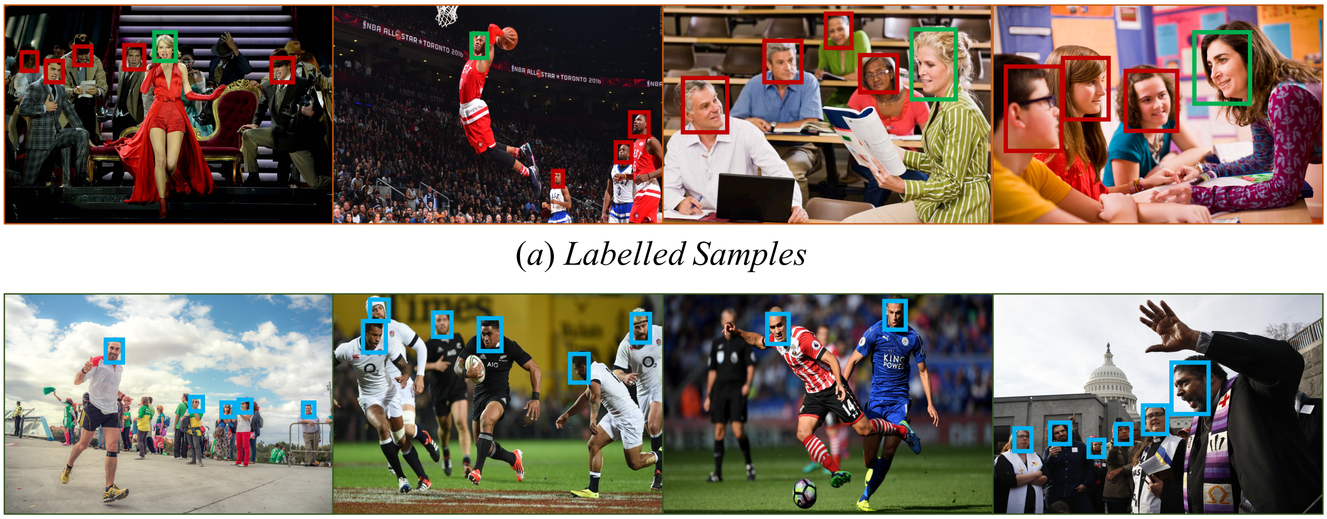
Supplementary Material for "Learning to Detect Important People in Unlabelled Images for Semi-supervised Important People Detection"

Fa-Ting Hong[1,4,5*] , Wei-Hong Li[3*] , and Wei-Shi Zheng[1,2,5†]

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Peng Cheng Laboratory, Shenzhen 518005, China
[3] VICO Group, School of Informatics, University of Edinburgh, United Kingdom
[4] Accuvision Technology Co. Ltd.
[5] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China.
hongft3@mail2.sysu.edu.cn, w.h.li@ed.ac.uk, wszheng@ieee.org

1. Details of Both EMS And ENCAA Datasets

We present some labelled and unlabelled examples of both our proposed datasets (*i.e.*, EMS and ENCAA) in Figure 1 and Figure 2.



