# Learning to Learn Relation for Important People Detection in Still Images

Wei-Hong Li[1,2], Fa-Ting Hong[1] and Wei-Shi Zheng[1]
[1]Sun Yat-sen University [2]University of Edinburgh

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

## Introduction

### ★ Introduction & Motivation

* Humans can easily recognize the importance of people in social event images, and they always focus on the most important individuals. (Thinking about how we record the basketball game)

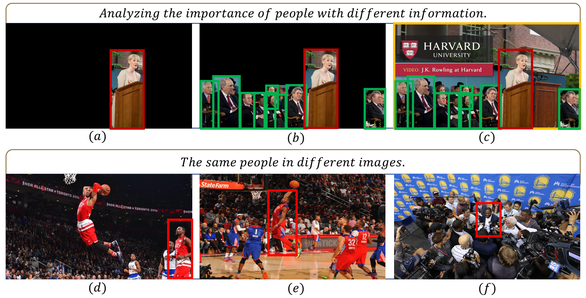
* Directly analyzing the importance from individual feature of persons is NOT enough and designing a network that can learn to model relations for important people detection remains unsolved.

### ★ Contributions

* The proposed POINT is the first to investigate deep learning for exploring and encoding the relation features and exploiting them for important people detection.

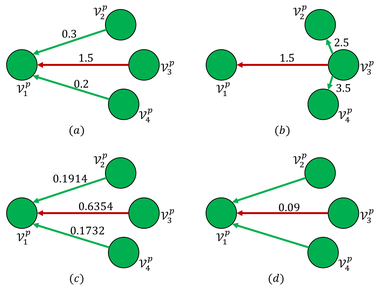
* We investigate the effect of various types of basic interaction functions on modeling pair-wise persons interactions and the effect of different types of information on important people detection.

* The POINT achieves state-of-the-art performance.

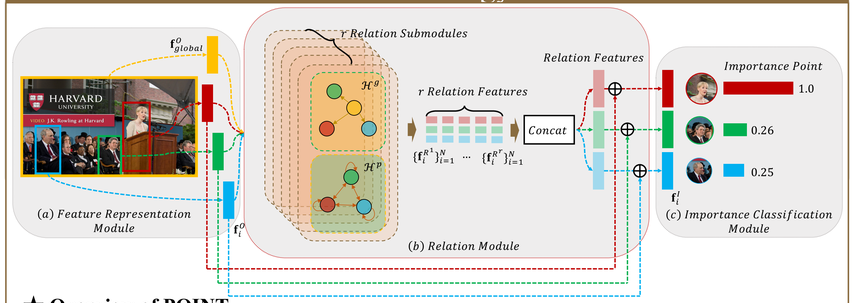



### ★ Visual comparison with other relation models

In this figure, (a) and (b) present the input person-person interactions of $\mathcal{V}_1^p$ and the output person-person interactions of $\mathcal{V}_3^p$. Our method (i.e., our relation modeling function weakens the effect of the interaction from $\mathcal{V}_3^p$ to $\mathcal{V}_1^p$ (the red link) as $\mathcal{V}_3^p$ has too many outputs (d). The attention model [18] (reference in paper) treats each node equally, and the interaction from $\mathcal{V}_3^p$ to $\mathcal{V}_1^p$ has a larger impact (c).




## Methodology



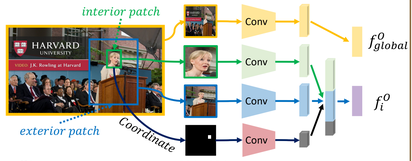

### ★ Overview of POINT

* POINT contains three main modules: feature representation module, relation module and importance classification module.

* It can be expressed as: $s_i = f^O(I, p_i|\theta^O) \circ f^R(\mathbf{f}_1, \ldots, \mathbf{f}_N, \mathbf{f}_{global}|\theta^R) \circ f^S(\mathbf{f}_i^I|\theta^S)$

### ★ Feature Representation Module

* Interior Feature
* Exterior Feature
* Global Feature
* Location feature from heat map




### ★ Relation Module

* **Person-person interaction graph:** $\mathcal{H}^p(\mathcal{V}^p, \mathcal{E}^p)$

  The person-person interaction module: $\mathcal{E}_{ji}^p = \max\{0, \boldsymbol{w}_P \cdot (\boldsymbol{W}_Q \mathbf{f}_i^O + \boldsymbol{W}_K \mathbf{f}_j^O)\}$

* **Event-person interaction graph:** $\mathcal{H}^g(\mathcal{V}^g, \mathcal{E}^g)$

  The event-person interaction module: $\mathcal{E}_i^g = \max\{0, \boldsymbol{w}_G \cdot (\mathbf{f}_i^O + \mathbf{f}_{global}^O)\}$

* **Estimate relations from two graphs**

  * Estimating the importance interaction among people as: $\hat{\mathcal{E}}_{ji}^p = \mathcal{E}_{ji}^p \cdot \mathcal{E}_i^g$

  * We estimate the relations among people by: $\mathcal{E}_{ji} = \frac{exp(\hat{\mathcal{E}}_{ji}^p)}{\sum_{k=1}^N exp(\hat{\mathcal{E}}_{jk}^p)}$

* **Encode importance feature from relations**

  * Relation feature: $\mathbf{f}_i^R = \sum_{j=1}^N \mathcal{E}_{ji} \cdot (\boldsymbol{W}_V \mathbf{f}_j^O)$

  * Importance feature: $\mathbf{f}_i^I = \mathbf{f}_i^O + Concat\left[\mathbf{f}_i^{R^1}, \ldots, \mathbf{f}_i^{R^r}\right], i = 1, \ldots, N$

### ★ Classification Module:

* two fully connected layers (i.e., $f^S(\mathbf{f}_i^I, \theta^S)$ to transform the importance feature into two scalar values indicating the probability of the person belonging to the important people or non-important people classes.

