

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

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

Abstract—The typical visual emotion recognition system involves multiple stages: subject detection, feature extraction, and emotion classification. However, this system may suffer from accumulative errors due to disjoint optimization and limited subject-context feature interaction. To address these limitations, this paper introduces a novel end-to-end Visual Emotion Recognition framework with subject-context Transformers (VERT). The VERT adopts a transformer-based architecture and predicts subjects' emotions and locations simultaneously. In detail, the transformer processes both subjects and contexts from low-level pixels to high-level features, enabling fine-grained feature interactions. Predicting emotions and locations simultaneously streamlines the multi-stage pipeline into an end-to-end one and eliminates handcrafted procedures, allowing for seamless optimization. Furthermore, to enhance the system's ability to capture contextual emotional cues, we propose the subject-context transformer that encompasses context decoupling and fusion. Specifically, the context queries are decoupled from subject queries and initialized with pre-trained queries of generic object detection for accurate and sufficient context hunting. Then, the subject queries integrate the valuable context information through spatial-semantic relational aggregation. Without bells and whistles, the VERT demonstrates its superiority on two popular datasets. Particularly, we achieve 91.81% 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.

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], [41], [42], [55], reflecting the fundamental role of emotions in daily communication [10]. VER has gained substantial attention within the multimedia community, particularly in the deep learning era [8], [22], [23], [26], [39], [45], [63].

Traditionally, the VER system comprises a sequence of key steps, including subject detection, feature extraction, and emotion classification. These steps predominantly revolve around the interpretation of facial expressions, featuring a multi-stage single-branch paradigm. As depicted in Fig. 1(a), 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. Research in this paradigm has primarily addressed challenges such as label uncertainty [5], [6], [20], [34], [46],

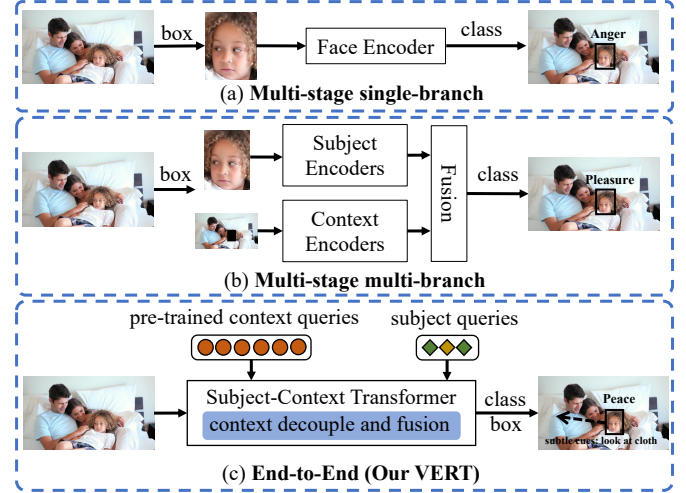


Fig. 1. Comparison of our VERT architecture with traditional multi-stage paradigms. (a) In the multi-stage single-branch pipeline, subject boxes are first detected and emotion classes are then inferred. (b) Within the multi-stage multi-branch paradigm, subject and context are detected, then encoded by different streams and fused together. (c) Our VERT is an end-to-end framework that adopts a transformer architecture, with subject-context transformers that contain context decouple and fusion, and predicts subjects' emotions and locations simultaneously. VERT can notice subtle cues, e.g. looking at a cloth, and provide a more accurate prediction than its multi-stage counterparts.

[49], [50], [56], [67], [68], micro expression [38], [60], and disentangled representation [58], [66].

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 is characterized as a multi-stage multi-branch paradigm. As illustrated in Fig. 1(b), it first identifies subjects and contexts within the image, then processes them through independent encoders, and subsequently integrates the resulting features or scores with a late fusion. Research mainly focuses on issues of dataset establishment [19], [21], emotional context interpretations [36], [54], and context processing [28], [53].