(*a*) *Labelled Samples*

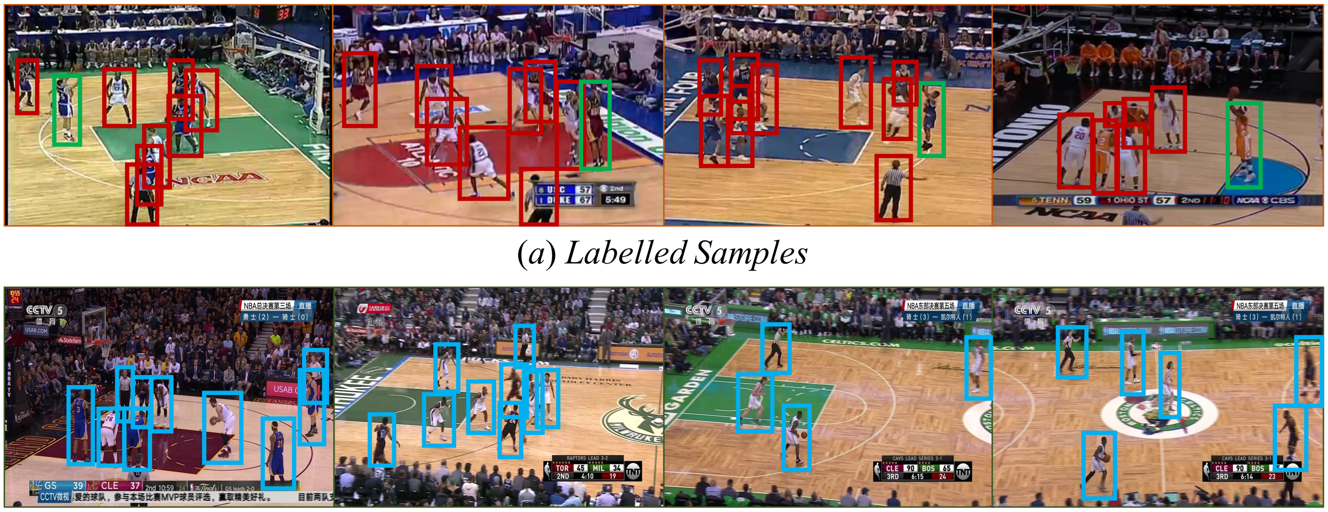
(*b*) *Unlabelled Samples*

Figure 1: Examples of EMS Dataset. People are detected by face detectors, where in labelled images, people marked with green face bounding boxes are annotated as important people and people marked by red face bounding boxes are non-important people. In unlabelled images, all people are also detected by the face detectors [3] (blue face bounding boxes).

2. Details of The Basic Model

In this work, we adopt the **deep imPOrtance relatIon NeTwork (POINT)** [4] as the basic model for our method. More specifically, POINT is an end-to-end network that automatically learns the relations among individuals in an image to encourage the network to formulate a more effective feature for important people detection. In particular, POINT has three

[*] Equal contribution.
[†] Corresponding author



(*a*) *Labelled Samples*

(*b*) *Unlabelled Samples*

Figure 2: Examples of ENCAA Dataset. People are detected by object detectors, where in labelled images, people marked with green body bounding boxes are annotated as important people and people marked by red body bounding boxes are non-important people. In unlabelled images, all people are also detected by the object detectors [6] (blue body bounding boxes).

components: 1) the **Feature Representation Module** containing ResNet-50, additional convolutional layers as well as fc layers, takes as input the social event images and detected individuals to encode features; 2) Fed with the encoded features, the **Relation Module** learns to model the relations graph among individuals and encode relation features from the graph; 3) taking as input the relation features encoded by the relation module, the **Classification Module** together with a Softmax operator transforms the relation feature into two scalar values indicating the probability of the "important" and "non-important" category.