Project Page

## Experimental Results

### ★ Datasets [10]

* MS Dataset: 2310 images from more than six types of scenes.

* NCAA Dataset: 9,736 frames of an event detection video dataset [13] covering 10 different types of events.

### ★ Comparisons with existing important people detection models

Table 1. The mAP (%) of Different Methods on both Datasets

| Method | Max-Face | Max-Pedestrian | Max-Saliency | Most-Center | Max-Scale | SVR-Person | VIP | Ramanathan's model [13] | PR | **Ours (POINT)** |
|---|---|---|---|---|---|---|---|---|---|---|
| MS Dataset | 35.7 | 30.7 | 40.3 | 50.9 | 73.9 | 75.9 | 76.1 | - - | 88.6 | **92.0** |
| NCAA Dataset | 31.4 | 24.7 | 26.4 | 30.0 | 31.8 | 64.5 | 53.2 | 61.8 | 74.1 | **97.3** |

### ★ Evaluations of different components in POINT & Comparison with existing attention model.

Table 2. The mAP (%) for Evaluating Different Components of our POINT on Both Datasets.

| Dataset | Method | mAP | Method | mAP |
|---|---|---|---|---|
| MS Dataset | $\text{Base}^{\text{Inter}}$ | 72.6 | $\text{POINT}^{\text{Inter}}$ | **76.5** |
| | $\text{Base}^{\text{Inter+Loca}}$ | 79.5 | $\text{POINT}^{\text{Inter+Loca}}$ | **85.6** |
| | $\text{Base}^{\text{Inter+Exter+Loca}}$ | **89.2** | $\text{POINT}^{\text{Inter+Exter+Loca}}$ | **92.0** |
| NCAA Dataset | $\text{Base}^{\text{Inter}}$ | 89.1 | $\text{POINT}^{\text{Inter}}$ | **90.3** |
| | $\text{Base}^{\text{Inter+Loca}}$ | 89.9 | $\text{POINT}^{\text{Inter+Loca}}$ | **93.9** |
| | $\text{Base}^{\text{Inter+Exter+Loca}}$ | **95.8** | $\text{POINT}^{\text{Inter+Exter+Loca}}$ | **97.3** |

Table 4. The mAP (%) for Comparison of our Method and the one in [18] for Estimating the Importance Relation on both Datasets.

| MS Dataset | | NCAA Dataset | |
|---|---|---|---|
| Method | mAP | Method | mAP |
| Attention [18] | 90.0 | Attention [18] | 95.8 |
| Ours (POINT) | **92.0** | Ours (POINT) | **97.3** |

### ★ Evaluation of $r$ & $N_r$ in POINT

Table 5. The mAP (%) for Evaluating the Effect of $r$ on Both Datasets

| Dataset | Baseline | Ours (POINT) | | | | | |
|---|---|---|---|---|---|---|---|
| | | $r$=1 | $r$=2 | $r$=4 | $r$=8 | $r$=16 | $r$=32 |
| MS Dataset | 89.2 | 90.7 | 91.4 | **92.0** | 91.4 | 91.8 | 91.4 |
| NCAA Dataset | 95.8 | 96.2 | 96.8 | **97.3** | 96.8 | 97.0 | 96.6 |

Table 6. The mAP (%) for Evaluating the Effect of $N_r$ on Both Datasets

| Dataset | Baseline | Ours (POINT) | | | |
|---|---|---|---|---|---|
| | | $N_r$=1 | $N_r$=2 | $N_r$=4 | $N_r$=6 |
| MS Dataset | 89.18 | 91.96 | **91.97** | 90.99 | 90.90 |
| NCAA Dataset | 95.84 | 97.28 | 97.24 | **97.29** | 96.02 |

### ★ Evaluation of different attention function for modeling interactions.

Table 7. The mAP (%) for Evaluating Different Types of Attention Functions on both Datasets.

| MS Dataset | | NCAA Dataset | |
|---|---|---|---|
| Method | mAP | Method | mAP |
| $\text{POINT}^{\text{Scaled Dot Product}}$ | 90.7 | $\text{POINT}^{\text{Scaled Dot Product}}$ | 96.2 |
| $\text{POINT}^{\text{Additive}}$ | **92.0** | $\text{POINT}^{\text{Additive}}$ | **97.3** |

### ★ Visual Comparisons

* POINT can can detect the important people in some complex cases (e.g. in the both 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 player as the important people.