Two in One Go: Single-stage Emotion Recognition with Decoupled Subject-context Transformer

Xinpeng Li[a], Teng Wang[b], Jian Zhao[c,d], Shuyi Mao[a],
Jinbao Wang[e], Feng Zheng[e], Xiaojiang Peng[a,†], Xuelong Li[c,d,†]
[a]College of Big Data and Internet, Shenzhen Technology University (SZTU), Shenzhen, China
[b]Department of Computer Science, The University of Hong Kong (HKU)
[c]Institute of Artificial Intelligence (TeleAI), China Telecom, China
[d]School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University (NWPU), Xi'an Shanxi, China
[e]School of Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, China

ABSTRACT

Emotion recognition aims to discern the emotional state of subjects within an image, relying on subject-centric and contextual visual cues. Current approaches typically follow a two-stage pipeline: first localize subjects by off-the-shelf detectors, then perform emotion classification through the late fusion of subject and context features. However, the complicated paradigm suffers from disjoint training stages and limited fine-grained interaction between subject-context elements. To address the challenge, we present a single-stage emotion recognition approach, employing a Decoupled Subject-Context Transformer (DSCT), for simultaneous subject localization and emotion classification. Rather than compartmentalizing training stages, we jointly leverage box and emotion signals as supervision to enrich subject-centric feature learning. Furthermore, we introduce DSCT to facilitate interactions between fine-grained subject-context cues in a "decouple-then-fuse" manner. The decoupled query tokens—subject queries and context queries—gradually intertwine across layers within DSCT, during which spatial and semantic relations are exploited and aggregated. We evaluate our single-stage framework on two widely used context-aware emotion recognition datasets, CAER-S and EMOTIC. Our approach surpasses two-stage alternatives with fewer parameter numbers, achieving a 3.39% accuracy improvement and a 6.46% average precision gain on CAER-S and EMOTIC datasets, respectively.

KEYWORDS

emotion recognition, single-stage framework, and decouple-fuse.

ACM Reference Format:
Xinpeng Li[a], Teng Wang[b], Jian Zhao[c,d], Shuyi Mao[a],, Jinbao Wang[e], Feng Zheng[e], Xiaojiang Peng[a,†], Xuelong Li[c,d,†], and [a]College of Big Data and Internet, Shenzhen Technology University (SZTU), Shenzhen, China, [b]Department of Computer Science, The University of Hong Kong (HKU), [c]Institute of Artificial Intelligence (TeleAI), China Telecom, China, [d]School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University (NWPU), Xi'an Shanxi, China, [e]School of Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, China . 2018. Two in One Go: Single-stage Emotion Recognition with Decoupled Subject-context Transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

[†]corresponding author.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Automatic human emotion recognition gets increasing research attention in the multimedia community, where studies include inferring emotions from speech [67, 78], image [45, 74] and multi-modalities [40, 42]. Its potential applications span across healthcare, driver surveillance, and diverse human-computer interaction systems [8, 46, 47, 63], reflecting the fundamental role of emotions [11].

In this paper, we focus on the problem of inferring the emotion of one person in a real-world image. Concretely, given an in-the-wild image, we aim to identify the subject's apparent discrete emotion categories (e.g. happy, sad, fearful, or neutral). Existing methods typically involve two stages: subject detection and emotion classification. Conventional approaches primarily emphasize facial cues [7, 51, 60–62, 75, 82], featuring a *two-stage without fusion* paradigm. As depicted in Fig. 1(a), a standard off-the-shelf detector indicates a facial region, and a dedicated face encoder extracts facial features for subsequent classification into distinct emotional categories. Recent advances have increasingly recognized the importance of contextual cues in emotion recognition, like body language, scene semantics, and social interactions [22, 24, 28, 37, 38, 45, 64, 65]. This system is characterized as a *two-stage with late fusion* paradigm. As illustrated in Fig. 1(b), it first identifies subjects and contexts within the image, processes them through independent encoders, and fuses the resulting features for emotion prediction.

While effective, existing approaches are hindered by two primary limitations. Firstly, the disjointed learning processes of emotion classifiers and subject detectors in a two-stage paradigm often result in inefficient computational efficiency. Illustrated in Fig. 2, existing methods' effectiveness is limited with many parameters. Secondly, existing paradigms may exhibit a restricted capacity for subject-context fusion, thereby falling short in addressing real-world images that are susceptible to nuanced contextual influences [36, 59].

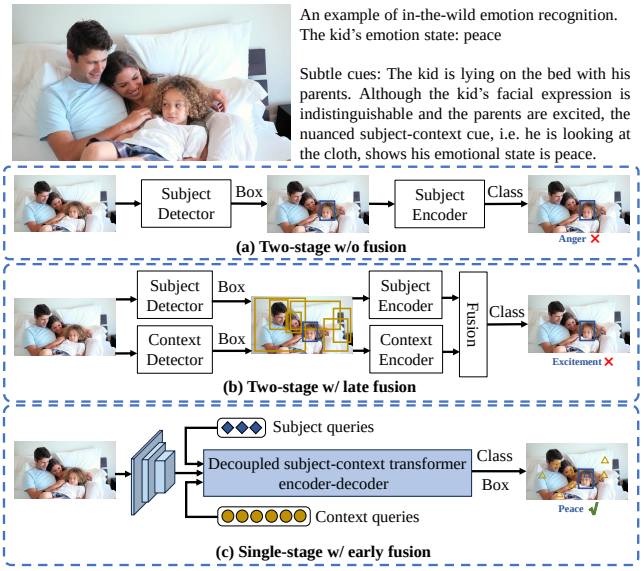

Figure 1: Motivation of single-stage framework. Contexts play a vital and nuanced role in emotion recognition. In (a) and (b), prior methods include two stages: subject without or with context (blue and gold rectangles) region localization and emotion classification without or with late fusion. In (c), we propose a single-stage framework for simultaneous localization and classification and decoupled subject-context transformer with early fusion. Our method notices useful and subtle emotional cues (blue and gold triangles).

Shown in Fig. 1, the first paradigm focuses solely on facial expressions, neglecting essential contextual cues, and the second paradigm's late fusion scheme misses fine-grained subject-context interaction, leading to sub-optimal emotion recognition.