3. Additional Explanation of Proposed Method

3.1. What if there are less than K individuals in the images

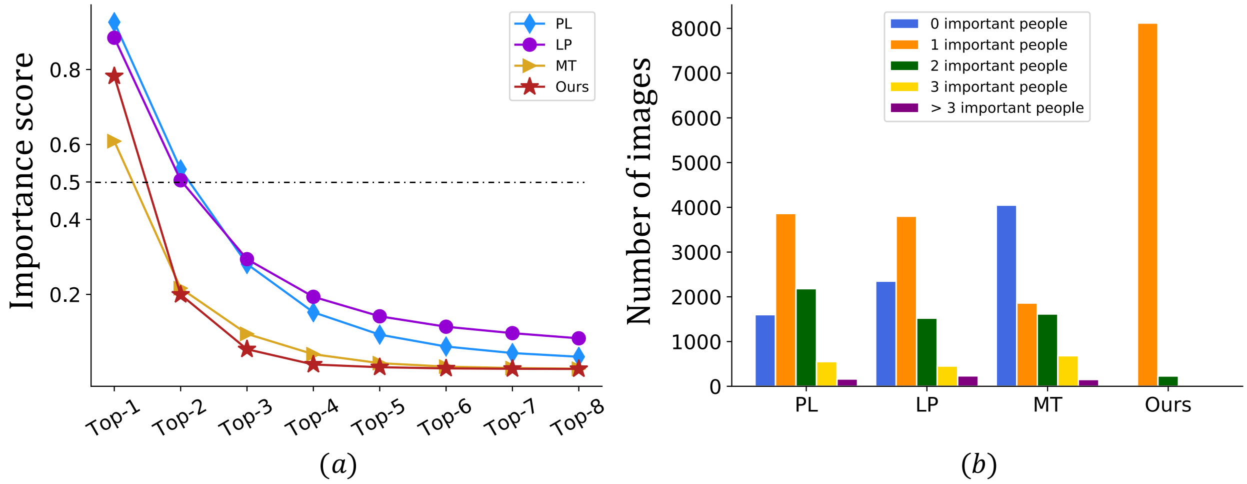
We split all detected people into important and non-important sets. If there are less than K individuals in the image, we will randomly duplicate some of the detected non-important people to ensure there are K individuals. In fact, according to POINT, there is on average $K = 8$ individuals in an image, which is the default setting for K in our paper; our method is not too much sensitive to K as the std of the performance by setting $K = 6, 8, 10, 12$ is 1.15%.

3.2. What happens if we use a pretrained model to generate pseudo-label for unlabeled data directly?

We show the experimental results of only using pretrained model (i.e. the starting model trained on available labelled data) and mixing the pseudo-label data with current labelled data, the performance is 79.80%, 80.47% and 78.12% for 33%, 66% and 100% labels on the EMS dataset, respectively. In comparison, our method obtains 87.81%, 88.44% and 89.79%, respectively. This shows that such a simple strategy is not sufficiently effective as it ignores the problem of using pseudo-label data we address.

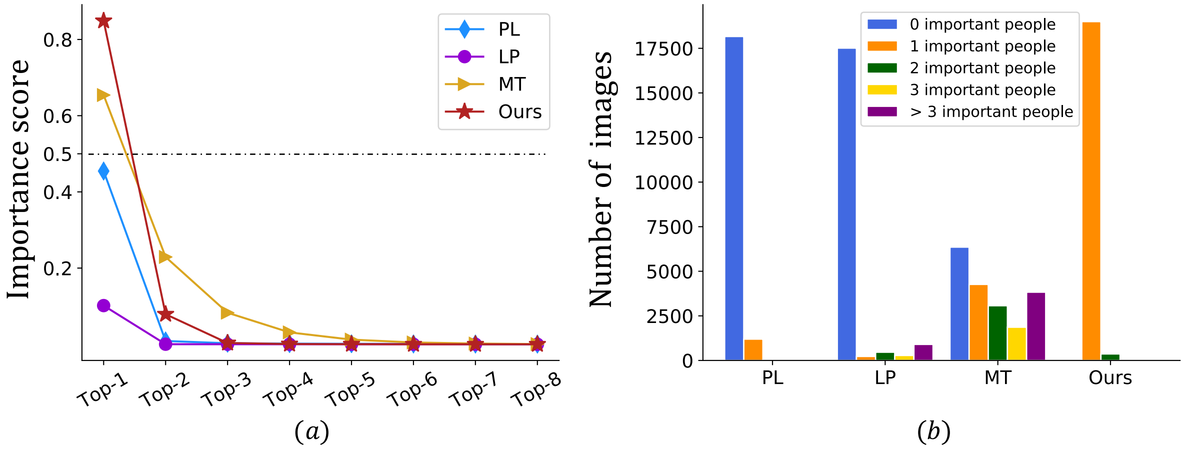
3.3. The starting labelled dataset would not cause bias.

In the paper, we reported the average results. We did implement our whole process five times, with different sets of starting labelled samples. The results with std are: $87.81 \pm 0.32\%$, $88.44 \pm 0.27\%$ and $89.79 \pm 0.30\%$ for 33%, 66% and 100% labels in the EMS, respectively.The results indicate the stability of our model.



(a) (b)

Figure 3: Fgiure (a) is the distribution of top 8 importance score in testing set in EMS datasets and Figure (b) is the statistics of unlabelled data's pseudo-labels on EMS dataset. Better view in color.