The two paradigms might suffer from accumulative errors and deficient feature interaction between subjects and contexts. The first issue stems from the disjoint nature of the multi-stage pipeline, including the separate detectors and encoders and non-differentiable handcrafted procedures like anchor design. Another concern pertains to the multi-branch strategy that independently preprocesses subjects and contexts and then delicately fuses them with a late fusion module.

We advocate for the development of an end-to-end frame-

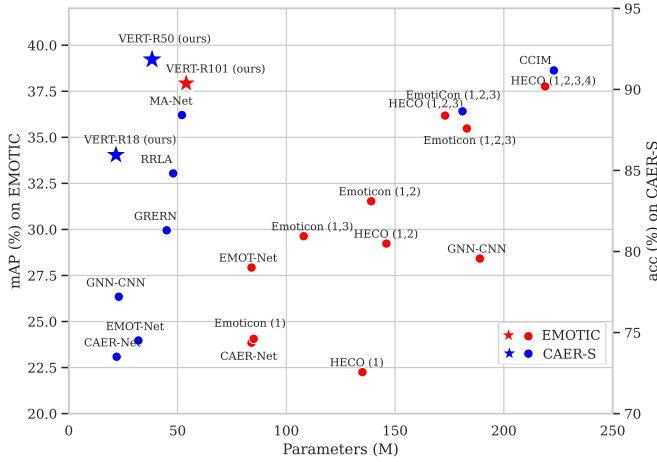


Fig. 2. Performance vs. model efficiency on EMOTIC and CAER-S. VERT achieves state-of-the-art performance with fewer parameters than prior arts.

work 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 [27]. Secondly, capturing contextual emotion cues is so challenging that it requires strong and delicate off-the-shelf detectors, resulting in a late fusion of contexts [36], [53], [54].

In response to these challenges, we introduce a novel end-to-end framework, VERT (Visual Emotion Recognition with subject-context Transformers). This framework aims to provide seamless optimization and fine-grained subject-context feature interaction. To this end, the VERT adopts a transformer architecture and predicts subjects' emotions and locations simultaneously. Illustrated in Fig. 1(c), the VERT can notice subtle cues, e.g. looking at a cloth, and provide a more accurate prediction than its multi-stage counterparts. Specifically, predicting emotions and locations simultaneously streamlines the multi-stage pipeline into an end-to-end one and eliminates handcrafted procedures, allowing for seamless optimization. Furthermore, the transformer processes both subjects and contexts from low-level pixels to high-level features, enabling fine-grained feature interactions. As depicted in Fig. 2, experimental results demonstrate that this end-to-end architecture surpasses multi-stage counterparts with fewer parameters.

Even though the end-to-end system is effective and efficient, it is still desirable to enhance the system's ability to capture sufficient contextual emotional cues since many studies show adding more emotional contexts boosts performance [19], [21], [36], [53], [54]. In this work, we introduce the effective and efficient component named subject-context transformer. The proposal encompasses context decomposition and spatial-semantic relational aggregation. The decomposition aims to help delineate more accurate and sufficient context information. As depicted in Fig. 1(c), the input queries include the decoupled context queries, which are initialized with pre-trained queries of common object detection, and subject queries. After being fed into the subject-context transformer,

the subject queries integrate the information of context queries to enhance representation. The spatial-semantic relational aggregation intends to filter the context queries to have more valid information and less disturbance. Specifically, we select the context queries with the smallest spatial distance from subject queries. Then we attentively aggregate them based on their semantic relevance with subject queries. Our assumption is that context queries are more valuable when they are in close spatial and semantic proximity to the subject query. Ultimately, the subject queries aggregated with the context information are used for the final prediction. This context decomposition and spatial-semantic relational fusion strategy have proven to be highly effective in capturing informative contexts.

Extensive experiments conducted on two widely-recognized benchmarks validate the efficacy of our approach. The VERT framework attains impressive results, achieving 91.81% on the CAER-S dataset [21] and 37.81% on EMOTIC [19]. Notably, on CAER-S, VERT not only surpasses similar-parameter methods by a substantial margin of 3.39% but also outperforms larger-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 larger-parameter counterparts. Furthermore, we provide valuable insights by visualizing network output, feature activation, and query positions, and by conducting ablation studies. These findings underscore VERT can discern meaningful and nuanced emotional cues within context through an end-to-end pipeline and a context decoupling and fusion strategy.

