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

Abstract

Humans can easily recognize the importance of people in social event images, and they always focus on the most important individuals. However, learning to learn the relation between people in an image, and inferring the most important person based on this relation, remains undeveloped. In this work, we propose a deep imPOrtance relatIon NeTwork (POINT) that combines both relation modeling and feature learning. In particular, we infer two types of interaction modules: the person-person interaction module that learns the interaction between people and the event-person interaction module that learns to describe how a person is involved in the event occurring in an image. We then estimate the importance relations among people from both interactions and encode the relation feature from the importance relations. In this way, POINT automatically learns several types of relation features in parallel, and we aggregate these relation features and the person's feature to form the importance feature for important people classification. Extensive experimental results show that our method is effective for important people detection and verify the efficacy of learning to learn relations for important people detection.

1. Introduction

In our daily lives, we often see wonderful live broadcasting as the cameraman can easily recognize the importance of people in an event and take shots or videos of the important people in the event to present what is occurring at the moment. Additionally, when a social event image is presented, humans can easily recognize the distinct importance of different faces (persons) in the event and focus on the most important people (e.g., when people are watching a basketball game, they are more likely to focus on the shooter or the player with the basketball). It is natural to ask

---

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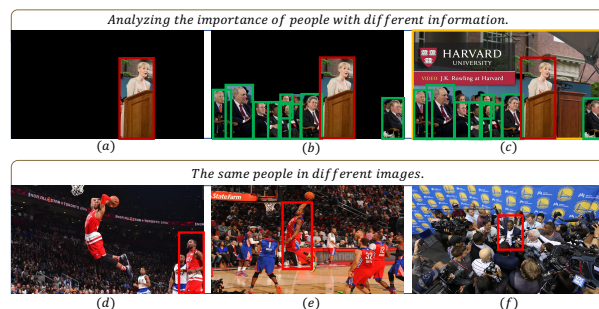†Corresponding author

Figure 1. Inferring the importance of persons from an image is inherently complex and difficult as it relates to diverse information (i.e., individual features of persons (Figure (a)), relations among persons (Figure (b)) and the event information (Figure(c)) from the whole image). The great visual variations lead to difficulties as well. The person in the red bounding box in all the images shown in the second row is the same person while he plays diverse roles in these images. He is the most important person in Figure (e) and (f) while his appearance, location and the event in both images are completely distinct. Comparing between Figure (d) and Figure (e), he wears the same clothes in both images but his importance in these images is different.

whether a computer vision model can be built to automatically detect the important people in event images. It is also known that correctly detecting the most important people in images can benefit other vision tasks such as event detection [13], event/activity recognition [13, 17] and image captioning [14].

Important people detection has only recently become the focus of research. To detect the important people in still images, a straightforward approach is to exploit classification or regression models to infer the importance of people directly from their individual features [14]. Another solution considers the relations among persons by estimating their interaction sequentially (i.e., sequential relation models [14, 10]). Solomon et al. [14] studied the relative importance between a pair of faces, either in the same image or separate images, and developed a regression model to pre-

dict the relative importance between any pair of faces using manually designed features. Li et al. [10] modeled all previously detected people in a hybrid interaction graph and developed PersonRank, a graphical model to rank the people from the interaction graphs.

Despite these efforts on important people detection, the problem remains challenging, as the **importance** of people is related to not only their appearance but also, more importantly, the relations among the people. Only relying on the appearance features is not effective. For instance, we would be unable to determine whether the lady with the red bounding box in Figure 1(c) is important or not if we were given the patch inside the red bounding box as shown in Figure 1(a). However, if we know who and how others individuals are interacting with the lady (Figure 1), it becomes easier to separate the lady from the others. Although relation modeling is important, the relation between two people in an image is still determined by customized features (e.g., [14, 10]). The customized features are highly affected by variations in pose, appearance and actions. How to automatically exploit the reliable and effective relation features that describe the relations between people is still unsolved.