To alleviate the limitation, we introduce a single-stage framework, employing a Decoupled Subject-Context Transformer (DSCT) with early fusion, for simultaneous subject localization and emotion classification, characterized as a *single-stage with early fusion* paradigm. As illustrated in Fig. 1(c), we adopt an encoder-decoder architecture with DSCT, where learnable queries are correlated with the global and multi-scale features for prediction. Rather than disjoint training stages, we jointly leverage box and emotion signals as supervision to enrich subject-centric feature learning, i.e., the framework is trained with a joint loss of classification and localization. Fig. 2 demonstrates that our method is effective and efficient, surpassing two-stage prior arts with fewer parameters.

Furthermore, we introduce DSCT to facilitate interactions between fine-grained subjects and context in a decouple-then-fuse manner. As depicted in Fig. 1(c), the queries are decomposed into subject and context queries to capture the subject's emotional signal, e.g., facial expression, and a wide range of contextual cues, e.g., body posture and gesture, agents, objects, and scene attributes. The decoupled query tokens—subject queries and context queries—gradually intertwine across layers within DSCT. For effective fusion, the spatial and semantic relations between context and subject information are exploited and aggregated. The spatial relation picks up contextual queries with short-range subject-context interaction, such as

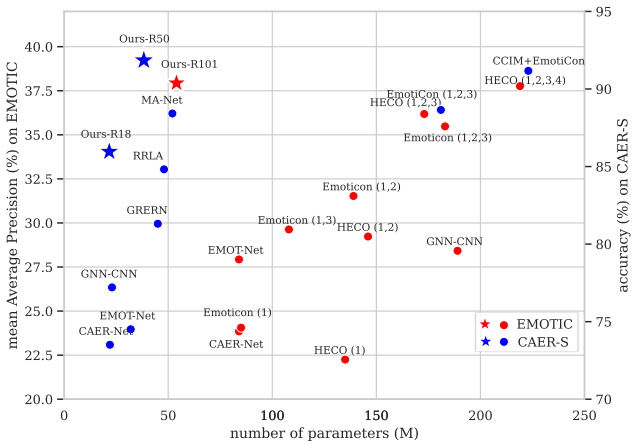

Figure 2: Performance vs. model efficiency of different methods on EMOTIC (red) and CAER-S (blue). Our proposed single-stage framework (star) achieves state-of-the-art performance with fewer parameters than two-stage prior arts (circle).

the subject between objects in hands and close agents. As complementary, the semantic relation chooses contextual queries with long-range subject-context interaction, like the subject between scene attributes and distant people. Fig. 1(c) shows that the single-stage framework notices useful and subtle emotional cues between the subject and context, e.g. the kid is looking at the father's clothes.

Extensive experiments are conducted on two standard context-aware emotion recognition benchmarks to validate the efficacy of our approach. The proposed framework attains impressive results, achieving 91.81% accuracy on the CAER-S dataset [24] and 37.81% mean average precision on EMOTIC [22]. In the case of similar parameter numbers, the proposal surpasses counterparts by a substantial margin of 3.39% accuracy and 6.46% average precision on CAER-S and EMOTIC respectively. Furthermore, we provide valuable insights by visualizing network output, feature map activation, and query selection, underscoring the proposal can discern useful and nuanced emotional cues of subject and context.

The main contributions can be summarized as follows:

- We present a novel single-stage framework for simultaneous subject localization and emotion classification to address the limitations of disjoint training stages.
- To facilitate fine-grained interactions between subjects and context, we introduce a new decoupled subject-context transformer to decouple and fuse queries across layers.
- The spatial and semantic relations are exploited and aggregated to capture the short-range and long-range subject-context interaction complementarily.
- Extensive experiments and visualization on two standard datasets show that the single-stage framework outperforms two-stage alternatives by a significant margin and excels in capturing useful and nuanced emotional cues.

2 RELATED WORK

Visual Emotion Recognition. The Visual Emotion Recognition (VER) task can be broadly categorized into two main paradigms. 1)

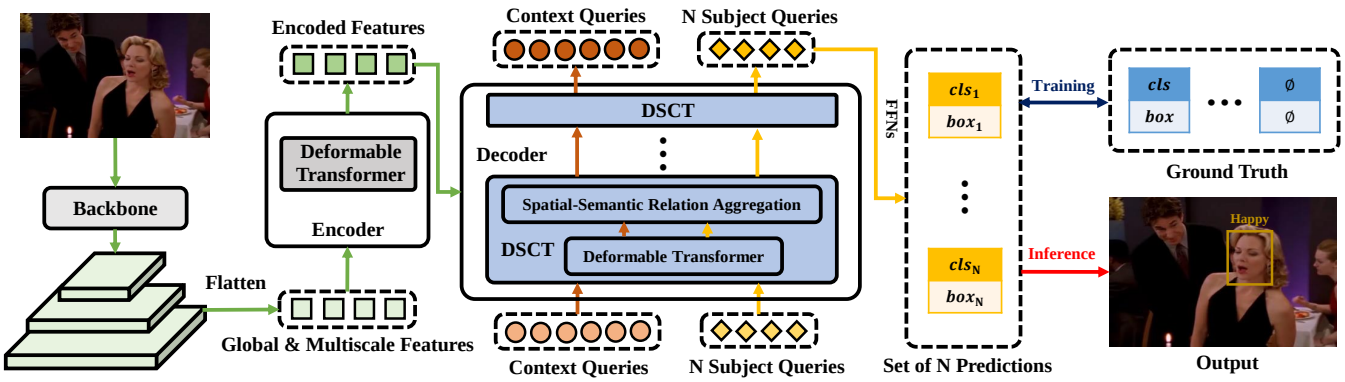

Figure 3: Overall architecture of our single-stage emotion recognition approach for simultaneous subject localization and emotion classification, employing a Decoupled Subject-Context Transformer (DSCT) with early subject-context fusion.

