Supplementary Material of Learning to Learn Relation for Important People Detection in Still Images

Wei-Hong Li[1,2*] , Fa-Ting Hong[1,3,4*] , and Wei-Shi Zheng[1,4†]
[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] VICO Group, School of Informatics, University of Edinburgh, United Kingdom
[3] Accuvision Technology Co. Ltd.
[4] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China.
w.h.li@ed.ac.uk, hongft3@mail2.sysu.edu.cn, wszheng@ieee.org

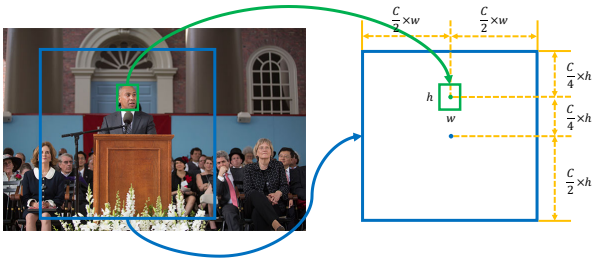

Figure 1. Illustration of extracting the exterior patch on the MS dataset.

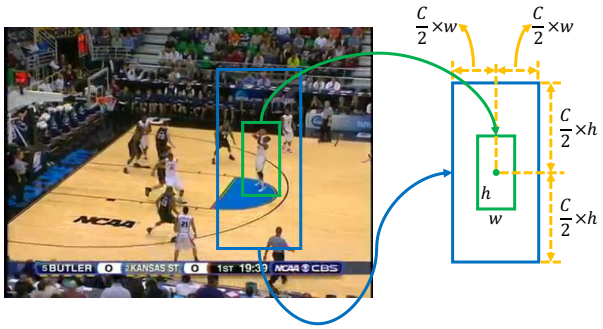

Figure 2. Illustration of extracting the exterior patch on the NCAA dataset.

1. Training Details

1.1. Experimental Settings

We implement our model using PyTorch on a machine with an E5 2686 2.3 GHz CPU, GTX 1080 Ti and 256 GB RAM. In our experiments, we adopt the ResNet-50 [1] as the backbone for feature representation. During training, the batch size is 3 (images). For each image, we randomly select 7 non-important persons and one important person for training (During test time, POINT takes as input all detected persons). Therefore, in each training batch, there are 24 person image sets (i.e., {the interior patch, the location heat map and the exterior patch}) and 3 whole images. The cross-entropy loss is applied as the final loss of our network. We then use the commonly used optimizer, the SGD, to optimize our model setting the weight decay to 5×10^{-4} and the momentum to 0.5. The learning rate is set to 2×10^{-3} at the beginning and it is controlled to decrease by a factor of 0.5 in every additional 50 epochs. On the MS dataset, $\sim 256k$ iterations(700 epochs) are performed, while $\sim 556k$ iterations(200 epochs) are performed on the NCAA dataset.

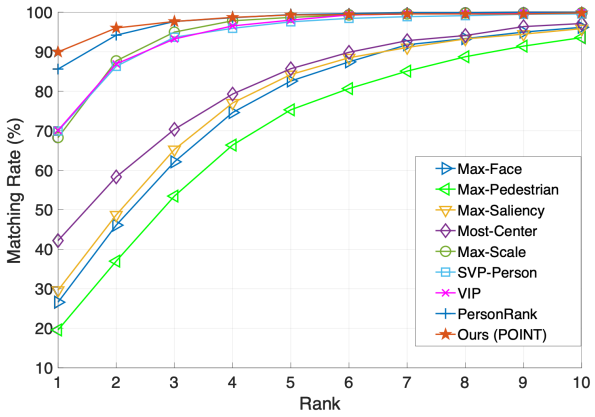

Figure 3. CMC curve on the MS Dataset.

1.2. Preprocessing

For training our POINT, we used the detected face bounding boxes and person's body bounding boxes which

*Equal contribution. Work done at Sun Yat-sen University.
†Corresponding author

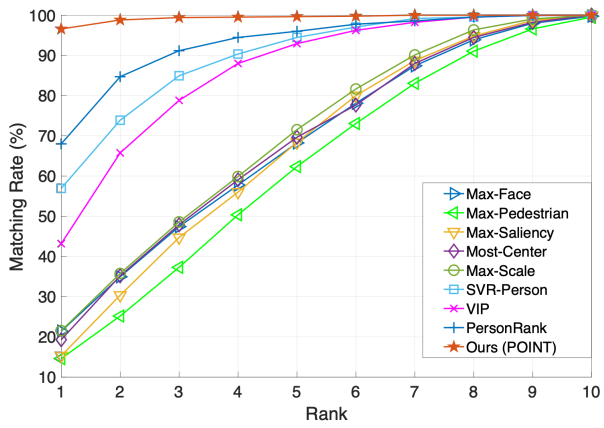

Figure 4. CMC curve on the NCAA Dataset.

are provided by the MS Dataset and the NCAA Dataset, respectively.