(a) (b)

Figure 4: Fgiure (a) is the distribution of top 8 importance score in testing set in ENCAA datasets, Figure (b) is the statistics of unlabelled data's pseudo-labels on ENCAA dataset. Better view in color.

4. Additional Results.

4.1. Statistics And Results on Testing Set of EMS.

We report the statistics results on ENCAA in Figure 4. In particular, Figure 4(b) shows that in LP, PL and MT, most of unlabelled images' pseudo-labels are all "non-important", resulting a bias to "non-important" class, *i.e.*, the top-1 importance scores in most testing images are less than ours or even the softmax classifier threshold (0.5). In contrast, Figure 4(b) verifies that our method prevents the problem of classifying all individuals as "non-important" pseudo-labels and thus achieves more robust results; for instance, as shown in Figure 4(a), the gap between the importance scores of the most important people predicted by our method and other people's is larger than related methods.

For EMS dataset, we calculate the number of important people in an unlabeled image according to their respective pseudo labels generated by different semi-supervised learning methods and report the statistics results in Figure 3(b). Additionally, we report the average distribution of top 8 importance scores in Figure 3(a). Note that, if there are N people in a testing

Dataset	EMS			ENCAA		
Percentage of labelled images	33 %	66 %	100 %	33 %	66 %	100 %
POINT (fully supervised)	83.36	85.97	88.48	84.60	88.21	89.75
Label Propagation (LP)	82.34	86.33	86.66	85.36	88.61	90.18
$\text{Ours}_{\text{LP}}^{\text{w/o ISW and EW}}$	86.68	87.47	88.41	88.50	90.59	91.51
$\text{Ours}_{\text{LP}}^{\text{w/o EW}}$	86.88	87.92	89.38	88.82	90.98	91.92
Ours_{LP}	**87.51**	**88.10**	**89.65**	**88.95**	**91.06**	**91.98**
Mean Teacher (MT)	84.50	86.29	87.55	83.33	84.66	87.55
$\text{Ours}_{\text{MT}}^{\text{w/o ISW and EW}}$	86.11	87.77	88.93	88.38	90.35	91.16
$\text{Ours}_{\text{MT}}^{\text{w/o EW}}$	86.59	88.29	89.49	88.95	90.63	91.37
Ours_{MT}	**87.23**	**88.56**	**90.72**	**88.97**	**90.93**	**91.62**

Table 1: Ablation study on both datasets. RankS represents ranking-based sampling while ISW and EW indicate importance score weight and effectiveness weight, respectively. $\text{Ours}_{\text{MT}}^{\text{w/o ISW and EW}}$ means our model using Mean Teacher for importance score estimation during pseudo-labelling and without using ISW and EW.

