R}$  are hyperparameters. For matching and training, we employ the focal loss  $\mathcal{L}_{cls}$  [27] and set  $\mathcal{L}_{box}$  as the  $l_1$  loss and generalized IoU loss [39].

In inference, we first set the mean of the class output logit as the score of each prediction, and then we employ NMS to remove duplicate predictions. For multi-label tasks, subject emotions are determined based on a threshold  $t$ :

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

where  $o$  represents the index list of the emotion class. In the case of multi-class tasks, subject emotions are determined as:

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

where  $o$  corresponds to the index of the emotion class.

**Discussion.** In the VER task, the visual emotions often hinge on subtle interaction between subjects and contexts. To capture these emotions, we pursue an end-to-end framework that is ideally suited for fine-grained interaction and seamless optimization. The DETR pipeline, including the above-mentioned one-stage processing, set-level prediction, and end-to-end training, aligns well with our demand.

#### IV. SUBJECT-CONTEXT TRANSFORMER

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. Instead, we introduce a novel subject-context transformer, which includes context decomposition and spatial-semantic relational aggregation.

**Context Decomposition.** The input queries are divided into subject and context queries. To initiate context queries with accurate and abundant context information, we employ pre-trained queries from existing DETR models. As illustrated in the left section of Fig. 4, the subject query (diamond-shaped)

primarily focuses on the subject area, while the context queries (circle-shaped) are distributed densely across the entire image.

**Spatial-Semantic Relational Aggregation.** To increase the representation of the subject query, we design a spatial-semantic relational aggregation component. The module takes subject and context queries as input and outputs the augmented subject query, as shown in the right part of Fig. 4.

In the first step, we select a subset of relevant context queries, discarding the irrelevant ones based on their spatial proximity to the subject query. Our assumption is that context queries are more valuable when they are in close spatial proximity to the subject query. We calculate the positional distance of the subject 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 distances between the subject query and context queries. We select the top  $K$  context queries that are nearest to the subject query, and the process can be formalized as:

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

where  $D$  represents the subset of selected context queries, which is later used for fusion with the subject query.

In the second step, we fuse the subject query with the selected  $K$  context queries based on semantic relevance, assuming that context queries are more relevant if they share semantic proximity with the subject query. We initially calculate the relevance between the features of the subject and context queries using inner products and apply a softmax function for normalization. These relevance values are employed to re-weight and aggregate all the context queries. Finally, we add the subject query to the aggregated context queries, and this operation can be expressed as:



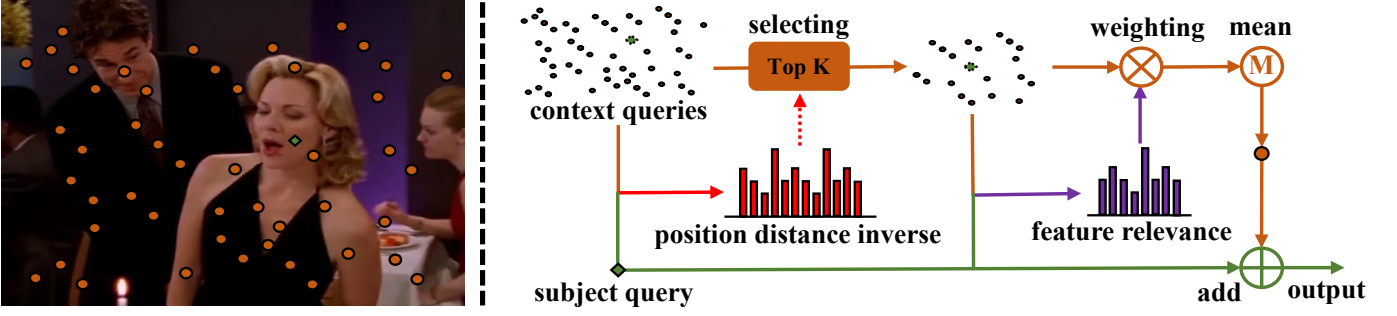


Fig. 4. Illustration of spatial-semantic relational aggregation module. The diamond and circle refer to the subject and context queries. The object with dotted lines is imaginary for clarity. The left figure shows the subject query attends to the agent area while the context queries are distributed across the images. The right part describes how the subject query integrates context queries through spatial-semantic relations. We select a subset of context queries with the smallest spatial distance from subject queries and attentively aggregate this subset based on the semantic relevance between subject and context queries. Ultimately, the subject query integrates the merged context query to capture sufficient contextual information.