Interior Patch. We first crop the interior patches from the images according to the bounding boxes $[x, y, w, h]$ provided by the datasets, where $[x, y]$ are the left and top coordinates and w and h are the width and height of the bounding box, respectively.

Location Heat Map. According to the center coordinate and the scale of the bounding box, we produce a 224×224 grid (i.e., the location heat map) where one or several cells corresponding to the persons coordinate are assigned as 1 and others are zero.

Exterior Patch. Additionally, we also extract an exterior patch for each detected person according to the bounding box. On the NCAA dataset, the exterior patch centered on the center of the people's body bounding box is C^2 times larger than the scale of the bounding box (Figure 2).

Different from the NCAA dataset, the MS dataset provides the face detection bounding boxes. On the MS dataset, we extract an exterior that covers the whole body of the detected person and some contextual information, and we crop the exterior patch that is centered on the location which is $\frac{C}{4}h$ lower than the center of the bounding box. The size of the exterior patch is C^2 times larger than the face bounding box (Figure 1).

Here, C is a hyperparameter for extracting the exterior patch (i.e., the size of the exterior patch is C^2 times larger than the face/body bounding box). It is trained on the validation set (i.e., we vary C from 2 to 8 with step 1 and select the best C base on the performance on the validation set), and it is set to 8 and 6 on the MS dataset and the NCAA dataset (the selected C is fixed during test time), respectively. As we mentioned above, the way to extract the exterior patch depends on what detector we use (face or person), e.g., if a face detector is used for detection, we extract the exterior patch in the way shown in Figure 1.

2. More Experimental Results

2.1. Detailed Results on both Datasets.

For further evaluating our POINT, we provide detailed experimental results of different methods for detecting important people in different events on both datasets in Table 1 and Table 2. Overall, it is worth noting that our POINT obtains considerable improvement over the PersonRank (PR) method [3] in most events on both datasets (e.g., Our POINT obtains an improvement of 39.0 % over the PR method for the '3 point succ' event on the NCAA dataset). These results verify the efficacy of our POINT method for extracting higher level semantic feature that embraces more effective information for important people detection, compared to those customized or deep features trained for other tasks (i.e., the PersonRank[3], the VIP [5] and Ramanathan's model [4]). This also indicates the effectiveness of incorporating the relation modeling with feature learning on important people detection.

2.2. CMC Curves for Evaluation.

We also plot the CMC curves of different methods on both datasets (Figure 3 and 4). From both figures, it is clear that our POINT performs better to retrieve the important people from still images compared to the VIP [5], the PersonRank [3] and Ramanathan's model [4]. In addition, our POINT achieves 96.58 % Rank-1 matching rate on the NCAA dataset, which significantly exceeds the best existing method, the PersonRank model, which obtains 68.0 % Rank-1 matching rate. This indicates that our method learns a more effective feature for important people detection.

2.3. More Quantitative Evaluations

Classification Average Accuracy. We formulate the important people detection as a classification task. The classification average accuracy of all people tested on both datasets is also reported: 97.40 % on the MS dataset and 98.58 % on the NCAA dataset.

Evaluation of the Structure of the POINT. Compared with the baseline we constructed in Section 4.3, our POINT has more parameters than the baseline. Here, to evaluate the effectiveness of our POINT structure, we formulate another baseline which uses the same amount of parameters as our POINT, namely, Base^p. Table 3 tabulates the results of the Base^p and our POINT. From Table 3, it is clear that simply increasing the number of parameters does not yield better important people detection performance. In other words, these results indicate that the improvement achieved by our POINT is NOT caused by using more parameters, thus verifying the efficacy of our POINT structure.