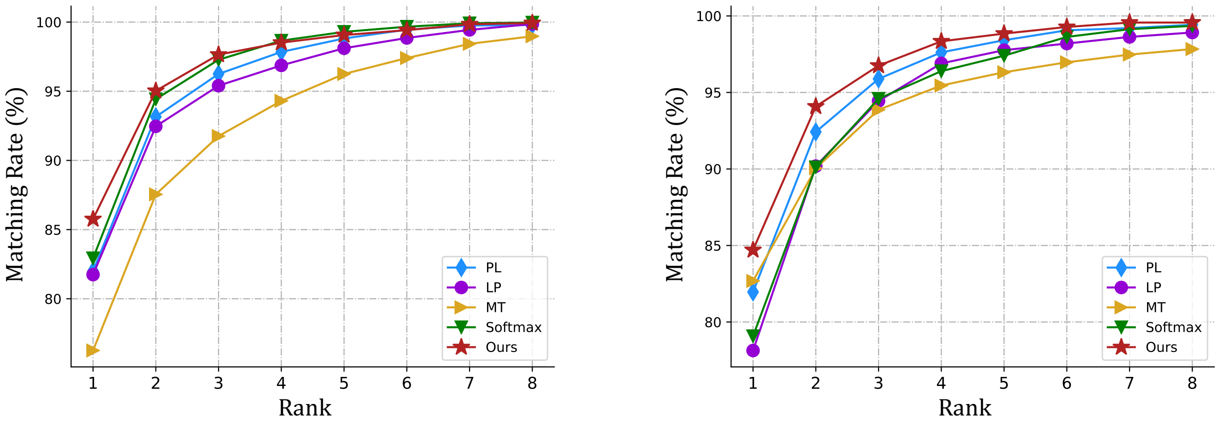
image and N is smaller than 8, we record top N importance score. From Figure 3(b), it is clear that most of unlabeled images's pseudo labels generated by related semi-supervised learning methods are all non-important (*i.e.*, the blue bar) and some of unlabelled images have more than 3 important people pseudo-labels (*i.e.*, the purple bar). This again clearly points out the imbalance pseudo-labelling problem (*i.e.*, assigning pseudo-labels of all detected persons in an image as "non-important" or "important") in recent state-of-the-art semi-supervised learning methods for important people detection. It is also demonstrated that by using our proposed method, our method yields a more reliable pseudo-labels (*i.e.*, Figure 3(b)), resulting in better testing prediction. In particular, by using our proposed strategies, we prevent the problem of assigning all "non-important" or all "important" pseudo-labels to people in unlabelled images. By using our model, on the testing data of the EMS dataset, the average most important people's score is above 0.5, the classification threshold in softmax classifier, and is higher than that of MT; the rest of people's score is below 0.5, which are classfied into "non-important". Though LP and PL can have high score for the most important people, they mis-classify the second important people whose score should be less than 0.5. Beyond this, we can see that the gap between the importance score of most important people and other people is larger than the compared methods.

4.2. Additional Ablation Study Results.

We also report additional ablation study results where we adopt Label Propagation (LP) and Mean Teacher (MT) for importance score estimation in our method during pseudo-labelling. Specially, "Ours_{MT}" implys our method using Mean Teacher for importance score estimation during pseudo-labelling. The results in Table 1 demonstrate that our method incorporating with existing semi-supervised learning is able to improve the performance over the one of the respective method, which strongly verifies that our method is generic and effective. It is also clear that three proposed strategies can consistently improves the performance regardless of which baseline is built on. These results strongly imply the effectiveness and stableness of all proposed strategies.

4.3. CMC Curves.

We also plot the Cumulative Matching Characteristics (CMC) curves [5] of different methods on EMS and ENCAA datasets (Figure 5. Compare with related semi-supervised methods, the results reported in the both figures show that ours method performs better for retrieving the important people from images, which indicates that our method is able to leverage the information at unlabelled data to benefit important people detection.



(*a*) CMC curve on EMS dataset

(b) CMC curve on ENCAA dataset

Figure 5: CMC curve on the EMS dataset (Figure (a))ENCAA dataset (Figure (b)).

4.4. Visual Comparisons

Pseudo-labels of unlabelled images generated by Different Methods. To better visualize the pseudo-labels generated by different methods, we present some examples of unlabelled images' pseudo-labels in Figure 6 and Figure 7. From both figures, it is worth noting that those recent semi-supervised would assign all "non-important" pseudo-labels to the people in an unlabelled image while our method is able to prevent this problem. Beyond this, we can clearly see that the pseudo-labels generated by our method are correct and this provides more reliable pseudo supervision for training.

The Generated Importance Score of Different Methods. To better illustrate how our method affects the results, we report people in testing images' importance score estimated by different methods in Figure 8 and Figure 9. From two figures, it is clear that our method yields correct and very robust importance score prediction for testing images. In contrast, those compared methods are unable to identify the important people in images (*e.g.*, Image 3 in Figure 8). Additionally, sometimes those compared method mis-classify all people as "non-important" category.