Our work makes several significant contributions:

- We introduce a fresh analysis of current multi-stage methods in VER, emphasizing the critical need for an end-to-end framework that can provide seamless optimization and nuanced feature interaction between subjects and contexts.
- We present the VERT framework, the first end-to-end architecture for the VER task. The simultaneous prediction of emotions and locations and the transformer-based architecture enable seamless training and fine-grained interaction.
- We propose the subject-context transformer, featuring context decomposition and spatial-semantic relational aggregation. The subject-context transformer is demonstrated to be excellent in capturing contextual emotional cues.
- Our empirical findings underscore the superiority of VERT over multi-stage alternatives 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], [26], [39], [45], [63]. 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], [20], [34], [46], [49], [50], [56], [67], [68], micro expressions [38], [60], and disentangled

representations [58], [66]. 2) *Context-Aware VER*. 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 [19], [21], [28], [36], [37], [47], [53], [54], [65]. 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: [19] suggests using the whole scene image, [21] recommends masking the face in the scene image, [36] extends this approach by incorporating inter-agent interactions, and [54] adds agent-object interactions. *Despite the effectiveness of these methods, they all share a common multi-stage process involving detection and feature extraction, sometimes 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-context interaction.*

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], [48], query [32], [51], [72], and bipartite matching [4], [24], [25], [61], [62]. Particularly, Deformable DETR [72] replaces original transformer attention with deformable attention, showing high efficacy and efficiency. The original DETR framework, along with its various adaptations, has not only brought forth a simple yet powerful end-to-end architecture for common object detection but has also been extended to other related tasks, including multiple-object tracking [59], action detection [33], human-object interaction [17], [18], person search [1], and instance segmentation [15], [52]. *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, as generic objects exhibit distinct and localized characteristics, but visual emotion might be ambiguous and scattered across the image, we introduce the subject-context transformer to capture sufficient contexts.*

III. VERT ARCHITECTURE

Current multi-stage approaches in VER may encounter cumulative errors stemming from disjoint optimization and insufficient subject-context feature interaction from late fusion. To address the limitations, we introduce a pioneering end-to-end framework, VERT. As depicted in Fig. 3, VERT processes the entire image directly with a transformer encoder-decoder architecture and predicts subject bounding boxes and emotion classes simultaneously with a set prediction strategy, facilitating seamless optimization and intricate feature interaction.

Fine-grained Feature Interaction. Based on deformable DETR [72], our VERT system consists of three main com-

ponents: 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 context's emotions) across 6 subject-context transformer layers. The Feed-Forward Networks (FFNs) generate a set of N subject predictions, including their corresponding classes and boxes, based on updated subject queries. Intuitively, the transformer processes the entire image from low-level pixels to high-level features, enabling fine-grained subject-context feature interactions.

End-to-End Optimization. To achieve end-to-end training, we adopt a set-level prediction approach, encapsulating several predictions or ground truth within a set. One element of the set includes 1) the coordinate of the subject bounding box and 2) the corresponding emotion class. For clarity, we denote the i -th element of ground truth set as $y_i = (c_i, b_i)$, where c_i represents the target emotion label and $b_i \in [0, 1]^4$ 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 a set of N predictions, where N is typically larger than the number of subjects in an image.

During training, since the prediction number is larger than the actual number of subjects in an image, we first pad the set of ground truths with \emptyset to ensure a consistent size. Following [2], we employ the bipartite matching that computes one-to-one associations between the set of predictions \hat{y} and the padded ground truths y by the following equation:

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

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}_{cls} and a box regression loss \mathcal{L}_{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:

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

where $\lambda_{\text{cls}}, \lambda_{\text{box}} \in \mathbb{R}$ are hyperparameters. For matching and training, we employ the focal loss \mathcal{L}_{cls} [30] and set \mathcal{L}_{box} as the l_1 loss and generalized IoU loss [44].