In this work, we cast the important people detection problem as learning the relation network among detected people in an image and inferring the most active person there. Thus, we attempt to develop a deep imPOrtance relatIon NeTwork (POINT) to allow machine learning to exploit the relations automatically. In POINT, we mainly introduce the relation module, which contains several relation submodules to automatically construct interaction graphs and model their importance relations from the interaction graphs. In each relation submodule, we form two types of interaction modules, the person-person interaction module and the event-person interaction module. The person-person interaction module describes the pairwise person interactions and the event-person interaction module indicates the probability of a person being involved in the event. We then introduce two methods to estimate the importance relations among persons from both the interaction graphs and encode the relation feature based on the importance relations. Finally, we concatenate the relation features from all relation submodules into one relation feature and employ the residual connection to aggregate the concatenated relation feature and the person feature, resulting in the importance feature for the final importance classification. In summary, the POINT method is a classification framework consisting of a feature representation module, a relation module and an importance classification module.

To the best of our knowledge, POINT is the first to investigate deep learning for exploring and encoding the relation features and exploiting them for important people detection. In our experiments, we investigate and discuss the effect of various types of basic interaction functions (i.e., additive function and scaled dot product function) on modeling pairwise persons interactions and the effect of different types of information on important people detection. The experimental results show that our deep relation network achieves state-of-the-art performance on two public datasets and verify its efficacy for important people detection.

2. Related Work

Persons and General Object Importance. Recently, the importance of generic object categories and persons has attracted increased attention and has been studied by several researchers [2, 5, 6, 7, 15, 8, 14, 10]. Solomon et al. [14] focused on studying the relative importance between a pair of faces, either in the same image or separate images, and developed a regression model for predicting the importance of faces. The authors designed customized features containing spatial and saliency information of faces for important face detection. In addition, Ramanathan et al. [13] trained an attention-based model with event recognition labels to assign attention/importance scores to all detected individuals to measure how related they were to basketball game videos. More specifically, they proposed utilizing spatial and appearance features of persons including temporal information to infer the importance score of all detected persons. Recently, Li et al. [10] modeled all detected people in a hybrid interaction graph by organizing the interaction among persons sequentially and developed PersonRank, a graphical model to rank the persons by inferring the importance scores of persons from person-person interactions constructed on four types of features that have been pretrained for other tasks.

Different from the aforementioned methods, which design both handcrafted relations as well as features, or those pretrained for other tasks, as far as we know, our work is the first to design a deep architecture to combine the learning of relations and features for important people detection. The relation module is learned to construct interaction graphs and automatically encode relation features. Thus, our network can not only encode more effective features from a persons individual information but also efficiently encode the relations from other people and the event in the image.

Relation Networks on Vision Tasks. Relation modeling is not limited to important people detection and has broad application, such as object detection [4], AI gaming [22], image captioning [20], video classification [19], and few-shot recognition [21]. Related to our method, Hu et al. [4] proposed adapting the attention module by embedding a new geometric weight and applying it in a typical object detection CNN model to enhance the features for object classification and duplicate removal. Zambaldi et al. [22] exploited the attention module to iteratively identify the relations between entities in a scene and to guide a model-free policy in a novel navigation and planning task called Box-World.
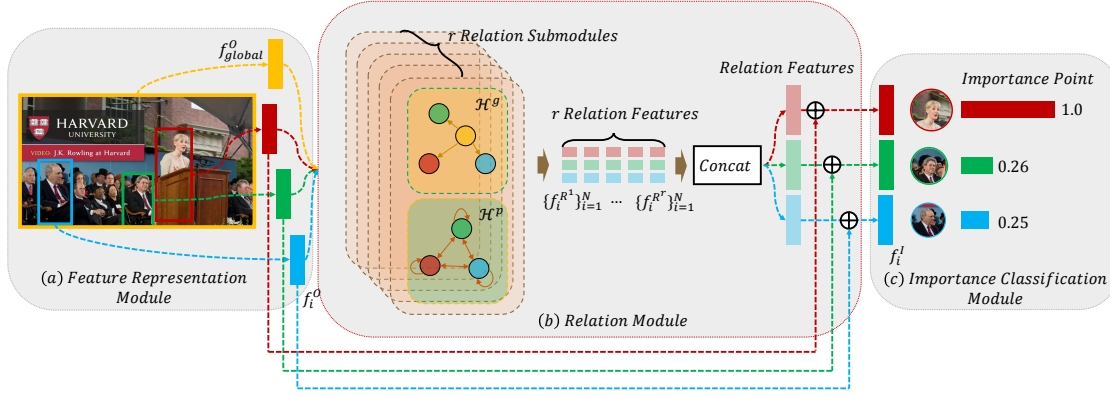
Figure 2. An illustration of our deep imPOrtance relatIons NeTworks (POINT). We exploit the feature representation module to extract the person feature for persons and the global feature for the whole image (Figure (a)). These features are fed into the relation module, which contains r relation submodules. In each relation submodule, we construct two interaction graphs and estimate importance relations from both graphs, which are used for encoding relation features. In this way, the POINT learns r relation features in parallel and these features are concatenated into a relation feature vector. We add this concatenated relation feature to the person feature, resulting in the importance feature. Finally, the importance classification module is employed to infer the importance point of people.