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

where  $a \in \mathbb{R}^{(1,d)}$  represents 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.

**Discussion.** In the context of object detection, queries typically focus on localized information, as objects are distinct and well-defined. In contrast, in the task of VER, emotional cues can be dispersed across the entire context of an image. To effectively capture these distributed emotional cues, we employ a decomposition and fusion strategy, dividing the input queries into context and subject queries. This approach empowers context queries to capture contextual information more effectively and increases the subject queries' representation.

## V. EXPERIMENTS

### A. Implements

In our experiments on CAER-S [20] and EMOTIC [18], we set the number of queries  $N$  to 4. The weights of VERT and context queries are initialized using Deformable DETR [67], which was pre-trained on COCO [28]. Our batch size is 32, and we use hyperparameters  $\theta_{box}$ ,  $\lambda_{box}$ ,  $\theta_{cls}$ , and  $\lambda_{cls}$  set to 5, 5, 2, and 5, respectively. The experiments were conducted using 8 GPUs of the NVIDIA Tesla A6000. All other architectural configurations, training strategies, and preprocessing steps follow those outlined in [67]. During evaluation, we compare the ground truth with the output subject that exhibits the highest bounding box overlap with the ground truth.

### B. Datasets

We conducted extensive experiments on two popular VER datasets: CAER-S [20] and EMOTIC [18].

The CAER-S dataset consists of 70,000 images, randomly divided into training (70%), validation (10%), and testing (20%) sets. Annotations include face bounding boxes and multi-class emotion labels. The dataset encompasses seven

emotion categories: Surprise, Fear, Disgust, Happiness, Sadness, Anger, and Neutral. Performance on this dataset is measured using overall accuracy [20].

The EMOTIC dataset [18] 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. Annotations include body and head bounding boxes, as well as multi-label emotion categories. EMOTIC encompasses 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. Performance on EMOTIC is evaluated based on the mean Average Precision (mAP) for all classes [18].

### C. Quantitative and Qualitative Results

The performance of various methods on CAER-S and EMOTIC datasets is presented in Table I and Table II. To facilitate a fair comparison, we categorize the methods into two groups: similar-parameter measures and significantly larger-parameter ones, with a line separating them in the tables. Additionally, the rightmost column presents detailed network configurations used for parameter calculations. Subscripts in EmotiCon [32] and HECO [49] correspond to specific context modalities as mentioned in their respective papers. The performance of the methods is sourced from their original papers or re-implemented results of other papers (indicated with \*). Our VERT framework outperforms similar-parameter methods by a notable margin, achieving a significant 3.39% improvement on CAER-S and an impressive 6.46% boost on EMOTIC. Notably, VERT even surpasses larger-parameter approaches on both datasets, underscoring its suitability for the VER task when compared to multi-stage and multi-branch methods.

We present the qualitative results in Fig. 5 and Fig. 6, depicting the bounding boxes and emotion classes output by our VERT framework, alongside those produced by the EMO-Net [18], a representative multi-stage multi-branch method. To enhance the clarity of VERT's output, we include visual indicators for subject queries' reference points (colored in red) and sampling locations. Outputs of different subjects are

Methods	Acc (%)	Param. (M)	Details
VERT-R18 (ours)	84.96	22	ResNet18
CAER-Net-S [20]	73.51	22	12-layers CNN
GNN-CNN [59]	77.21	23	VGG16
EfficientFace [65]	85.87	25	MobileNet28, ResNet18
EMOT-Net [18]	74.51	32	ResNet18 x 2
SIB-Net [26]	74.56	33	ResNet18 x 3
<b>VERT-R50 (ours)</b>	<b>91.81</b>	39	ResNet50
GRERN [11]	81.31	45	ResNet101
RRLA [25]	84.82	48	ResNet50, RCNN50
MA-Net [64]	88.42	52	Multi-Scale ResNet18
EmotiCon [32]	88.65	181	OpenPose, RobustTP, Megadepth
VRD [14]	90.49	380	{VGG19, ResNet50, FRCNN50} × 2
CCIM [48]	91.17	223	OpenPose, RobustTP, Megadepth, ResNet101

Table I. Performance of different approaches on the CAER-S [20].