Two-stage without fusion. Traditional methods focus on utilizing subject-centric regions while treating contextual areas as noise, as observed in various studies [9, 27, 43, 50, 75]. The pipeline includes subject detection and emotion classification. These studies primarily address challenges associated with label uncertainty [5, 6, 23, 34, 52, 55, 56, 66, 80, 81], micro expressions [41, 71], and disentangled representations [69, 79]. 2) *Two-stage with late fusion.* In recent years, the research has paid increasing attention to context-aware emotion recognition, which emphasizes the use of multiple contexts for more robust emotion classification [22, 24, 28, 38, 39, 53, 64, 65, 77]. In addition to two-stage components, the pipeline includes multi-branch and late fusion characteristics. Typically, a multiple-stream architecture, followed by a fusion network, is employed to independently encode the subject and context information. *Despite the effectiveness of these methods, they suffer from disjoint training stages and limited interaction between fine-grained subject-context elements. In contrast, we present a single-stage approach with an early fusion, employing a Decoupled Subject-Context Transformer, for simultaneous subject localization and emotion classification.*

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 [10, 54], query [32, 58, 86], and bipartite matching [4, 25, 26, 72, 73]. 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 [70], action detection [33], human-object interaction [19, 20], person search [1], and instance segmentation [17, 57]. *We propose the adaptation and modification for VER: First, since we suggest a single-stage framework, we adopt deformable DETR for simultaneous subject localization and emotion classification; Second, as generic objects exhibit distinct and localized characteristics, but contexts are essential and nuanced-related for VER, we introduce a decoupled subject-context transformer to capture contextual interaction.*

3 METHOD

3.1 Single-stage Framework

Current two-stage approaches for in-the-wild emotion recognition may suffer low efficiency from disjoint training stages and limited interaction between fine-grained subject-context elements. To address the limitation, we introduce a single-stage framework, employing a Decoupled Subject-Context Transformers with early fusion, for subject localization and emotion classification.

Architecture. As shown in Fig.3, the system handles an entire image through a CNN backbone, an encoder with deformable transformers, and a decoder with novel Decoupled Subject-Context Transformers (DSCT). Given an image, we extract multi-scale features through the backbone, flatten them in spatial dimensions, and supplement position encoding and level embeddings. The encoder subsequently encodes the global and multi-scale features through six deformable transformers. After that, the decoder correlates the given learnable queries with encoded features with six DSCTs. Finally, the Feed-Forward Networks (FFNs) transform a set of N subject queries into N final predictions, including emotion classes and bounding boxes. We defer to the supplementary material the detailed definition of the architecture, which follows deformable DETR [86]. We jointly train classification and localization to enrich subject-centric feature learning. Furthermore, DSCTs facilitate fine-grained subject-context interactions by early fusion.

Queries. Each learnable query is a concatenation of 256-dimension spatial and 256-dimension semantic embeddings. The spatial embedding is decoded into the 2-d normalized coordinate of the reference point and the semantic one into 1) the bounding box as relative offsets w.r.t. the reference point and 2) the corresponding emotion class of the subject. The semantic embeddings of queries adaptively integrate multi-scale image features by sampling locations around the reference points. We refer the reader to the supplementary material for detailed definitions, which follow deformable DETR [86]. We adopt a set of N queries for prediction, where N is typically larger than the average subject number per image in a dataset.

Optimization. We use set-level prediction [2] that encapsulates several predictions or ground truth within a set. For clarity, we

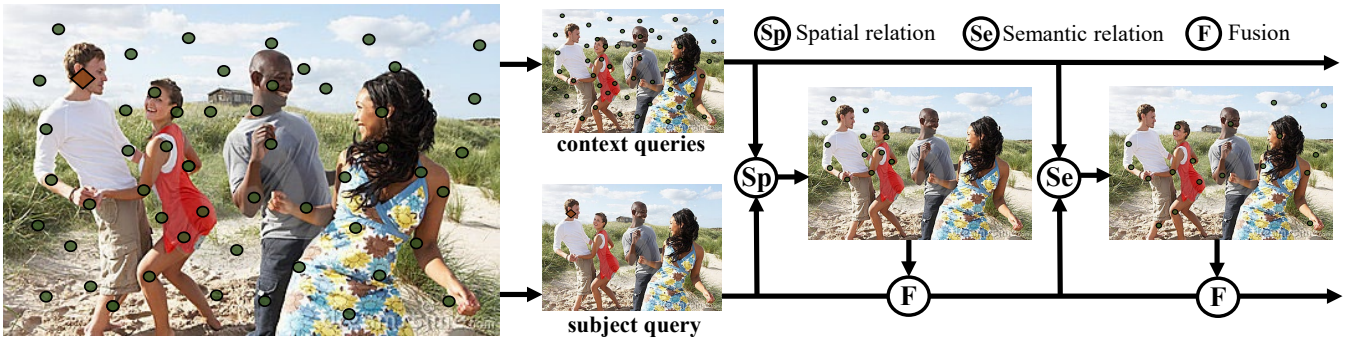

Figure 4: Illustration of the DSCT. The left figure shows the reference points of the subject (orange diamond) and context queries (green circle). The right part describes the spatial-semantic relational aggregation.

denote the i-th element of a set as $(\text{cls}_i, \text{box}_i)$, where cls_i represents the categorical emotion label and $\text{box}_i \in [0,1]^4$ specifies the normalized center coordinate and box's height and width.

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. We employ the bipartite matching [2] that computes one-to-one associations between the set of predictions $\hat{y}$ and the padded ground truths y:

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

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

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

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:

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

where $\lambda_{\text{cls}}, \lambda_{\text{box}} \in \mathbb{R}$ are hyperparameters. For matching and training, we employ the focal loss [30] for $\mathcal{L}_{\text{cls}}$ and set $\mathcal{L}_{\text{box}}$ as the l_1 loss and generalized IoU loss [49].

During inference, we set the mean of the class output logit as the score of each prediction. For multi-label tasks, subject emotions are determined with a threshold t:

$$o = \{ i \mid \hat{y}_i > t \}, \tag{4}$$

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

$$o = \arg\max_{i} \{ \hat{y}_i \}, \tag{5}$$

where o corresponds to the index of the emotion class.

Discussion. Current emotion recognition approaches usually include two steps of localization and classification, which suffer from disjoint training stages. Therefore, we pursue a single-stage framework to simultaneously recognize the subject's bounding box and emotion class. The deformable DETR pipeline, including the above-mentioned one-stage processing and joint classification and localization loss, aligns well with our demand.

3.2 Decoupled Subject-Context Transformer

To facilitate interactions between fine-grained subject and contextual elements, we introduce a novel Decoupled Subject-Context Transformer (DSCT), which treats queries in a "decouple-then-fuse" manner and exploits spatial-semantic relational aggregation.

Decouple then Fuse. Before the DSCT, the queries are decomposed into subject and context queries. As shown in Fig. 3, we directly adopt N subject queries and context queries as input queries of the decoder, where all queries have the same tensor size. In DSCT, both types of queries are correlated with multi-scale image features through the base deformable transformer, and then subject queries integrate context queries by spatial-sentimental relational aggregation before output. The subject and context queries are fused in an early way through all DSCT layers. As illustrated in the left section of Fig. 4, the reference point of the subject query primarily attends to the subject area to capture the subject's emotional signal, e.g., facial expression, while the reference points of the context queries are distributed across the entire image to pick up extensive and subtle contextual cues, e.g., body posture and gesture, surrounding agents, and scene attributes like grass and sky.