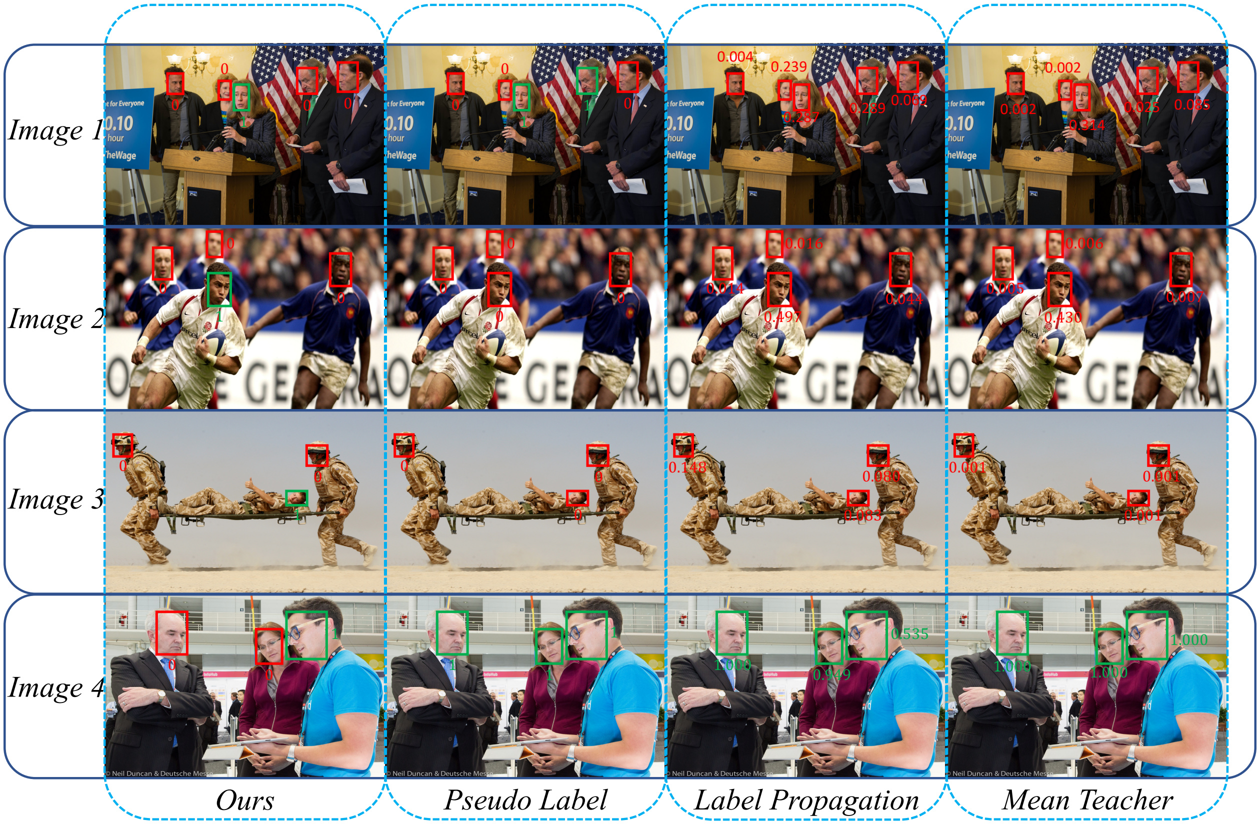


Figure 6: Pseudo-labels of unlabelled images generated by our method and the compared methods (*i.e.*Pseudo Label [2], Label Propagation [1] and Mean Teacher [7]) on EMS. The people with green bounding boxes are treated as "important" people, and people with red bounding boxes are classified as "non-important" people (the softmax classifier threshold is 0.5) during pseudo-labelling.

References

[1] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Computer Vision and Pattern Recognition*, 2019.

[2] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning Workshop on Challenges in Representation Learning*, 2013.

[3] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Computer Vision and Pattern Recognition*, 2019.

[4] Wei-Hong Li, Fa-Ting Hong, and Wei-Shi Zheng. Learning to learn relation for important people detection in still images. In *Computer Vision and Pattern Recognition*, 2019.

[5] Wei-Hong Li, Benchao Li, and Wei-Shi Zheng. Personrank: detecting important people in images. In *International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018.

[6] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[7] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, 2017.