Methods	mAP (%)	Param. (M)	Details
VERT-R50 (ours)	37.26	39	ResNet50
<b>VERT-R101 (ours)</b>	<b>37.81</b>	58	ResNet101
EMOT-Net [18]	27.93	84	YOLO, ResNet18*2
CAER-Net [18]	20.84	84	YOLO, 12-layers CNN
EmotiCon <sub>(1)</sub> [32]	31.35	85	OpenPose, 15-layers CNN
EmotiCon <sub>(1,3)</sub> [32]	35.28	108	OpenPose, 15-layers CNN, Medadepth
HECO <sub>(1)</sub> [49]	22.25	135	YOLO, Alphapose, ResNet18
EmotiCon <sub>(1,2)</sub> [32]	32.03	139	OpenPose, RobustTP, ResNet18
HECO <sub>(1,2)</sub> [49]	36.18	146	YOLO, Alphapose, ResNet18 × 2
HECO <sub>(1,2,3)</sub> [49]	34.93	173	YOLO, Alphapose, ResNet50, ResNet18 × 2
EmotiCon <sub>(1,2,3)</sub> [32]	32.03	183	OpenPose, RobustTP, ResNet18, Megadepth
GCN-CNN [59]	28.16	189	YOLO, VGG16, 6-layers GCN
HECO <sub>(1,2,3,4)</sub> [49]	37.76	219	YOLO, Alphapose, {ResNet18, ResNet50} × 2, Faster RCNN50
EmotionCLIP [60]	32.91	577	YOLO, ViT_b_32

Table II. Performance of existing methods on the EMOTIC [18].

color-coded for differentiation. These visualizations illustrate that VERT consistently yields high-quality results, showcasing superior classification accuracy when compared to EMO-Net.


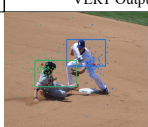
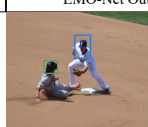
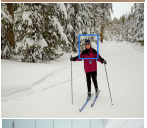


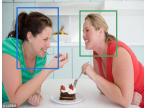

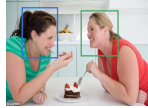
Ground Truth	VERT Output	EMO-Net Output
		
		
		

Fig. 5. The output comparison of VERT and EMO-Net on EMOTIC [18].

#### D. Why the End-to-End Framework Enhances Performance?

In this section, we conduct an evaluation on images with varying subject numbers and visualize feature maps of methods of different paradigms. We aim to investigate the reasons behind the performance boost of the end-to-end framework. We select EMO-Net [18] as a representative multi-stage multi-

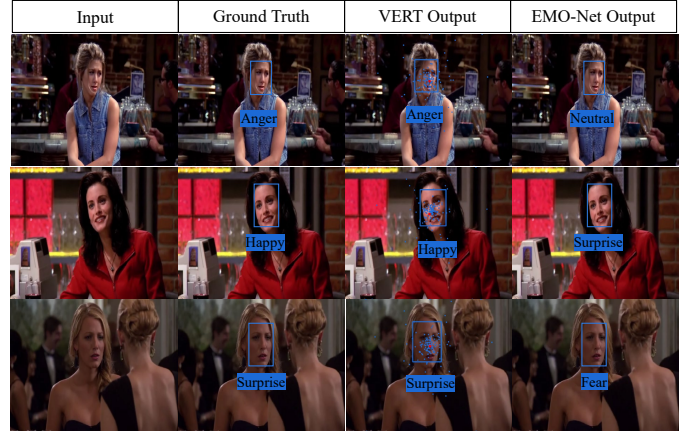


Fig. 6. The output comparison of VERT and EMO-Net on CAER-S [20].

Subject #	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.

branch method. For a fair comparison, we re-implement EMO-Net using ResNet50 as the backbone to achieve similar parameters to VERT (referred to as EMO-Net-R50).

Table III presents the performance on EMOTIC for images with different subject numbers. We also include results of the method that masks subjects in the context [20] for comparison (denoted as EMO-Net-R50-M). As the subject number in an image increases, the complexity and subtle interaction of the context also rise. Notably, VERT maintains stable performance with increasing subject numbers, while EMO-Net-R50's performance deteriorates. This observation can be attributed to the shortcomings of multi-stage multi-branch methods in handling complex contexts and capturing subtle interactions.

In Fig. 7, we illustrate the feature maps generated by EMO-Net-R50 and VERT. The selected feature maps originate from the final layer of ResNet50. A clear distinction emerges: EMO-Net-R50 exhibits a tendency to emphasize a few large regions, while VERT consistently places importance on smaller, more intricate areas. This observation suggests that VERT excels in the precise handling of contextual information compared to conventional multi-stage multi-branch methods, thereby leading to superior performance in the VER task.