Spatial-Semantic Relational Aggregation. As shown in the right part of Fig. 4, the DSCT fuses context queries based on their spatial-semantic relationships w.r.t. the subject query.

The DSCT first picks up queries with the short-range subject-context interaction, such as subject and objects in hands and close agents, based on the relative spatial distance. For each context query, the distance is calculated as the Euclidean distance between reference points of the context and subject queries. For the subject and context queries, we denote their coordinate vectors of the reference points as p_S^n, where $n = 1, ..., N$ and N is the total number of the subject queries, and p_C^m, where $m = 1, ..., M$ and M is the total number of the context queries. The relative spatial distance d_m^n for a pair of subject and context query is computed as:

$$d_m^n = \| p_C^m - p_S^n \|_2. \tag{6}$$

Then we select K_{sp} queries of the shortest spatial distance from total M context queries for each subject query.

As complementary, the queries with long-range subject-context interaction, like subject and scene attributes and distant people, are chosen via semantic relevance. For each context query, the relevance is calculated as the similarity between semantic embeddings of context and subject queries. For the subject and context queries, we denote their semantic embeddings as E_S^n, where $n = 1, ..., N$ and N is the total number of the subject queries, and E_C^m, where $m = 1, ..., M$ and M is the total number of the context queries. The semantic relevance r_m^n for is calculated as:

$$r_m^n = \text{dot}(E_C^m, E_S^n). \tag{7}$$

Since the semantic embeddings are processed with image features in the same architecture, we can measure their similarity without the transforming matrices in previous methods [28, 65]. Then we select K_{sm} queries of the smallest semantic relevance from total M context queries for each subject query.

Finally, we adopt relevance re-weighting fusion to integrate the context queries into the subject query. We denote their semantic embeddings as E_S and E_C^k, where $k = 1, ..., K_{sm} + K_{sp}$. The attention weight w^k is computed by the dot product of E_S and E_C^k. Then the softmax function makes the sum of attention weights to be 1. After that, the fused contextual subject query $\hat{E}_S$ is defined as:

$$\hat{E}_S = \sum_{k=1}^{K_{sm}+K_{sp}} w^k E_C^k + E_S. \tag{8}$$

Discussion. In the object detection task, queries can capture effective information from distinctive and localized areas. While in the task of emotion recognition, contexts have essential and nuanced influence, we introduce the novel DSCT to capture fine-grained subject-context interaction, where spatial and semantic relationships are exploited and aggregated.

4 EXPERIMENTS

4.1 Implementations

We set the number of queries N to 4 and 9 for CAER-S [24] and EMOTIC [22]. We set N to 4 and 9. To facilitate the training, we initialize the weights of architecture and borrow 300 context queries from Deformable DETR [86], 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. For evaluation, we first use non-max-suppression to remove the duplicate subjects and then select the subject that exhibits the highest bounding box overlap with the ground truth. The experiments were conducted using 8 GPUs of the NVIDIA Tesla A6000. We show details about architectural configurations, training strategies, and preprocessing steps, which follow those outlined in [86], in the supplementary material.

4.2 Datasets

We conducted extensive experiments on two typical and popular context-aware emotion recognition datasets in real-world scenarios, namely the CAER-S [24] and EMOTIC [22].

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) [24].

The EMOTIC dataset [22] 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 [22].

4.3 Quantitative and Qualitative Results

The performance of various methods on CAER-S and EMOTIC datasets is presented in Table 1 and Table 2. 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. For the methods without released code, we count their parameters through their backbone configuration in paper and present the details in the rightmost column. Subscripts in EmotiCon [38] and HECO [65] 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 proposal 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, the proposal even surpasses larger-parameter approaches, underscoring its effectiveness and efficiency for emotion recognition when compared to two-stage methods.

We present the qualitative results in Fig. 5 and Fig. 6, depicting the bounding boxes and emotion classes output by our proposal framework, alongside those produced by the EMO-Net [22], a representative two-stage late-fusion method. To enhance the clarity of the proposal's output, we include visual indicators for subject queries' reference points and sampling locations. Outputs of different subjects are color-coded for differentiation. These visualizations illustrate that the proposal 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".

4.4 Visualization and Analysis

Classification and localization. We conducted experiments to fine-tune θ_{cls} and λ_{cls} on the EMOTIC dataset [22] and keep $\theta_{\text{box}} = 5$ and $\lambda_{\text{box}} = 5$. The results of these hyperparameter experiments are thoughtfully presented in Table 3. Notably, the most compelling performance is achieved when $\theta_{\text{cls}} : \theta_{\text{box}}$ is set to 2:5, and $\lambda_{\text{cls}} : \lambda_{\text{box}}$ is set to 5:5. When the ratio of classification and localization coefficients raises to 3:1, the performance drops significantly by 1.25%.

Methods	Acc (%)	Param.(M)	Backbone
With <100M parameters			
Ours-R18	84.96	22	ResNet18
CAER-Net-S [24]	73.51	22	12-layer CNN
GNN-CNN [76]	77.21	23	VGG16
EfficientFace [84]	85.87	25	MobileNet28, ResNet18
EMOT-Net [22]	74.51	32	ResNet18 × 2
SIB-Net [29]	74.56	33	ResNet18 × 3
Ours-R50	**91.81**	39	ResNet50
GRERN [13]	81.31	45	ResNet101
RRLA [28]	84.82	48	ResNet50, RCNN50
MA-Net [83]	88.42	52	Multi-Scale ResNet18
With >100M parameters			
EmotiCon [38]	88.65	181	OpenPose, RobustTP, Megadepth
VRD [16]	90.49	380	{VGG19, ResNet50, FRCNN50} × 2
CCIM+EmotiCon [64]	91.17	223	OpenPose, RobustTP, Megadepth, ResNet101

Table 1: Performance and model efficiency on the CAER-S.