Methods for Embedding Location Information. Both the individual location feature and the location relation feature are useful for important people detection, which is verified

Table 1. The mAP (%) for the Evaluation of the Different Methods on the MS Dataset

Method	Max-Face	Max-Pedestrian	Max-Saliency	Most-Center	Max-Scale	SVR-Person	VIP	PR	**Ours (POINT)**
Lecture/Speech	36.4	28.2	38.3	39.4	77.8	79.9	69.6	90.2	**94.2**
Demonstration	29.9	27.2	45.4	59.0	75.3	77.5	84.3	92.0	**93.9**
Interview	36.9	36.6	36.8	59.6	78.5	77.7	85.0	90.2	**95.1**
Sports	35.8	33.8	40.1	60.9	67.4	69.0	79.5	83.6	**87.5**
military	37.9	28.5	43.3	42.3	62.6	75.4	67.7	86.5	**88.5**
Meeting	43.2	36.2	45.1	58.6	69.0	57.5	67.9	**76.5**	75.1
Other	35.3	31.4	37.5	57.9	74.5	69.6	76.8	86.7	**91.9**
Total	35.7	30.7	40.3	50.9	73.9	75.9	76.1	88.6	**92.0**

Table 2. The mAP (%) for the Evaluation of the Different Methods on the NCAA Dataset.

Events	Max-Face	Max-Pedestrian	Max-Saliency	Most-Center	Max-Scale	SVR-Person	VIP	Ramanathan's model[4]	PR	**Ours (POINT)**
3-point succ.	35.7	29.3	12.8	14.6	26.7	56.5	47.9	51.9	71.0	**100.0**
3-point fail.	32.5	27.4	15.9	12.8	24.8	58.4	48.1	54.5	75.2	**100.0**
free-throw succ.	33.3	37.3	13.8	11.4	63.6	86.8	55.3	77.2	**94.4**	92.9
free-throw fail.	30.9	24.1	10.1	9.6	81.8	71.7	63.9	68.5	94.6	**100.0**
layup succ.	38.6	22.0	35.8	53.4	34.9	67.1	55.0	62.7	75.3	**92.5**
layup fail.	32.5	23.1	37.0	44.3	41.4	64.3	55.6	60.5	74.3	**98.1**
2-point succ.	25.6	22.1	29.9	32.2	30.7	65.9	58.6	55.4	71.6	**96.6**
2-point fail.	24.8	21.2	29.8	31.3	24.8	65.9	51.6	54.2	68.4	**95.6**
slam dunk succ.	41.8	26.6	45.2	52.2	37.0	78.4	78.3	68.6	89.7	**95.0**
slam dunk fail.	38.5	36.5	59.4	81.3	40.6	**100.0**	59.4	64.5	81.3	**100.0**
Total	31.4	24.7	26.4	30.0	31.8	64.5	53.2	61.8	74.1	**97.3**

Table 3. The mAP (%) of Our Methods and the Base^p on Both Datasets.

MS Dataset		NCAA Dataset	
Method	mAP	Method	mAP
Base^p	88.9	Base^p	95.2
Ours (POINT)	**92.0**	Ours (POINT)	**97.3**

by the results in Table 2 in the paper. We further evaluate our location embedding method with another approach in [2]. The results on both datasets are shown in Table 4. It is clearly shown that our approach for embedding location information performs better than the location embedding method in [2] for important people detection (e.g., 92.0% vs 88.9%, respectively, on the MS dataset).

Table 4. The mAP (%) for evaluating Methods of Embeding Location Information on Both Datasets.

MS Dataset		NCAA Dataset	
Method	mAP	Method	mAP
POINT$^{\text{LocaEmbed}[2]}$	89.7	POINT$^{\text{LocaEmbed}[2]}$	96.6
Ours (POINT)	**92.0**	Ours (POINT)	**97.3**

2.4. Visual Results

Visual Comparisons on Both Datasets. In this section, selected visual results are reported to further evaluate our POINT. The comparison results are shown in Figure 5. In Figure 5, compared with the PersonRank (PR), it is clear that our POINT can detect the important people in complex cases (e.g. in the second image in the second row, the defender and the shooter are very closed and our POINT can correctly assign most points to the shooter while the PersonRank (PR) usually pick the defender or other players as the most important people). This again verifies the fact that our POINT approach is able to extract higher level semantic feature that is effective for important people detection compared with those methods using customized or deep feature pretrained in other tasks. In addition, this indicates the effectiveness of incorporating the relation modeling with feature learning on important people detection.