#### E. Why the Context Decomposition and Fusion Works?

In this section, we analyze the effectiveness of context decomposition and fusion strategy on the EMOTIC dataset. Specifically, we compare the performance and visualization of two approaches: expansion-selection and query decomposition with fusion. In the expansion-selection approach, the number of queries is significantly increased to 300 to cover the global context, and the most distinct query is chosen for the final prediction. Instead, our proposed query decomposition and fusion method utilizes 300 decoupled context queries to capture global contexts and integrate them into each subject query. As

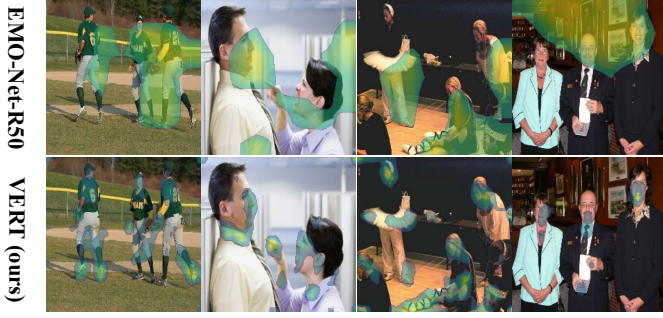


Fig. 7. The visualization of feature maps from VERT and EMO-Net-R50.

depicted in Fig. 8, the queries and the performance of these two strategies are illustrated. The subject queries are drawn as big circles with different colors and the context queries as small circles with yellow. Experimental results reveal that the decomposition-fusion strategy leads to a performance improvement of 0.71%, while the expansion-selection method results in a decline in accuracy of 4.69%. This suggests that, in contrast to object detection tasks where objects are distinct and well-defined, emotional cues are distributed throughout images. Therefore, fusing the decoupled contexts enhances the performance of VER. Additionally, the observed deterioration in the expansion-selection strategy can be attributed to the presence of numerous queries competing for optimization.

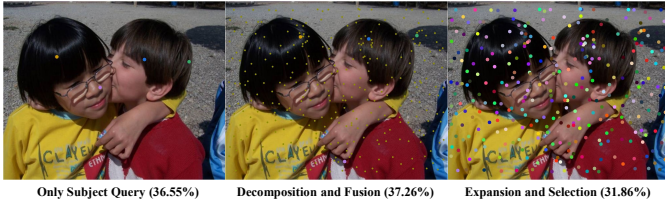


Fig. 8. The comparison of different query processing strategies.

#### F. Ablation Study

**Subject Query Number.** We conduct experiments to investigate the impact of query number setting. Table IV presents the performance of different query numbers. Fig. 9 offers a visualization of bounding boxes, reference points, and sampling locations corresponding to different subject query numbers. Here, the reference points are depicted in red, while different subject queries are color-coded for differentiation.

Table IV shows the VERT achieves the best result on EMOTIC and CAER-S when the query number is 4 and 9 respectively. Notably, we observe a consistent performance stability trend as the query number increases, ranging from one to six on EMOTIC and from one to ten on CAER-S. This phenomenon can be attributed to the intrinsic nature of visual emotions. Unlike distinct, well-defined objects, visual emotions are dispersed globally across an image. Hence, a single query possesses the capacity to capture comparable emotional information to multiple queries distributed in different places. Fig. 9 illustrates that the queries are strategically distributed to encompass distinct regions and subjects within the image.



Fig. 9. The query position visualization of different query numbers.

Set Number	1	2	3	4	5
EMOTIC (mAP %)	37.14	36.71	36.61	<b>37.26</b>	36.70
CAER-S (Acc %)	91.78	91.57	91.39	91.57	91.47
Set Number	6	7	8	9	10
EMOTIC (mAP %)	36.97	36.03	36.26	35.92	35.68
CAER-S (Acc %)	91.52	91.59	91.42	<b>91.81</b>	91.44

Table IV. The query number study on EMOTIC [18] and CAER-S [20].