Methods	mAP (%)	Param.(M)	Backbone
With <100M parameters			
Ours-R50	37.26	39	ResNet50
Ours-R101	**37.81**	58	ResNet101
EMOT-Net [22]	27.93	84	YOLO, ResNet18
CAER-Net [22]	20.84	84	YOLO, CNN-12
EmotiCon$_{(1)}$ [38]	31.35	85	OpenPose, 15-layer CNN
With >100M parameters			
EmotiCon$_{(1,3)}$ [38]	35.28	108	OpenPose, 15-layer CNN, Medadepth
HECO$_{(1)}$ [65]	22.25	135	YOLO, Alphapose, ResNet18
EmotiCon$_{(1,2)}$ [38]	32.03	139	OpenPose, RobustTP, ResNet18
HECO$_{(1,2)}$ [65]	36.18	146	YOLO, Alphapose, ResNet18 × 2
HECO$_{(1,2,3)}$ [65]	34.93	173	YOLO, Alphapose, ResNet50, ResNet18 × 2
EmotiCon$_{(1,2,3)}$ [38]	32.03	183	OpenPose, RobustTP, ResNet18, Megadepth
GCN-CNN [76]	28.16	189	YOLO, VGG16, 6-layer GCN
HECO$_{(1,2,3,4)}$ [65]	37.76	219	YOLO, Alphapose, {ResNet18, ResNet50} ×2, Faster RCNN50
EmotionCLIP [77]	32.91	577	YOLO, ViT_b_32

Table 2: Performance and model efficiency on the EMOTIC.

We can see adding appropriate localization loss can boost classification performance and facilitate subject-centric feature learning. Besides, we noticed that θ_{cls} has minimal impact on performance, whereas λ_{cls} significantly influences the results.

Feature map activation. We visualize feature map activation of methods of different paradigms. We select EMO-Net [22] as a representative two-stage method. For fairness, we re-implement

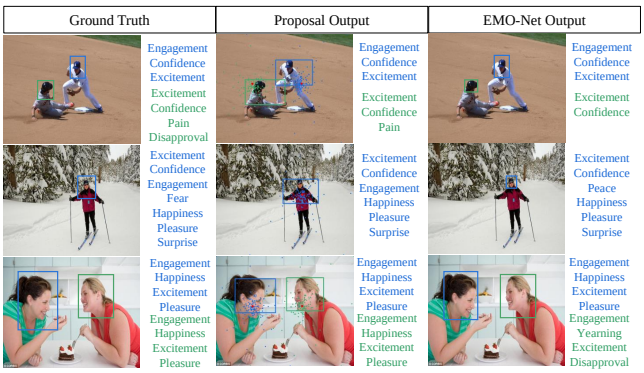

Figure 5: The output visualization on EMOTIC.

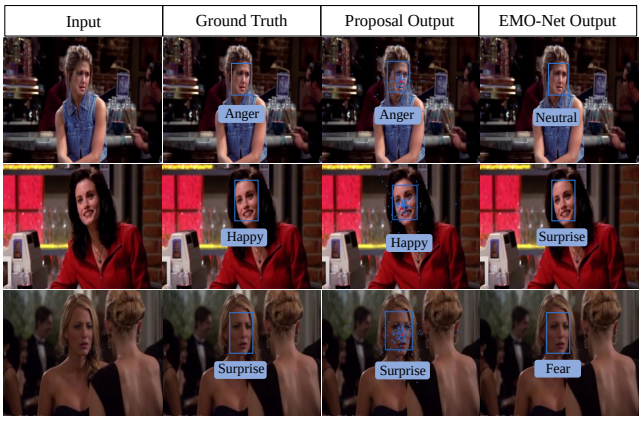

Figure 6: The output visualization on CAER-S.

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 3: Performance of different classification coefficients.

EMO-Net using ResNet50 as the backbone to achieve a similar model complexity to the proposal (referred to as EMO-Net-R50). The selected feature maps originate from the final layer of the ResNet50 backbone. In Fig. 7, a clear distinction emerges: EMO-Net-R50 exhibits a tendency to emphasize a few large regions, while the proposal consistently places importance on smaller, more intricate areas. This observation suggests that the proposal excels in the precise handling of fine-grained subject-context cues compared to conventional two-stage methods of coarse-grained cues.

Sampling positions of queries We visualize the reference point (blue star) and feature sampling positions (grey circle) of the normal subject queries and contextual subject queries of the DSCT in Fig. 8. For the normal subject query, the sampling positions only rely on its own sampling points [86]. For the contextual subject query, the sampling positions rely on both the sampling points of

Figure 7: The visualization of feature maps activation.

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
Ours-R50	36.91	35.20	31.20	40.96	35.97

Table 4: Performance on images with multiple subjects.

the subject query and context queries. We can see the sampling positions of normal subject queries only cover the subject area while the contextual ones are densely distributed across the image. The quantitative result shows the contextual query of DSCT outperforms the normal one with a 0.71% precision improvement. The gain can be attributed to aggregating extensive contextual cues, which are essential for in-the-wild emotion recognition.

Figure 8: The visualization of the reference point (blue star) and sampling positions (grey circle) of queries.

Multiple subjects. We conduct an evaluation on images with varying subject numbers. We select EMO-Net-R50 and EMO-Net-R50-M, which further masks subjects in the context [24], for comparison. Table 4 presents the performance on EMOTIC for images with different subject numbers. As the subject number in an image increases, the complexity of subject-context interaction also rises. Notably, the proposal maintains stable performance with increasing subject numbers, while EMO-Net-R50's performance deteriorates. This observation verifies our early fusion proposal can handle complex interactions better than late fusion two-stage methods.

4.5 Ablation Study

Subject Query Number. We conduct experiments to investigate the impact of query number setting. Table 5 presents the performance of different query numbers. Fig. 9 offers a visualization of

Figure 9: The position visualization of subject queries.

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 5: Ablation study on subject query number.

Baseline	Decouple-fuse	Spatial	Semantic	mAP (%)
✓	×	×	×	36.55
✓	✓	×	×	36.73
✓	✓	✓	×	37.11
✓	✓	×	✓	36.85
✓	✓	✓	✓	**37.26**

Table 6: Ablation study on DSCT components.

bounding boxes (colorful rectangles), reference points (red circles), and sampling locations (colorful circles) corresponding to different subject query numbers. Table 5 shows the proposal 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. Fig. 9 also indicates that different subject queries attend to separate subject areas.

Components of DSCT. We assess the impact of components of DSCT on the EMOTIC dataset [22], and the results are displayed in Table 6. The categories include "baseline" (only subject queries), "Decouple-fuse" (decouple queries and then fuse), "Spatial" (select K_{sp} context queries with shortest spatial distance), and "Semantic" (select K_{se} context queries with semantic relevance). As we can see, the DSCT enhances performance by 0.71%, highlighting the importance of sufficient contextual interaction and fusion. Specifically, decoupling and fusing context queries improve performance by 0.18%, and selecting context queries based on spatial and semantic relation boosts the result by an extra 0.56% and 0.12%. The results demonstrate the effectiveness of each proposed component.