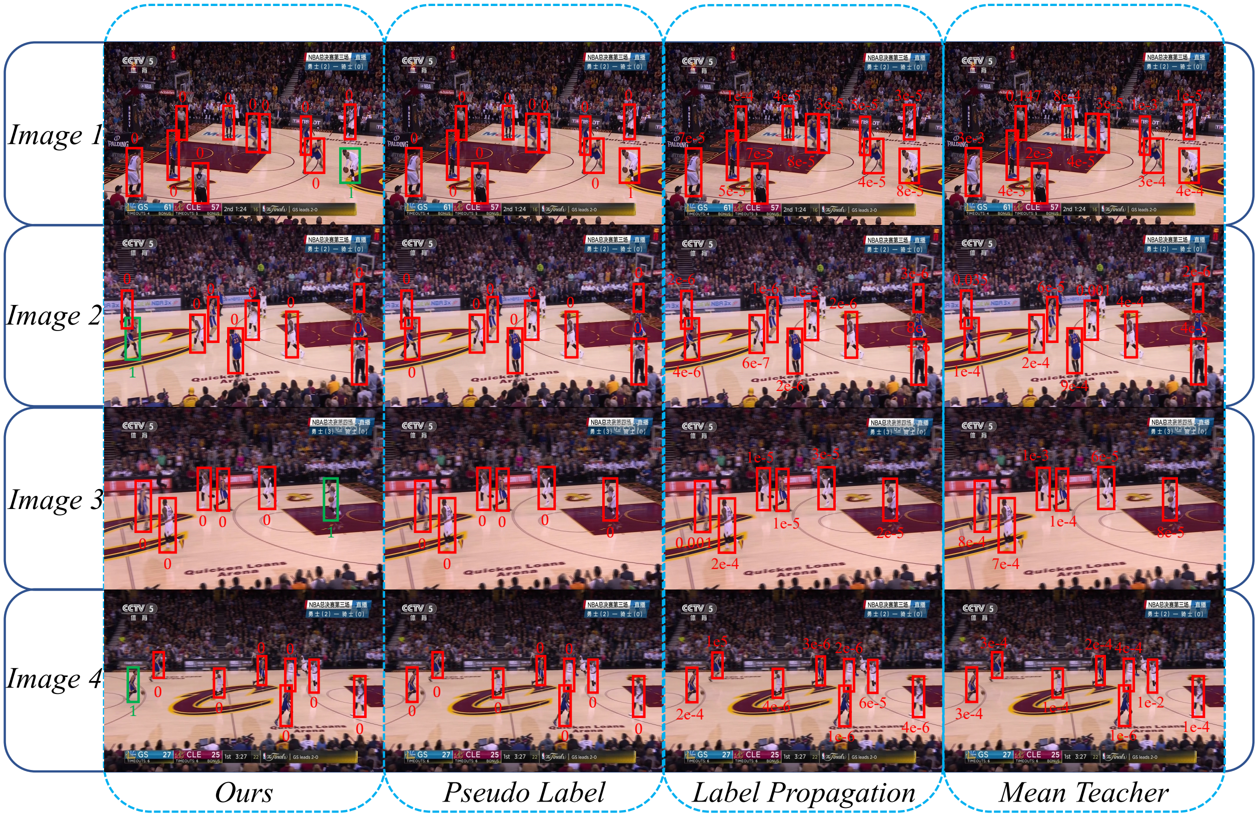


Figure 7: Pseudo-labels of unlabelled images generated by our method and the compared methods on ENCAA. The people with green bounding boxes are treated as "important" people, and people with red bounding boxes are classified as "non-important" people (the softmax classifier threshold is 0.5) during pseudo-labelling.