**Coefficient of Classification.** Visual motion recognition and object detection are fundamentally distinct tasks, with the former emphasizing classification accuracy and the latter emphasizing box accuracy. Therefore, optimizing the coefficient of classification becomes a pivotal consideration for achieving superior performance. To this end, we conducted experiments to fine-tune  $\theta_{cls}$  and  $\lambda_{cls}$  on the EMOTIC dataset [18]. The results of these hyperparameter experiments are thoughtfully presented in Table V. Notably, the most compelling performance is achieved when  $\theta_{cls}$  is set to 2, and  $\lambda_{cls}$  is set to 5. In contrast to the original configuration of deformable DETR, VERT assigns greater importance to classification. Additionally, we noticed that  $\theta_{cls}$  has minimal impact on performance, whereas  $\lambda_{cls}$  significantly influences the results.

$\theta_{cls}$	5	10	15	2	2	2
$\lambda_{cls}$	2	2	2	5	10	15
mAP (%)	35.41	35.89	35.41	<b>36.01</b>	34.99	34.64
$\theta_{cls}$	2	5	10	15	1	1
$\lambda_{cls}$	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.

**Subject-Context Transformer.** The proposed subject-context transformer consists of decoupled context queries and spatial-semantic relational aggregation. We assess their impact through experiments on the EMOTIC dataset [18], and the results are displayed in Table VI. The categories include “Subject” (VERT with only subject queries), “Context” (decoupled context queries without fusion), “Fusion” (fusing subject and mean context queries), “Spatial” (selecting top K context queries based on spatial relation), and “Semantic” (re-weighting context queries with semantic relation).

As we can see, the subject-context transformer significantly enhances performance by 0.71%, highlighting the importance of sufficient contextual information. Specifically, fusing context queries improves performance by 0.18%, re-weighting context queries with semantic relation boosts the result by an extra 0.08%, and selecting the top K context queries based on



Subject	Context	Fusion	Spatial	Semantic	mAP (%)
✓	×	×	×	×	36.55
✓	✓	×	×	×	36.21
✓	✓	✓	×	×	36.73
✓	✓	✓	×	✓	36.85
✓	✓	✓	✓	✓	<b>37.26</b>

Table VI. The performance of different components combination.

spatial relation contributes a substantial 0.41% gain. However, adding decoupled context queries without fusion introduces disturbance and leads to a 0.34% deterioration. The results demonstrate the effectiveness of each proposed component.

**Spatial Relational Selection.** In this study, we investigated the setting of spatial relational selection by conducting experiments with varying values of “K” on the EMOTIC [18]. The results, presented in Table VII, indicate that the optimal performance is achieved when “K” is set to 100. The best performance is 0.41% higher than the one when “K” is 300 (no selection). This observation suggests that not all contextual information is equally valuable for effective emotion recognition. The selection of the top 100 context queries based on spatial relations appears to be a suitable strategy for this task.

K	25	50	100	150	200	250	300
mAP %	37.01	37.25	<b>37.26</b>	37.01	36.94	36.91	36.85

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

**Semantic Relational Re-weighting.** We explored various re-weighting strategies for context queries and conducted experiments on the EMOTIC dataset. The results, shown in Table VIII, indicate the performance under different re-weighting strategies: Semantic (based on feature relevance with the subject query), Equal (equal values), Spatial (based on position distance inverse with the subject query), and Attentive (attentive weights from learning). The findings reveal that the best performance is achieved when employing semantic relational re-weighting. Conversely, the spatial relational re-weighting approach performs less effectively than the mean strategy. This can be attributed to the dispersed nature of contextual emotions, which often exhibit limited spatial correlation with the subject. In conclusion, the semantic relational re-weighting strategy emerges as the most reliable choice for suppressing noise and enhancing valuable information.

Strategy	Semantic	Equal	Spatial	Attentive
mAP %	<b>37.26</b>	36.73	36.48	36.93

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

**Feature Extractor.** To evaluate the influence of feature extractors, we conducted experiments with various backbone architectures for VERT on both EMOTIC and CAER-S datasets. The results, as depicted in Table IX, include performance metrics and corresponding parameter counts. Specifically, “R” refers to ResNet [13], and “WR” designates Wide ResNet [52]. For ResNet50, we utilized a pre-trained backbone from deformable DETR as the initialization. The optimal performance on EMOTIC is attained when using ResNet-101 as the backbone, while on CAER-S, ResNet-

50 yields the best results. Interestingly, it’s evident that the relationship between performance and parameter counts is not linear on both datasets. Moreover, the performance on CAER-S seems to be more influenced by the choice of pre-trained initialization rather than the number of parameters.