Additionally, some failure cases are reported in Figure 6. From these images and the results in Table 1, we find that on the MS dataset, PersonRank and our POINT obtain relatively lower mAP on detecting important people in a meeting compared with other events. The reason is that there are very limited images of this event (38 images for training). In addition, in both images shown in the second row of Fig-

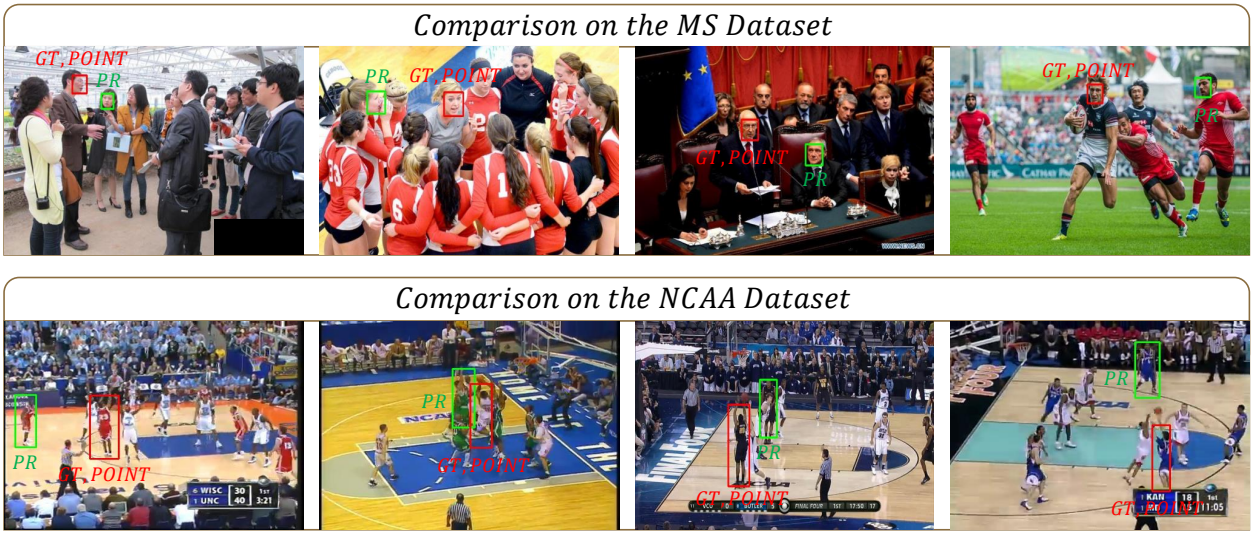

Figure 5. Selected visual results of detecting important people and comparison with related work (i.e., PersonRank (PR)) on Both Datasets.

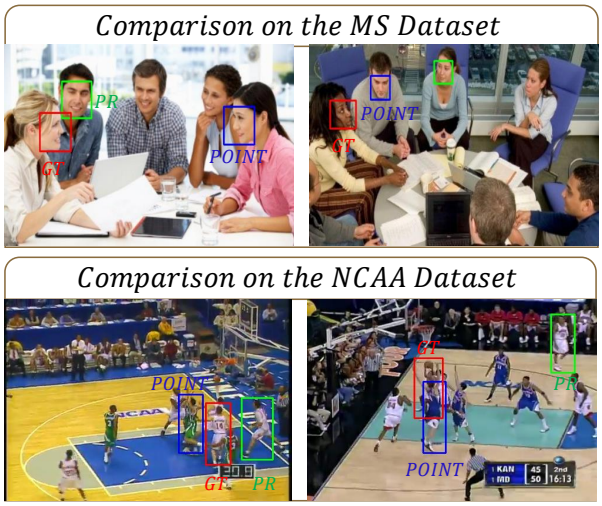

Figure 6. Failure cases on both datasets.

ure 6, both PersonRank and POINT are unable to correctly locate the important people due to the heavy occlusion.

Visual Importance Relations and Interaction Graphs. In POINT, we mainly construct two interaction graphs, the person-person interaction graph and the event-person interaction graph, and estimate the importance relations from both graphs. The visual importance relations and interaction graphs of some selected images on both datasets are reported in Figure 7 and Figure 8. For description convenience, we mainly present the interaction graphs and importance relations among the important people and three other people. From both Figure 7 and Figure 8, it is clear that our POINT is able to automatically and correctly model both the person-person interaction graph and the event-person graph and learn the importance relations among people. For instance, in Image 1 in Figure 8, both $\mathcal{V}_2^p$ and $\mathcal{V}_3^p$ are less involved in the event and the event-person interactions of both people learned by the POINT are relatively lower than both $\mathcal{V}_1^p$ and $\mathcal{V}_4^p$. In addition, the importance relations among these four players show that the $\mathcal{V}_4^p$ receives the most input importance relations and has less output importance relations than other players. This affects the relation feature modeling and importance feature model, and finally yields the result that $\mathcal{V}_4^p$ is the most important people in the image (i.e., the last row in Figure 8). These interactions and importance relations are learned automatically by the POINT without extra supervisions. These results again verify that POINT can learn to automatically model interactions and importance relations to encode importance feature for important people detection. And thus this indicates the efficacy of the relation modeling of our POINT on important people detection.

