

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

Anonymous Authors

1 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:

$$f_q^{ms} = \sum_{m=1}^M W_m \left[\sum_{l=1}^L \sum_{k=1}^K A_{mlqk} \cdot W'_m \Phi_{mlqk} \right], \quad (1)$$

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 A_{mlqk} is normalized such that $\sum_{k,l} A_{mlqk} = 1$.

2 DETAILED ARCHITECTURE

Encoder. We employ the multi-scale deformable attention module in place of the standard encoder layer. In accordance with [11], 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 [5] (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.

3 MORE IMPLEMENTATION DETAILS

ImageNet [3] pre-trained ResNet-50 [5] 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[1], our models are trained using the Adam optimizer [6], 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.

4 ADDITIONAL RESULTS

Classification vs. Localization. We conducted experiments to fine-tune λ_{box} on the EMOTIC dataset [7] while maintaining θ_{cls} , θ_{box} , and λ_{cls} constant. The results are showcased in Table 1, and the loss curves are depicted in Figure 1. The optimal outcome is attained when $\lambda_{\text{box}} = 5$, surpassing the performance achieved with $\lambda_{\text{box}} = 1$ 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 1, 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 1: 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

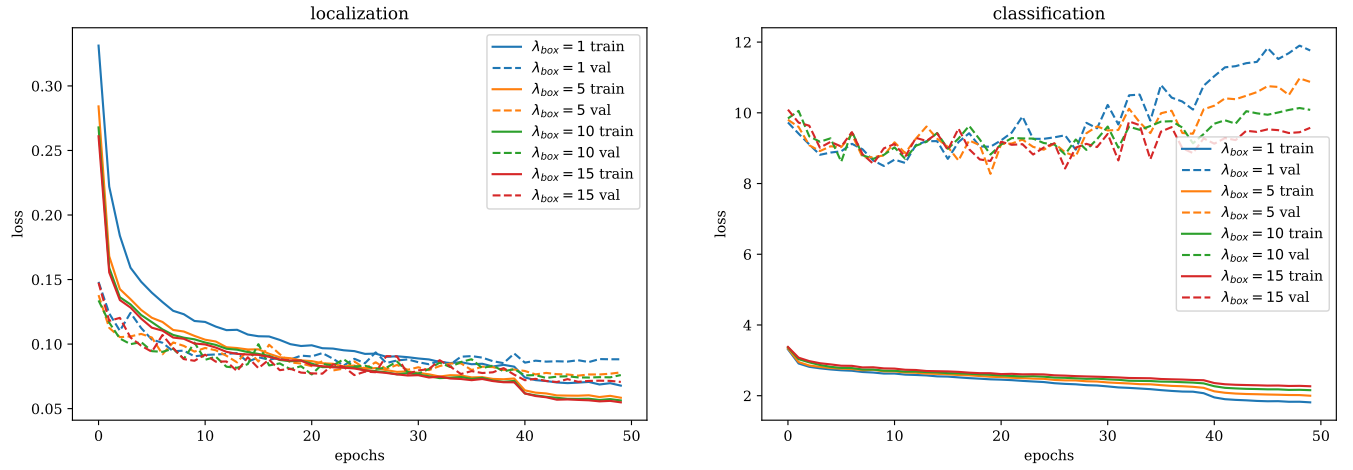


Figure 1: The loss of classification and localization during training.

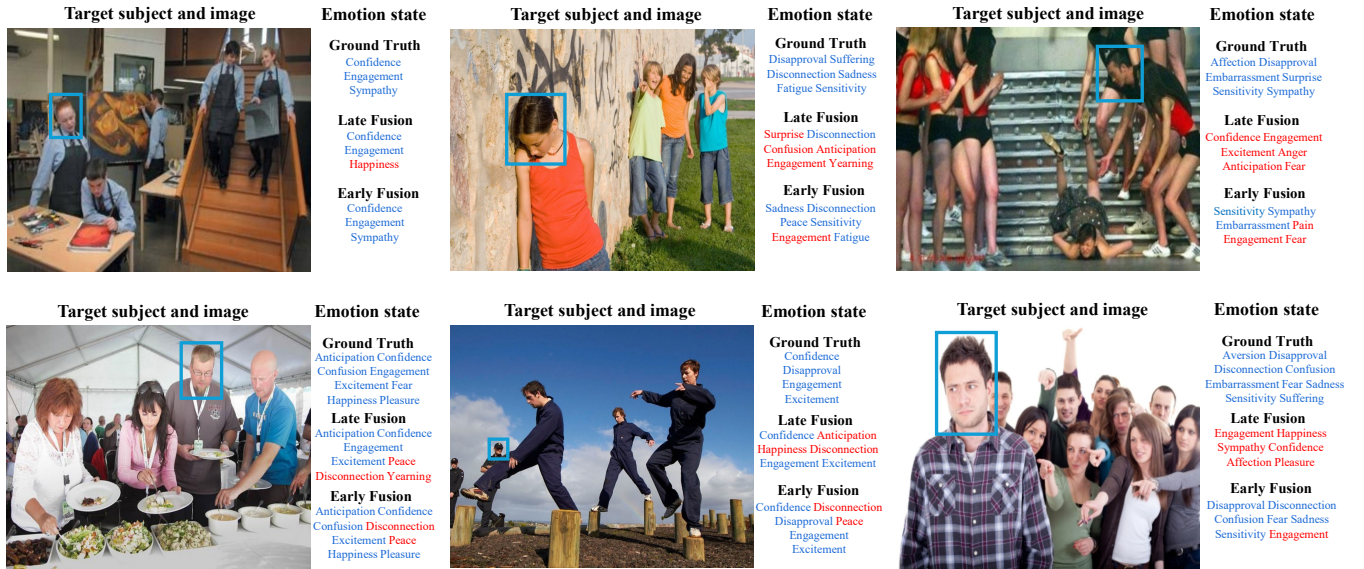


Figure 2: The output comparison of early and late fusion (incorrectly inferred emotions are marked in red).

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 2 showcases the performance on EMOTIC (mAP %) 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 2. 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.

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 2: Performance on images with multiple subjects.

Multiple Modalities. In some studies, the incorporation of multiple modalities has been proposed to enhance context-based emotion recognition [8, 9]. To investigate the potential benefits of including additional modalities in the proposal, 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 [10], Deeplabv3 [2], and MaskRCNN [4], 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 3, 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 3: Ablation study on adding different modalities.

REFERENCES

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [7] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *TPAMI* 42, 11 (2019), 2755–2766.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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).