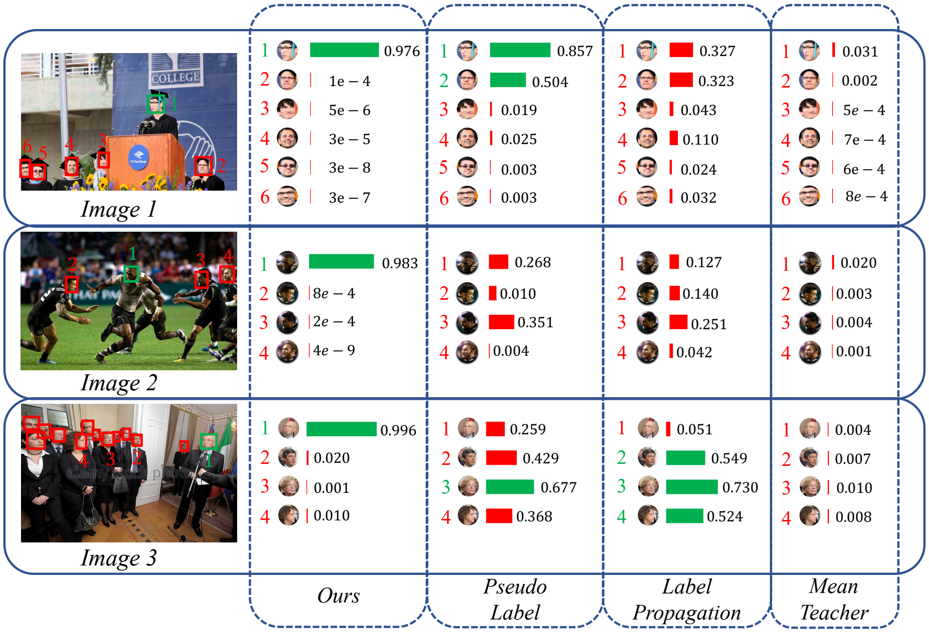


Figure 8: The importance score estimated by different methods on EMS. The images shown in the first column are testing images with ground-truth importance annotations, where people with green bounding boxes are the ground-truth "important" people and people with red bounding boxes are "non-important" people. In the second to fifth column, we report the importance score prediction of sampled people in respective testing images estimated by different methods.

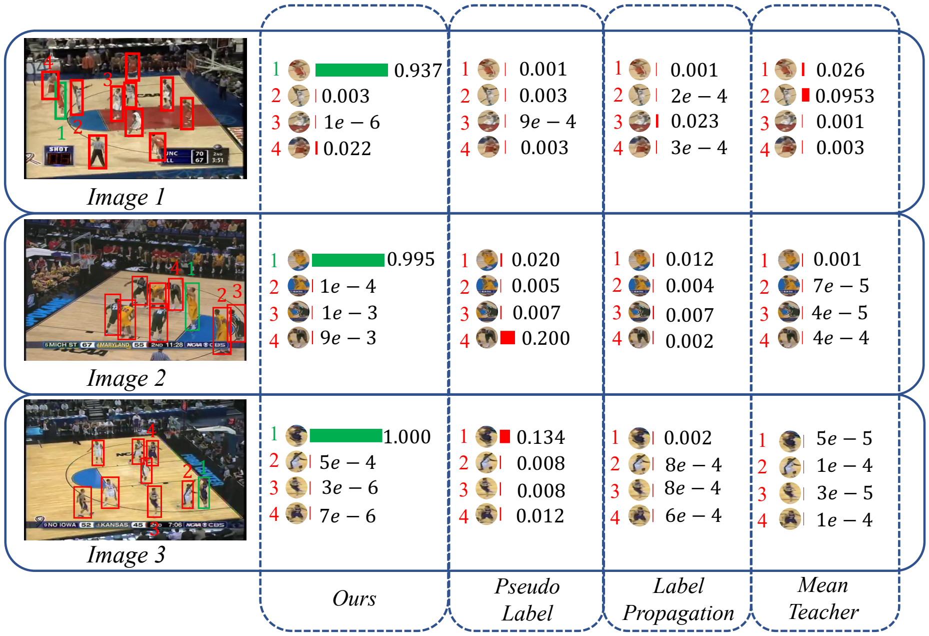


Figure 9: The importance score estimated by different methods on ENCAA. The images shown in the first column are testing images with ground-truth importance annotations, where people with green bounding boxes are the ground-truth "important" people and people with red bounding boxes are "non-important" people. In the second to fifth column, we report the importance score prediction of sampled people in respective testing images estimated by different methods.