During inference, we first set the mean of the class output logit as the score of each prediction, and then we employ Non-Maximum Suppression (NMS) to remove duplicate predictions. For multi-label tasks, subject emotions are determined by the following equation with a threshold t :

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

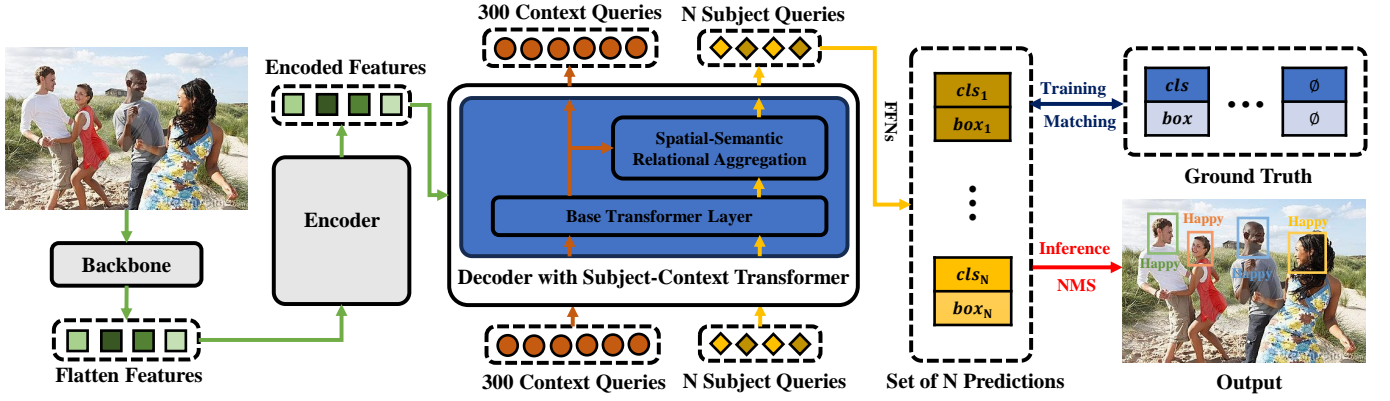


Fig. 3. Overall pipeline of VERT. The entire image is processed by the backbone and encoder, producing encoded features with fine-grained subject-context feature interaction. Then, the decoder is fed with the encoded features, subject queries, and context queries. In the decoder, the subject queries integrate the context queries through spatial-semantic relational aggregation to capture comprehensive context information. Finally, individual FFNs transform the encoded subject queries into a set of N emotion class and subject box predictions. During training, we adopt Hungarian Matching between predictions and padded ground truths to achieve end-to-end training. During inference, we output the high-quality positions and emotions of subjects after applying NMS.

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 subject-context 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 contexts and attentive learning to fuse them. As depicted in Fig. 3, we introduce a novel subject-context transformer, which includes context decomposition and spatial-semantic relational aggregation.

Context Decomposition and Fusion. To improve our system's ability to capture contextual cues, the subject-context transformer adopts a context decomposition and fusion strategy. The context queries are decoupled from subject queries and initialized with pre-trained queries from common object detection, allowing the easier capture of accurate and sufficient contextual information. 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. The subject queries later integrate the information of context queries.

Spatial-Semantic Relational Aggregation. To enhance the valuable information and avoid the disturbance from the context queries, we design a spatial-semantic relational aggregation component, 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. Let vector s and C_i represent the position of the subject query and the i -th point in the context query set C respectively. The spatial distance d_i of the subject and context queries is calculated as:

$$d_i = \|s - C_i\|_2, \quad (6)$$

where d is the distance vector that contains the spatial distances between the subject query and context queries. We select the top K context queries that are nearest to the subject query, which is represented by D .