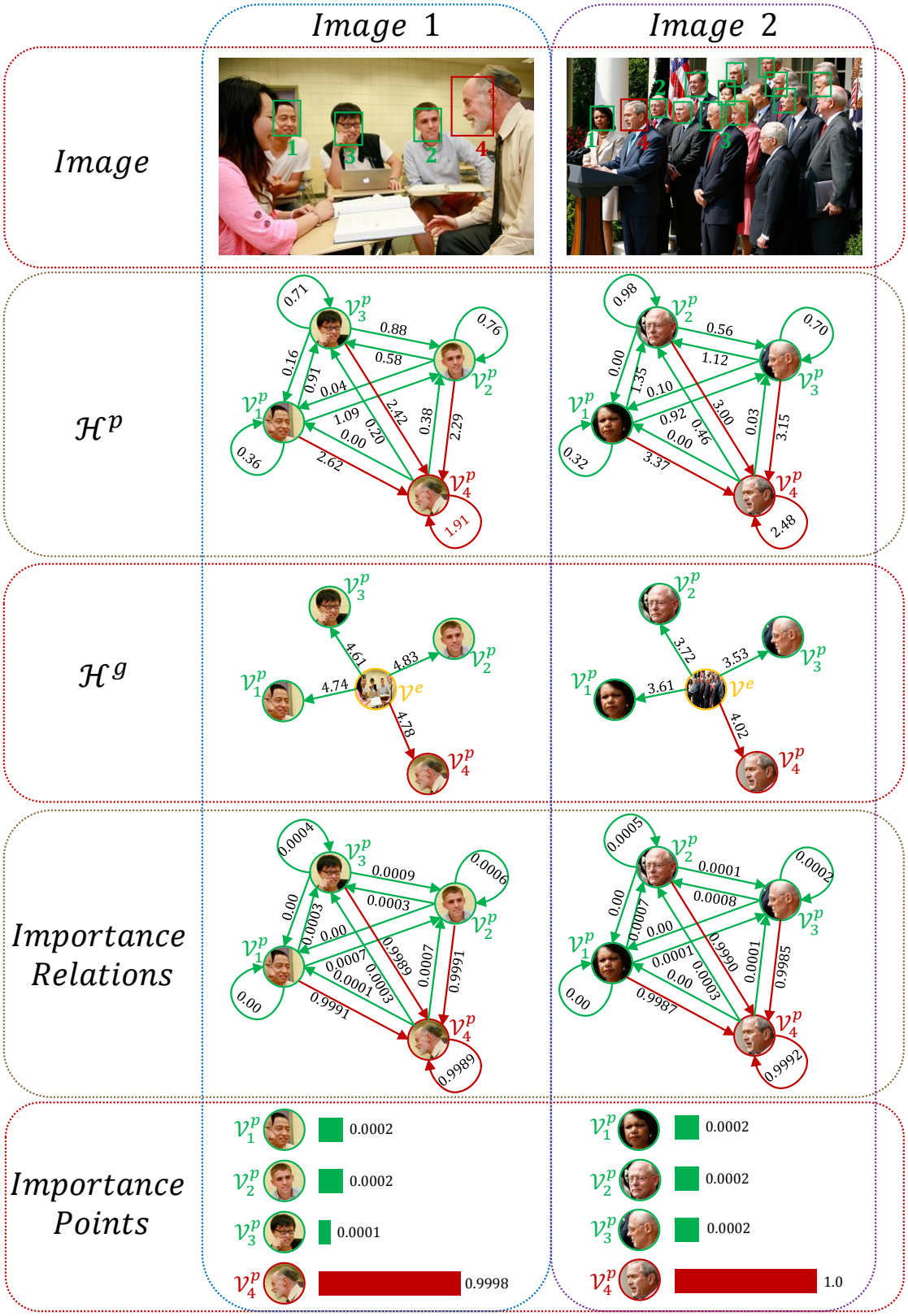

Figure 7. Visual importance relations and two interaction graphs on the MS dataset.