Selection of Spatial and Semantic Relation. We conduct experiments with varying values of K_{sp} and K_{se} on EMOTIC [22] to evaluate the sensitivity of Spatial and Semantic Relation parameters. As shown in Figure 11, the optimal performance is achieved when K_{sp} is set to 100. The best performance is 0.41% higher than the one when K_{sp} is 300. This observation suggests that not all contextual information is valuable for effective emotion recognition. The selection of the 100 closest context queries w.r.t. the subject

Figure 10: The position visualization of the subject query (blue star) and selected context queries (grey spots).

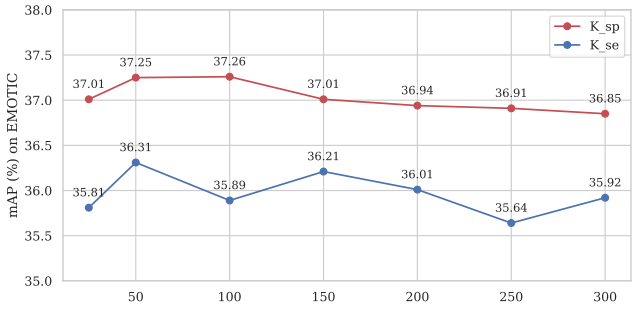

Figure 11: The evaluation of K_{sp} and K_{se} on EMOTIC [22]

query also aligns with gradual interaction decay in [65]. Figure 10 shows spatial relational selection keeps contextual cues such as objects in hands and close agents for short-range interactions. Besides, the optimal performance is achieved when K_{se} is set to 50, which is 0.39% higher than the one when K_{se} is 300. This suggests that some contextual information is a disturbance. Figure 10 shows semantic relational selection preserves contextual signals like scene attributes and distant people for long-range interactions. We can also see two selections are complementary from Figure 10.

Re-weighting Strategy. We conduct experiments of re-weighting strategies for context query fusion on the EMOTIC dataset. The results, shown in Table 7, indicate the performance under different re-weighting strategies: Semantic (semantic relevance weights based on semantic relation w.r.t the subject query), Equal (equal weights), Spatial (spatial distance weights based on spatial relation w.r.t the subject query), and Attentive (attentive weights learned from the subject query through a linear layer). The findings reveal that the best performance is achieved when employing semantic re-weighting, which performs better than equal re-weighting by 0.53%. The observation aligns well with previous studies that contexts have variant contributions to emotion recognition [28, 65].

Re-weighting strategy	Semantic	Equal	Spatial	Attentive
mAP %	**37.26**	36.73	36.48	36.93

Table 7: The ablation study on re-weighting strategy.

Fusion location. We conduct experiments of different locations for subject-context fusion on the EMOTIC dataset and show the results in Table 8. The decoder has six layers, and we perform the query fusion in different layers. The best performance is achieved when fusing in the 1st to 6th layers, which exceeds fusing only in the 6th layer by 0.57%. It can be explained by that early fusion can capture fine-grained subject-context interaction to boost performance. The observation aligns with the psychological studies [36, 59].

locations	1st to 6th	2nd to 6th	4th to 6th	6th
mAP %	**37.01**	36.58	36.82	36.44

Table 8: The ablation study on fusion location.

Feature Extractor. To evaluate the influence of feature extractors, we conducted experiments with various backbone architectures for the proposal on both EMOTIC and CAER-S datasets. The results, as depicted in Table 9, include performance metrics and corresponding parameter counts. Specifically, "R" refers to ResNet [15], and "WR" designates Wide ResNet [68]. 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 significantly influenced by the choice of pre-trained initialization.

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 9: Ablation study on feature extractor.

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 proposal, we conducted experiments on the CAER-S dataset by leveraging the detrex platform [48]. For equitable comparisons, all architectures utilize ResNet50 as the backbone. The results, summarized in Table 10, reveal the performance of proposal 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 [73]	89.27
Anchor-DETR [58]	90.42	Deformable DETR [86]	91.81
DAB-DETR [10]	87.49	DAB-D-DETR [10]	91.39
DN-DETR [25]	89.79	H-D-DETR [18]	91.65
Conditional-DETR [35]	90.07	DETA [44]	91.77

Table 10: Ablation study on different DETR architectures.

5 CONCLUSION

This paper introduces a single-stage visual emotion recognition framework with Decoupled Subject-Context Transformers (DSCT). The proposal predicts subjects' emotions and locations simultaneously and processes subject and context emotional cues by early fusion. We evaluate our single-stage framework on two widely used context-aware emotion recognition datasets, CAER-S and EMOTIC. Our approach surpasses two-stage alternatives with fewer parameter numbers, achieving a 3.39% accuracy improvement and a 6.46% average precision gain on CAER-S and EMOTIC datasets, respectively. We observe that the joint training of localization and classification can facilitate subject-centric feature learning. Besides, we find that early fusion improves handling the fine-grained subject-context interaction, e.g. multiple subjects in one scene. We also explore the spatial and semantic relationships between subject and contextual cues for more effective interaction and fusion.

REFERENCES

[1] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. 2022. PSTR: End-to-End One-Step Person Search With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9458–9467.

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

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[4] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. 2022. 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. 2020. 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*. 13984–13993.

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

[7] Yuedong Chen, Xu Yang, Tat-Jen Cham, and Jianfei Cai. 2022. Towards unbiased visual emotion recognition via causal intervention. In *Proceedings of the 30th ACM International Conference on Multimedia*. 60–69.

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

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

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

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

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

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

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

[17] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. 2021. Istr: End-to-end instance segmentation with transformers. *arXiv preprint arXiv:2105.00637* (2021).

[18] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. 2023. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19702–19712.

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

[20] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. 2022. 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*. 19578–19587.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

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

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

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

[26] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. 2023. 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*. 3041–3050.

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

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

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

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

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

[32] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. 2022. 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. 2021. 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. 2022. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In *European Conference on Computer Vision*. Springer, 648–665.

[35] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3651–3660.

[36] Batja Mesquita and Michael Boiger. 2014. Emotions in context: A sociodynamic model of emotions. *Emotion Review* 6, 4 (2014), 298–302.

[37] Trisha Mittal, Aniket Bera, and Dinesh Manocha. 2021. Multimodal and context-aware emotion perception model with multiplicative fusion. *IEEE MultiMedia* 28, 2 (2021), 67–75.

[38] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. In *CVPR*. 14234–14243.