Then, we integrate the subject query with the selected context queries D based on semantic relevance, assuming that context queries are more relevant if they share semantic proximity with the subject query. We calculate the semantic relevance by using inner products and applying a softmax function for normalization. These values are employed to re-weight and aggregate the context queries, which will be added to the subject query. The operation can be expressed as:

$$b = \text{softmax}(a [D_1^T, \dots, D_n^T]) \begin{bmatrix} D_1 \\ \vdots \\ D_n \end{bmatrix} + a, \quad (7)$$

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 $b \in \mathbb{R}^{(1,d)}$ is the feature of the output subject query.

Discussion. In the object detection task, queries can capture effective information from the distinctive and localized areas. While in the task of VER, emotional cues can be dispersed across the entire image as contexts. To take advantage of the context information, we leverage the decouple context queries to capture sufficient and accurate contextual emotion cues.

V. EXPERIMENTS

A. Implement

We set the number of queries N to 4 and 9 for CAER-S [21] and EMOTIC [19]. The weights of VERT and context

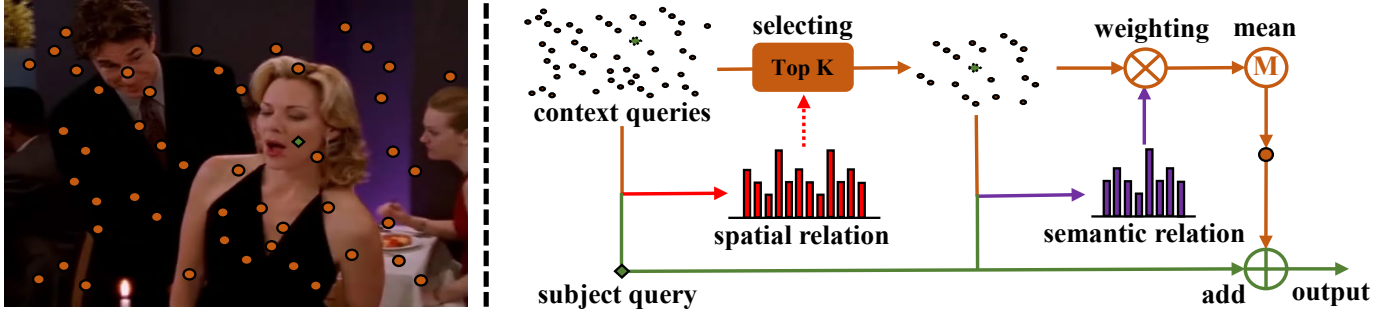


Fig. 4. Illustration of spatial-semantic relational aggregation module. The diamond and circle refer to the subject and context queries. The object with dotted lines is imaginary for clarity. The left figure shows the subject query attends to the agent area while the context queries are distributed across 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 and valuable contextual information.

queries are initialized using Deformable DETR [72], which was pre-trained on COCO [31]. Our batch size is 32, and we set hyperparameters θ_{box} , λ_{box} , θ_{cls} , and λ_{cls} 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 [72]. 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 [21] and EMOTIC [19].

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 (acc) [21].

The EMOTIC dataset [19] 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 [19].

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 based on the number of parameters: similar-parameter measures and larger-parameter ones, using a threshold of 100 Million (M) parameters. Additionally, the rightmost column presents detailed network configurations used for parameter

Methods	Acc (%)	Param. (M)	Details
With <100M parameters			
VERT-R18 (ours)	84.96	22	ResNet18
CAER-Net-S [21]	73.51	22	12-layer CNN
GNN-CNN [64]	77.21	23	VGG16
EfficientFace [70]	85.87	25	MobileNet28, ResNet18
EMOT-Net [19]	74.51	32	ResNet18 \times 2
SIB-Net [29]	74.56	33	ResNet18 \times 3
VERT-R50 (ours)	91.81	39	ResNet50
GRERN [11]	81.31	45	ResNet101
RRLA [28]	84.82	48	ResNet50, RCNN50
MA-Net [69]	88.42	52	Multi-Scale ResNet18
With >100M parameters			
EmotiCon [36]	88.65	181	OpenPose, RobustTP, Megadept
VRD [14]	90.49	380	{VGG19, ResNet50, FRCNN50} \times 2
CCIM [53]	91.17	223	OpenPose, RobustTP, Megadept, ResNet101