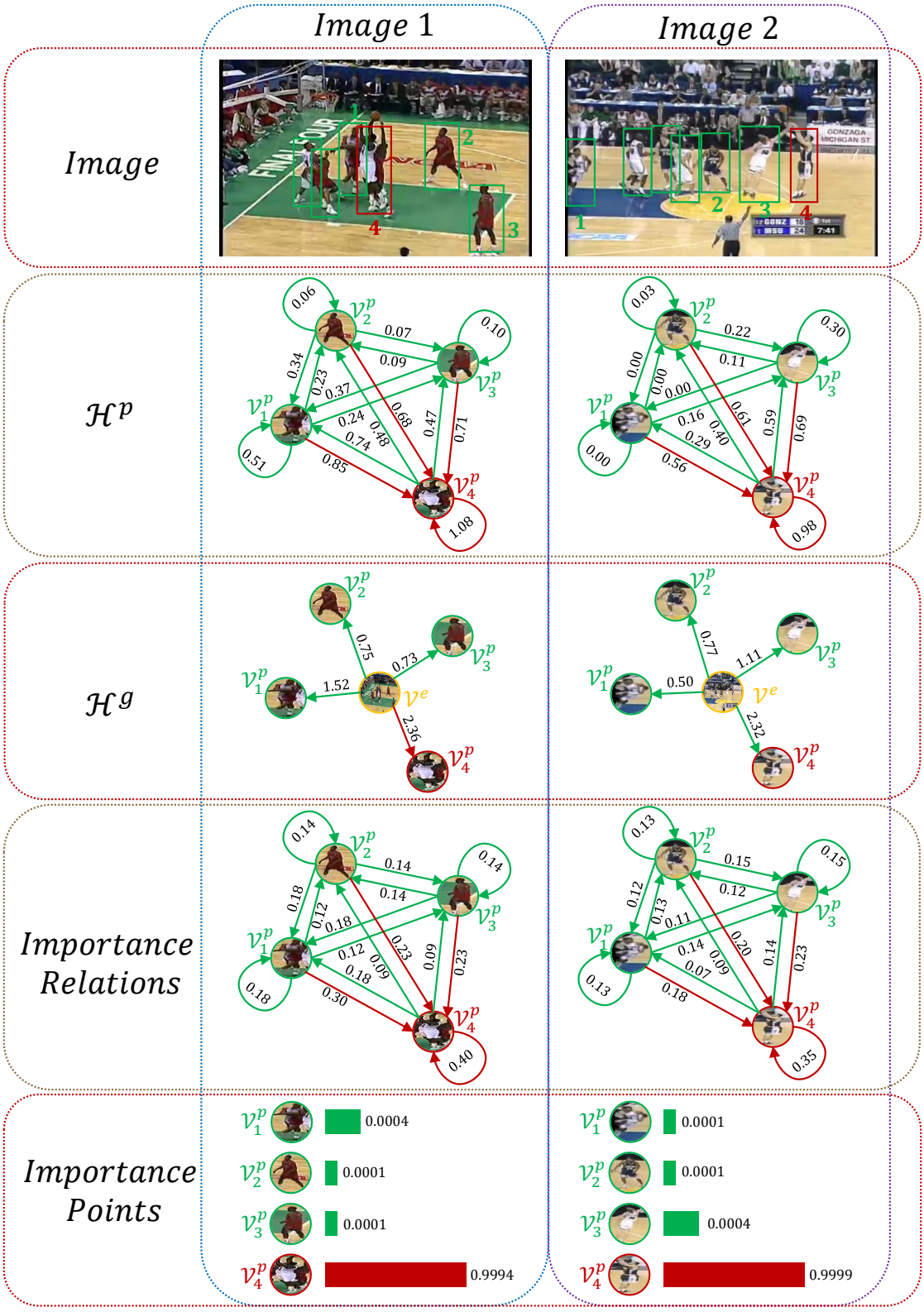

Figure 8. Visual importance relations and two interaction graphs on the NCAA dataset.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv*, 2015. 1

[2] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Computer Vision and Pattern Recognition*, 2018. 3

[3] Wei-Hong Li, Benchao Li, and Wei-Shi Zheng. Personrank: Detecting important people in images. In *International Conference on Automatic Face & Gesture Recognition*, 2018. 2

[4] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. *Computer Vision and Pattern Recognition*, 2016. 2, 3

[5] Clint Solomon Mathialagan, Andrew C Gallagher, and Dhruv Batra. Vip: Finding important people in images. In *Computer Vision and Pattern Recognition*, 2015. 2