Backbone	Param. (M)	EMOTIC (mAP %)	CAER-S (Acc %)
R18	22	35.68	84.98
R34	33	34.40	84.52
R50	39	37.26	<b>91.81</b>
R101	58	<b>37.81</b>	85.39
R152	74	37.32	83.88
WR50	82	34.46	-
WR101	140	34.07	-

Table IX. Performance of different backbones on the EMOTIC and CAER-S.

**Multiple Modalities.** In some studies, the incorporation of multiple modalities has been proposed to enhance context-based emotion recognition [32], [49]. To investigate the potential benefits of including additional modalities in VERT, we conducted experiments on the EMOTIC dataset. Specifically, we introduced three modalities: “Scene,” “Semantic,” and “Instance” corresponding to scene classification, semantic segmentation, and instance segmentation, respectively. The networks employed for these modalities are Places365 [66], Deeplabv3 [3], and MaskRCNN [12], all featuring a ResNet50 backbone. We extracted multi-scale features from these networks and integrated them with VERT’s features while keeping the parameters of the other modality networks frozen. The results, as summarized in Table X, indicate that the inclusion of additional modalities leads to a decline in accuracy. This suggests that introducing other modalities might introduce noise or redundancy to VERT, which is already adept at capturing fine-grained information for visual emotion recognition.

Modality	None	Scene	Semantic	Instance
mAP %	<b>37.26</b>	32.08	34.01	34.11

Table X. The result of adding different modalities for VERT.

**DETR-like Architecture.** The family of DETR-like architectures has gained significant momentum in the object detection task. To assess the impact of incorporating different DETR-like architectures into VERT, we conducted experiments on the CAER-S dataset, leveraging the detrax platform [38]. The results, summarized in Table XI, reveal the performance of VERT with various DETR-like architectures. This exploration provides insights into how different architecture choices influence VERT’s performance.

Architecture	Acc %	Architecture	Acc %
DETR	89.74	Deformable DETR	91.81
Anchor-DETR	-	Conditional-DETR	-
DAB-DETR	-	DN-DETR	-
DINO (RN)	-	DINO (Swin)	-
H-Deformable-DETR	-	DETA	-

Table XI. The result of adopting different DETR architectures for VERT.

## G. Conclusion

This paper introduces VERT, an end-to-end Visual Emotion Recognition system with subject-context Transformers. By



overcoming the constraints of traditional pipelines, VERT enables seamless optimization and fine-grained feature interaction, leading to superior performance on two popular datasets. The results underscore the significance of capturing fine-grained interaction information in contexts for effective VER. Future work will extend these insights to explore subtle information within video datasets.

## REFERENCES

- [1] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. Pstr: End-to-end one-step person search with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9458–9467, 2022.
- [2] 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.
- [3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, and Thibaut Ehret. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008.
- [8] 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.
- [9] 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.
- [10] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.
- [11] Qinquan Gao, Hanxin Zeng, Gen Li, and Tong Tong. Graph reasoning-based emotion recognition network. *IEEE Access*, 9:6488–6497, 2021.
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [14] 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.
- [15] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. Istr: End-to-end instance segmentation with transformers. *arXiv preprint arXiv:2105.00637*, 2021.
- [16] 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.
- [17] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19578–19587, 2022.
- [18] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *TPAMI*, 42(11):2755–2766, 2019.
- [19] 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.
- [20] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *CVPR*, pages 10143–10152, 2019.
- [21] 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.
- [22] 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.
- [23] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.
- [24] 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.
- [25] Weixin Li, Xuan Dong, and Yunhong Wang. Human emotion recognition with relational region-level analysis. *IEEE Transactions on Affective Computing*, 2021.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] Tianhe Ren, Shilong Liu, Feng Li, Hao Zhang, Ailing Zeng, Jie Yang, Xingyu Liao, Ding Jia, Hongyang Li, He Cao, Jianan Wang, Zhaoyang

- Zeng, Xianbiao Qi, Yuhui Yuan, Jianwei Yang, and Lei Zhang. detrex: Benchmarking detection transformers, 2023.
- [39] 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.
- [40] Philipp V Rouast, Marc Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 2019.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] Weijie Wang, Nicu Sebe, and Bruno Lepri. Rethinking the learning paradigm for facial expression recognition. *arXiv preprint arXiv:2209.15402*, 2022.
- [46] 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.
- [47] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021.
- [48] Dingkan 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.
- [49] Dingkan 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.
- [50] 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.
- [51] 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.
- [52] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.
- [67] 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.