In this work, we have the different purpose of building a relation network for important people detection, while the related relation models are not suitable for our task. In particular, in previous works, they learn the relation that describes the appearance and location similarity between two objects/entities to find the similar objects. These relational models will bias the important people detection model to detect the people with certain appearance or the people in a specific location, but not for purpose of telling how people are interacting with each other and who is the most active one. In our experiments, we have shown only using appearance features or specific location is not effective for important people detection (see Table 1, the SVR-person only using appearance and location information is not effective.) For estimating the important relations, we introduce two interaction modules (i.e. person-person and event-person interactions) to learn the interactions that describe the relation between two people and how people are involved in the event occurring in an image automatically.

3. Approach

Detecting important people in still images is a more challenging task than conventional people detection as it requires extracting higher semantic information than other detection tasks. In this work, under the same setting as that in previous works [13, 14, 10][1], we aim to design a deep relation network called the deep imPOrtance relatIons NeTwork (POINT) (Section 3.1), which learns to build the relations and combines the relation modeling with feature learning for important people detection. We briefly introduce the ar-

---
[1]Similar to the aforementioned works [13, 14, 10], we assume that all persons appearing in images are successfully detected by existing state-of-the-art person (face or pedestrian) detectors.
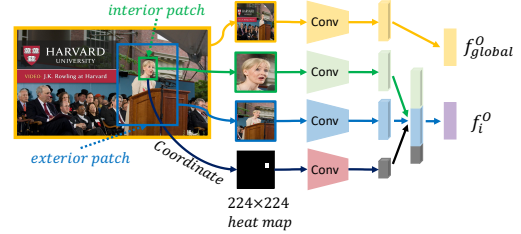
Figure 3. The feature representation module.

chitecture of the proposed POINT (Section 3.1) before detailing three specific modules and loss (Section 3.2, Section 3.3 and Section 3.4).

3.1. Overview

An illustration of our proposed model's architecture is shown in Figure 2. Given a social event image $\mathbf{I}$ and all detected (N) persons $\{\mathbf{p}_i\}_{i=1}^{N}$, to analyze the importance of these persons, we build our POINT as a classification pipeline. Our model processes an arbitrary number of detected people in parallel (as opposed to sequential relation modeling [14, 10]) and is fully differentiable (as opposed to the previous relation models using customized features [10]). For the ith person $\mathbf{p}_i$ in an image, its label (i.e., important or non-important person) $\mathbf{s}_i$ is estimated by:

$$\mathbf{s}_i = f^O(\mathbf{I}; \mathbf{p}_i | \theta^O) \circ f^R(\mathbf{f}_1^O, ..., \mathbf{f}_N^O, \mathbf{f}_{global}^O | \theta^R) \circ f^S(\mathbf{f}_i^I | \theta^S), \ (1)$$