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

[40] Dung Nguyen, Duc Thanh Nguyen, Rui Zeng, Thanh Thi Nguyen, Son N Tran, Thin Nguyen, Sridha Sridharan, and Clinton Fookes. 2021. Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. *IEEE Transactions on Multimedia* 24 (2021), 1313–1324.

[41] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. 2023. Micron-BERT: BERT-based Facial Micro-Expression Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1482–1492.

[42] Weizhi Nie, Minjie Ren, Jie Nie, and Sicheng Zhao. 2020. C-GCN: Correlation based graph convolutional network for audio-video emotion recognition. *IEEE Transactions on Multimedia* 23 (2020), 3793–3804.

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

[44] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. 2022. NMS Strikes Back. *arXiv preprint arXiv:2212.06137* (2022).

[45] Mijanur Palash and Bharat Bhargava. 2023. EMERSK-Explainable Multimodal Emotion Recognition with Situational Knowledge. *IEEE Transactions on Multimedia* (2023).

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

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

[48] 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. 2023. detrex: Benchmarking Detection Transformers. arXiv:2306.07265 [cs.CV]

[49] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. 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*. 658–666.

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

[51] Delian Ruan, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. 2020. Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2833–2841.

[52] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. 2021. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6248–6257.

[53] Dhruv Srivastava, Aditya Kumar Singh, and Makarand Tapaswi. 2023. How You Feelin'? Learning Emotions and Mental States in Movie Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2517–2528.

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

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

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

[57] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8741–8750.

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

[59] Matthias J Wieser and Tobias Brosch. 2012. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in psychology* 3 (2012), 35406.

[60] Yi Wu, Shangfei Wang, and Yanan Chang. 2023. Patch-Aware Representation Learning for Facial Expression Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6143–6151.

[61] Zhenqian Wu, Yazhou Ren, Xiaorong Pu, Zhifeng Hao, and Lifang He. 2023. Generative Neutral Features-Disentangled Learning for Facial Expression Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4300–4308.

[62] Zhen Xing, Weimin Tan, Ruian He, Yangle Lin, and Bo Yan. 2022. Co-completion for occluded facial expression recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 130–140.

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

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

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

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

[67] Jiaxin Ye, Yujie Wei, Xin-Cheng Wen, Chenglong Ma, Zhizhong Huang, Kunhong Liu, and Hongming Shan. 2023. Emo-DNA: Emotion Decoupling and Alignment Learning for Cross-Corpus Speech Emotion Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5956–5965.

[68] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).

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

[70] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. 2022. 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*. Springer, 659–675.

[71] Zhijun Zhai, Jianhui Zhao, Chengjiang Long, Wenju Xu, Shuangjiang He, and Huijuan Zhao. 2023. 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*. 22086–22095.

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

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

[74] Haimin Zhang and Min Xu. 2021. Recognition of emotions in user-generated videos through frame-level adaptation and emotion intensity learning. *IEEE Transactions on Multimedia* 25 (2021), 881–891.

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

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

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

[78] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. 2017. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia* 20, 6 (2017), 1576–1590.

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

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

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

[82] Zengqun Zhao and Qingshan Liu. 2021. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1553–1561.

[83] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the

wild. *IEEE Transactions on Image Processing* 30 (2021), 6544–6556.
[84] Zengqun Zhao, Qingshan Liu, and Feng Zhou. 2021. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 3510–3519.
[85] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.
[86] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).

A PRELIMINARIES

Sampling Locations. The core of deformable attention is to reduce computation cost by attending to a small set of key sampling points of spatial locations around a reference point. Given a multi-scale input feature map $\{x^l\}_{l=1}^L$ where $x^l \in \mathbb{R}^{C\times H_l\times W_l}$, the K sampling locations for each attention head and each feature level are generated from the semantic embedding of each *query* element $z_q \in \mathbb{R}^C$. Because the direct prediction of coordinates of sampling location is difficult to learn, it is formulated as a prediction of a reference point $r_q \in [0,1]^2$ along with K sampling offsets $\Delta r_q \in \mathbb{R}^{M\times L\times K\times 2}$. So, the k^{th} sampling location at l^{th} feature level and m^{th} attention head for query element z_q is defined by $p_{mlqk} = \phi_l(r_q) + \Delta r_{mlqk}$ where $\phi_l(\cdot)$ is a function for rescaling the coordinate of reference point to the input feature map of the l^{th} level.

Deformable Attention Module. Given a multi-scale input feature map $\{x^l\}_{l=1}^L$, the multi-scale deformable attention $f_q^{ms} = \text{MSDeformAttn}(z_q, p_q, \{x^l\}_{l=1}^L)$ for query element z_q is calculated using a set of predicted sampling locations p_q as follows:

$$f_q^{ms} = \sum_{m=1}^{M} W_m \Big[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk}\cdot W'_m \Phi_{mlqk} \Big], \tag{9}$$

where l, k and m index the input feature level, the sampling location and the attention head, respectively, while A_{mlqk} indicates an attention weight for the k^{th} sampling location at the l^{th} feature level and the m^{th} attention head. Φ_{mlqk} means the sampled k^{th} key element at l^{th} feature level and m^{th} attention head using the sampling location, which is obtained by bilinear interpolation as $\Phi_{mlqk} = x^l(p_{mlqk}) = x^l(\phi_l(r_q) + \Delta r_{mlqk})$. W_m and W'_m serve as learnable embedding parameters for the m^{th} attention head, and $Amlqk$ is normalized such that $\sum_{k,l} A_{mlqk} = 1$.

B DETAILED ARCHITECTURE

Encoder. We employ the multi-scale deformable attention module in place of the standard encoder layer. In accordance with [86], the encoder both takes in and produces multi-scale feature maps with matching resolutions. Within the encoder, we derive multi-scale feature maps $\{x^l\}_{l=1}^{L-1}$ ($L = 4$) from the output feature maps of stages C_3 to C_5 in ResNet [15] (modified by a 1×1 convolution). Each C_l has a resolution 2^l lower than the original image. The lowest resolution feature map x^L is acquired through a 3×3 convolution with a stride of 2 on the final C_5 stage, labeled as C_6. All multi-scale feature maps consist of $C = 256$ channels. To determine the feature level of each query pixel, we introduce a scale-level embedding, referred to as e_l, to the feature representation, in addition to the positional embedding. Unlike the positional embedding with predetermined encodings, the scale-level embeddings $\{e_l\}_{l=1}^L$ are initialized randomly and trained alongside the network.