Table I. Performance of different approaches on the CAER-S [21].

calculations. Subscripts in EmotiCon [36] and HECO [54] 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. 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, 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 [19], 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 color-coded for differentiation. These visualizations illustrate that VERT consistently yields high-quality results, showcasing superior classification accuracy when compared to EMO-Net. In EMOTIC, the EMO-Net might neglect subtle emotions like "Pain", or produce wrong even opposite emotions like "Disapproval". In CAER-S, when the facial expression is not distinguishing, the EMO-Net is confused with "Anger" and "Neutral", "Happy" and "Surprise", or "Surprise" and "Fear".

Methods	mAP (%)	Param. (M)	Details
With <100M parameters			
VERT-R50 (ours)	37.26	39	ResNet50
VERT-R101 (ours)	37.81	58	ResNet101
EMOT-Net [19]	27.93	84	YOLO, Alphaspose, ResNet18 $\times 2$
CAER-Net [19]	20.84	84	YOLO, 12-layer CNN
EmotiCon ₍₁₎ [36]	31.35	85	OpenPose, 15-layer CNN
With >100M parameters			
EmotiCon _(1,3) [36]	35.28	108	OpenPose, 15-layer CNN, Medadepth
HECO ₍₁₎ [54]	22.25	135	YOLO, Alphaspose, ResNet18
EmotiCon _(1,2) [36]	32.03	139	OpenPose, RobustTP, ResNet18
HECO _(1,2) [54]	36.18	146	YOLO, Alphaspose, ResNet18 $\times 2$
HECO _(1,2,3) [54]	34.93	173	YOLO, Alphaspose, ResNet50, ResNet18 $\times 2$
EmotiCon _(1,2,3) [36]	32.03	183	OpenPose, RobustTP, ResNet18, Medadepth
GCN-CNN [64]	28.16	189	YOLO, VGG16, 6-layer GCN
HECO _(1,2,3,4) [54]	37.76	219	YOLO, Alphaspose, {ResNet18, ResNet50} $\times 2$, Faster RCNN50
EmotionCLIP [65]	32.91	577	YOLO, ViT _{b_32}

Table II. Performance of existing methods on the EMOTIC [19].


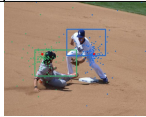
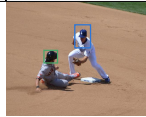



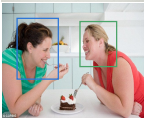

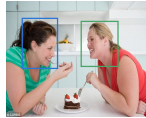
Ground Truth	VERT Output	EMO-Net Output
		
		
		

Fig. 5. The output comparison of VERT and EMO-Net on EMOTIC [19].

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 [19] as a representative multi-stage multi-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 [21] for comparison (denoted as EMO-Net-R50-M). As the subject number in an image increases, the complexity of subtle subject-context interaction also rises. 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 the ResNet50 backbone. 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

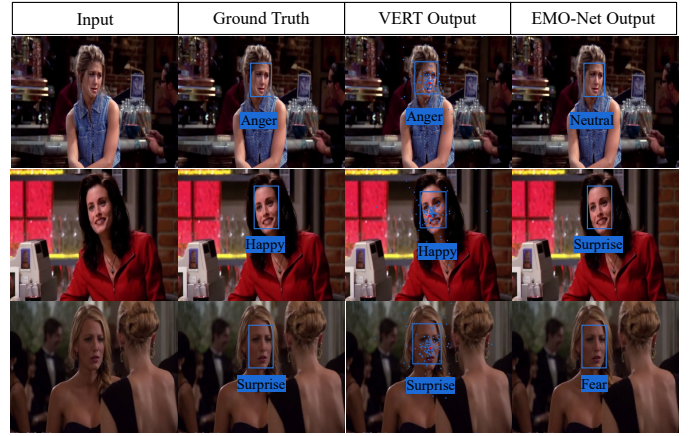


Fig. 6. The output comparison of VERT and EMO-Net on CAER-S [21].

