Self-supervised Social Relation Representation for Human Group Detection

Supplementary Material

Jiacheng Li[1*], Ruize Han[1*†], Haomin Yan[1], Zekun Qian[1], Wei Feng[1], and Song Wang[2]

[1] Intelligence and Computing College, Tianjin University, China
[2] University of South Carolina, Columbia, USA
{threeswords, han_ruize, yan_hm, clarkqian, wfeng}@tju.edu.cn, songwang@cec.sc.edu

1 Qualitative results

We provide several qualitative results as below (the subjects bounded by the same-color boxes are in the same group), from which we can see that the proposed method can well handle the complex scenes with background clutter. From the failure case we find that the identification of some groups need more long-term information.

(a) successful case (b) failure case (ours) (c) failure case (GT)

2 Computation cost

We discuss the computation cost of our method. We show the model size (number of parameters) and Flops of our method in Table 1. To be more clear, we report the cost of our model in the shared embedding network, and the heads of two stages (as shown in Fig. 2 in the paper), separately. We can first see that the shared embedding network dominates the model parameters, which is trained at Stage 1 in a self-supervised manner without the annotations. This alleviates the difficulty of the main model parameter training required abundant labeled data. Note that, the Flops on two datasets are quite different since the input sizes (number of subjects) are different. For the whole model, compared to the mainstream networks, e.g., ResNet50 (25M parameters, 3.9G FLOPs), the proposed model is not very gigantic.

Table 1. Computation cost of our method.

Component	# Para.	Flops (PANDA)	Flops (JRDB-Group)
Shared (ϕ)	1.179 M	6.598 G	1.415 G
S1 Head	0.232 M	0.169 G	0.034 G
S2 Head	0.528 M	0.470 G	0.158 G

3 Parameter study

We provide the comparison results to show the sensitivity of our model to the hyper-parameters, i.e., the percentage of swapped persons. As shown in table below, we increase/decrease the percentage from 10% on PANDA to 5% and 20%, and 20% on JRDB-Group to 10% and 30%, respectively, and report the results. We can see that, although with a relatively large range of parameter tuning, the proposed method is not very sensitive. We also surprisingly find that, on PANDA dataset, the selection of 5% of subjects to swap produces a slightly better result. Note that, this parameter study is conducted during the rebuttal stage in ECCV 2022. We include the new results in this supplementary material, but maintain the original settings and results in the paper.

Table 2. Sensitivity of our model to the hyper-parameters.

Percentage	$\mathcal{F}$ (PANDA)	Percentage	$\mathcal{F}$ (JRDB-Group)
5%	53.8	10%	52.8
10% (Ours)	53.2	20% (Ours)	56.9
20%	48.3	30%	53.7

4 Visualization

As shown in the figure below, we illustrate the the stacked attentions map $\mathbf{U}$, the predicted and ground-truth group relation matrix $\hat{\mathbf{R}}$ and $\mathbf{R}$, as described in Fig. 4 in the paper. We can see from the attention maps that response scores for two subject in or not in the same social group are discriminative. The predicted $\hat{\mathbf{R}}$ generated from $\mathbf{U}$ can effectively estimate the social relations among the subjects.