Decoder. In our approach, we employ the Decoupled Subject-Context Transformer (DSCT) across all decoder layers. Our methodology encompasses three key components: Deformable Attention, Self-Attention Modules, and Spatial-Semantic Relational Aggregation. Deformable attention facilitates the extraction of features from feature maps, self-attention modules enable queries to interact with each other, while spatial-semantic relational aggregation exploits spatial-semantic relationships for the fusion of subject and context.

C MORE IMPLEMENTATION DETAILS

ImageNet [12] pre-trained ResNet-50 [15] serves as the backbone for our ablation experiments. By default, deformable attentions utilize $M = 8$ and $K = 4$. Parameters of the deformable Transformer encoder are shared across different feature levels. Training models last for 50 epochs by default, with a learning rate decay at the 40th epoch by a factor of 0.1. Similar to DETR[2], our models are trained using the Adam optimizer [21], with a base learning rate of 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 10^{-4}. The learning rates of the linear projections, responsible for predicting query reference points and sampling offsets, undergo a 0.1 multiplication.

We incorporate scale augmentation, adjusting the size of input images so that the shortest side ranges from 480 to 800 pixels, while the longest side is at most 1333 pixels. To facilitate the learning of global relationships through encoder self-attention, we also introduce random crop augmentations during training. Specifically, there's a 0.5 probability of cropping a training image to a random rectangular patch, which is then resized to 800-1333 pixels.

D ADDITIONAL RESULTS

Classification vs. Localization. We conducted experiments to fine-tune λ_{box} on the EMOTIC dataset [22] while maintaining θ_{cls}, θ_{box}, and λ_{cls} constant. The results are showcased in Table 11, and the loss curves are depicted in Figure 12. The optimal outcome is attained when $\lambda_{box} = 5$, surpassing the performance achieved with $\lambda_{box} = 1$ by 1.06% in average precision. This underscores the affirmative influence of integrating a localization loss on subject-centric feature acquisition. As illustrated in Figure 12, the supplementary localization task enhances performance by mitigating the risk of classification over-fitting during model training.

λ_{box}	1	5	10	15
mAP (%)	35.61	**36.67**	36.45	36.61

Table 11: Performance of different localization coefficients.

Comparison of Early Fusion and Late Fusion. We investigate the efficacy of early fusion and late fusion by conducting an evaluation on images featuring varying numbers of subjects. We employ DSCT for all layers and for the 6th layer, representing early fusion and late fusion, respectively. It is worth noting that in late fusion, the queries still incorporate multi-scale image features from the encoder. However, the impact of fusing more low-level features can be inferred by employing DSCTs for early layers of the decoder. Table 12 showcases the performance on EMOTIC (mAP

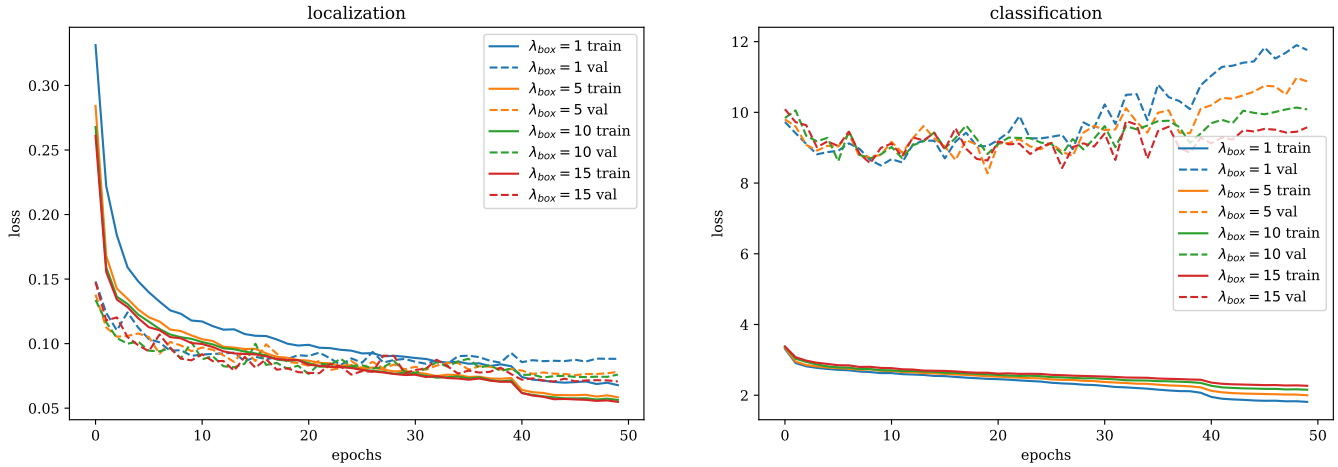

Figure 12: The loss of classification and localization during training.

Figure 13: The output comparison of early and late fusion (incorrectly inferred emotions are marked in red).

%) for images with different subject counts. With an increase in the number of subjects in an image, the subtlety and complexity of subject-context interaction also escalate. Early fusion demonstrates superior performance when the number of faces exceeds four, affirming that the proposed early fusion mechanism adeptly handles subtle subject-context interactions. Furthermore, we visually depict output examples of early and late fusion in Figure 12. Early fusion excels in discerning nuanced emotional states such as sympathy, confusion, disapproval, sensitivity, and embarrassment, which are inferred through fine-grained interactions among agents.

Multiple Modalities. In some studies, the incorporation of multiple modalities has been proposed to enhance context-based emotion recognition [38, 65]. To investigate the potential benefits of including additional modalities in the proposal, we conducted experiments on the EMOTIC dataset. Specifically, we introduced three

Subject #	1	2	3	4	>=5
Image #	2444	938	234	37	29
Late fusion	**36.94**	35.02	**31.21**	40.52	35.36
Early fusion	36.91	**35.20**	31.20	**40.96**	**35.97**

Table 12: Performance on images with multiple subjects.

modalities: "Scene", "Semantic", and "Instance" corresponding to scene classification, semantic segmentation, and instance segmentation, respectively. The networks employed for these modalities are Places365 [85], Deeplabv3 [3], and MaskRCNN [14], all adopting a ResNet50 backbone. We extracted multi-scale features from these networks and integrated them with the proposal's features while keeping the parameters of the other modality networks frozen. The results, as summarized in Table 13, 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 the proposal, which is adept at capturing fine-grained cues.

Modality	None	Scene	Semantic	Instance
mAP %	**37.26**	32.08	34.01	34.11

Table 13: Ablation study on adding different modalities.