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.

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 three approaches: 4 subject queries, 4 subject and 300 context queries, and 300 subject queries. In the original architecture, the subject queries contain the subject and context information, and the most distinct query is chosen for the final prediction. An intuitive way to capture more context information is to expand the subject queries from 4 to 300. Instead, our proposed query decomposition and fusion method utilizes 300 decoupled context queries to capture global contexts and integrate them into each subject query. The query position and the performance of these two strategies are illustrated in Fig. 8. The subject queries are drawn as big circles with different colors and the context queries as small circles with grey.

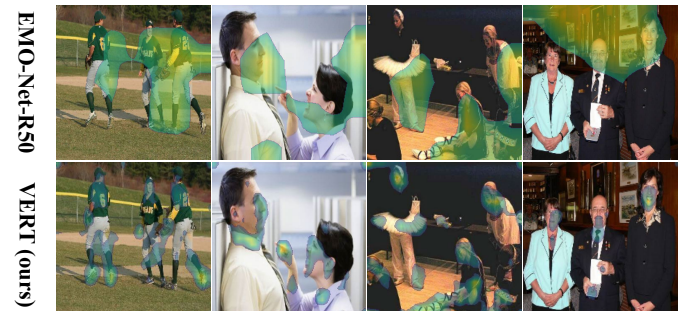


Fig. 7. The visualization of feature maps from VERT and EMO-Net-R50.

Compared to 4 subject queries, with more queries for contexts or subjects, the query positions are densely distributed across the image. 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 1 to 6 on EMOTIC and from 1 to 10 on CAER-S. From Fig. 9, with only a few subject queries, the learned query positions are distributed sparsely however cover the main area with an emotion, explaining the stability of the performance.

Coefficient of Classification. Visual emotion 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 θ_{cls} and λ_{cls} on the EMOTIC dataset [19]. The



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	37.26	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	91.81	91.44

Table IV. The query number study on EMOTIC [19] and CAER-S [21].

results of these hyperparameter experiments are thoughtfully presented in Table V. Notably, the most compelling performance is achieved when θ_{cls} is set to 2, and λ_{cls} is set to 5. In contrast to the original configuration of deformable DETR, VERT assigns greater importance to classification. Additionally, we noticed that θ_{cls} has minimal impact on performance, whereas λ_{cls} significantly influences the results.

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.

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 [19], 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 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 [19]. 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

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

Table VI. The performance of different components combination.

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	37.26	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 semantic relation with the subject query), Equal (equal values), Spatial (based on spatial relation 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 %	37.26	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 [57]. 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	91.81
R101	58	37.81	85.39
R152	74	37.32	83.88
WR50	82	34.46	85.71
WR101	140	34.07	83.16

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 [36], [54]. 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 [71], Deeplabv3 [3], and MaskRCNN [12], all adopting 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 %	37.26	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 by leveraging the detrax platform [43]. For equitable comparisons, all architectures utilize ResNet50 as the backbone. The results, summarized in Table XI, reveal the performance of VERT with various DETR-like architectures. Notably, there is a performance gap of around 1% between DETR-based and deformable DETR-based architectures. However, performances remain comparable within DETR-based and deformable DETR-based architectures even though they adopt different techniques.

Architecture	Acc %	Architecture	Acc %
DETR [2]	89.74	DINO [62]	89.27
Anchor-DETR [51]	90.42	Deformable DETR [72]	91.81
DAB-DETR [9]	87.49	DAB-D-DETR [9]	91.39
DN-DETR [24]	89.79	H-D-DETR [16]	91.65
Conditional-DETR [35]	90.07	DETA [40]	91.77

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. The VERT predicts subjects’ emotions and locations simultaneously with a set prediction strategy and processes the entire image directly with a transformer-based architecture. The proposal achieves 91.81% on CAER-S and 37.81% on EMOTIC, outperforming alternatives with similar parameter numbers by 3.39% on CAER-S and 6.46% on EMOTIC. The remarkable results underscore the significance of seamless optimization and fine-grained interaction information for effective VER. Future work can extend the insights to explore subtle information within video datasets. Additionally, this paper proposes the subject-context transformer that includes context decomposition and spatial-semantic relational aggregation. The quantitative and qualitative results, showing 0.71% improvement in EMOTIC, prove the subject-context transformer excels in capturing accurate and sufficient context information in the context of VER. The decoupling strategy of queries can be extended to information augmentation of other vision tasks.

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 Ehrette. 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] Qinqian 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] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Dets with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023.
- [17] 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.
- [18] 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.
- [19] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *TPAMI*, 42(11):2755–2766, 2019.
- [20] 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.
- [21] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *CVPR*, pages 10143–10152, 2019.
- [22] SangEun Lee, Chaeun Ryu, and Eunil Park. Osanet: Object semantic attention network for visual sentiment analysis. *IEEE Transactions on Multimedia*, 2022.
- [23] Chenchen Li, Jialin Wang, Hongwei Wang, Miao Zhao, Wenjie Li, and Xiaotie Deng. Visual-textual emotion analysis with deep coupled video and danmu neural networks. *IEEE Transactions on Multimedia*, 22(6):1634–1646, 2019.
- [24] 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.
- [25] 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.
- [26] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.
- [27] 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.
- [28] Weixin Li, Xuan Dong, and Yunhong Wang. Human emotion recognition with relational region-level analysis. *IEEE Transactions on Affective Computing*, 2021.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] Depu Meng, Xiaokang Chen, ZeJia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. Nms strikes back. *arXiv preprint arXiv:2212.06137*, 2022.
- [41] 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.

- [42] 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.
- [43] 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. detr: Benchmarking detection transformers, 2023.
- [44] 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.
- [45] Philipp V Rouast, Marc Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 2019.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] Weijie Wang, Nicu Sebe, and Bruno Lepri. Rethinking the learning paradigm for facial expression recognition. *arXiv preprint arXiv:2209.15402*, 2022.
- [51] 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.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [58] 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.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] 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.
- [67] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021.
- [68] 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.
- [69] 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.
- [70] 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.
- [71] 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.
- [72] 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.



Xinpeng Li obtained his bachelor's and master's degrees from South China University of Technology, Guangzhou, China, in 2018 and 2021. He is currently a research assistant at Shenzhen Technology University, Shenzhen, China. His research focuses on computer affective vision which aims to help machines understand emotions in vision media.



Teng Wang is currently a Ph.D. candidate at the Department of Computer Science, the University of Hong Kong. He obtained his bachelor's and master's degrees from Sun Yat-sen University, in 2017 and 2020. His research interests lie in vision-language multimodal learning and video understanding.



Shuyi Mao is currently a Computer Vision Researcher in the AI Lab of Lenovo Research, Shenzhen, China. He received his bachelor's degree in automation from Dongguan University of Technology, Dongguan, China, in 2020, and his master's degree in computer technology from Shenzhen University, Shenzhen, China, in 2023. His research interests include image processing and computer vision.



Jinbao Wang received his Ph.D. degree from the University of Chinese Academy of Sciences (UCAS) in 2019. He is currently an Assistant Professor at the Southern University of Science and Technology (SUSTech), Shenzhen, China. His research interests include image anomaly detection, graph representation learning, and computer vision.



Xiaojiang Peng (Member, IEEE) is currently an Associate Professor at the College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China. He received his Ph.D. degree from Southwest Jiaotong University. He was an Associate Professor at Shenzhen Institutes of Advanced Technology Chinese Academy of Sciences, from 2017 to 2020. His research interests include multi-modal affective computing, facial expression recognition, and deep learning. He was selected by Stanford University as one of the top 2% scientists in 2021.



Feng Zheng (M'19) received his Ph.D. degree from the University of Sheffield, UK, in 2017. He is currently an Associate Professor in the Department of Computer Science and Engineering at Southern University of Science and Technology, Shenzhen, China. His research interests include machine learning, computer vision, and cross-media intelligence.