where $\circ$ denotes module composition, $\mathbf{f}_i^I$ is the importance feature of $\mathbf{p}_i$ and $f^S(\mathbf{f}_i^I | \theta^S)$ is the importance classification module parameterized by θ^S, which follows the relation module $f^R(\cdot)$ parameterized by the parameter groups θ^R. In addition, the feature representation module $f^O(\mathbf{I}; \mathbf{p}_i | \theta^O)$ parameterized by θ^O is employed to extract the person fea-
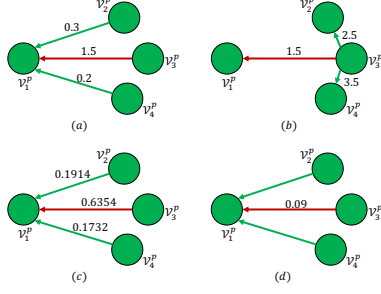
Figure 4. Figure (a) and (b) present the input person-person interactions of $\mathcal{V}_1^p$ and the output person-person interactions of $\mathcal{V}_3^p$. Our method (i.e., Eq. (4)) weakens the effect of the interaction from $\mathcal{V}_3^p$ to $\mathcal{V}_1^p$ (the red link) as $\mathcal{V}_3^p$ has too many outputs (Figure (d)). The attention model [18] treats each node equally, and the interaction from $\mathcal{V}_3^p$ to $\mathcal{V}_1^p$ has a larger impact (Figure (c)).

ture $\mathbf{f}_i^O$ of $\mathbf{p}_i$ and the global feature $\mathbf{f}_{global}^O$ of the whole image $\mathbf{I}$ and $\circ$ is the operator to connect three modules.

The relation module $f^R(\cdot)$ exploits the input features of persons $\{\mathbf{p}_i\}_{i=1}^N$ and the global features $\mathbf{f}_{global}^O$ to automatically construct interaction graphs and encode effective relation features. Similar to existing attention modules [18] and relational modules [4, 22], we adopt the residual connection to aggregate the person feature and the relation feature, resulting in a final importance feature $\mathbf{f}_i^I$, which comprises individual information, the relation information from other persons and the event information in an image. The details of each module are described in Section 3.2, Section 3.3 and Section 3.4.

3.2. Feature Representation Module

Since feature representation is the first step in important people detection, we require the feature representation module (Figure 3) to be capable of extracting effective features from local to global information (i.e., the people's interior/individual information, the exterior/contextual information around the people and the global information illustrating the event cues). As with most vision works [16, 10, 13, 14], it is natural to use the information inside the bounding box of the detected person, called the interior patch in this work, to represent the person's interior/individual feature. The location is also an indispensable element of a person's individual feature for illustrating the importance of the person and the coordinate of the person in the image is included in our feature. The reason is, from the photographer's perspective, when the images of an event are captured, the photographer tends to place the important people in the center of the image, and the important people usually look clearer than other people in the image. Additionally, the exterior/contextual information around each person must be considered for analyzing the importance of persons as this more global information, for instance, some

objects that the person uses can aid in distinguishing the important people from the non-important people. For this purpose, for each person, we crop an exterior patch[2], which is an image patch inside a box that is centered on the person's bounding box and is C^2 times larger than the scale of the person's bounding box.

In this work, we use the ResNet-50 to extract features from each interior and exterior patch because it has demonstrated its superiority in terms of important people detection [10] and other vision tasks such as object detection [11]. As shown in Figure 3, for each person in an image, we feed the interior and the exterior patches into separate Resnet-50s, transforming them into two $7 \times 7 \times 2048$ features (i.e., the interior feature and the exterior feature). While the coordinate is a four dimensional vector, we produce a heat map, which is a 224×224 grid where one or several cells correspond to the person's coordinate are assigned as 1 and the others zero. We apply convolutional kernels to this heat map to produce a $7 \times 7 \times 256$ feature. Then, we concatenate the interior, the exterior and the location features, resulting in a $7 \times 7 \times 4352$ feature and employ two convolutional layers with one fully-connected (fc) layer to transform this concatenated feature into a 1024 dimensional vector $\mathbf{f}_i^O$, called the *person feature*.

As the important person is inevitably related to the event that the person is involved in, the global information that represents this event should be considered as well. Similar to the interior and the exterior features, the whole image (denoted as the global patch) is fed into another deep network, which comprises the convolutional layers of the ResNet-50, two additional convolutional layers and one fc layer for encoding a 1024 dimensional $\mathbf{f}_{global}^O$. We call this feature the *global feature*.

3.3. Relation Module

Given the person feature and the global feature, we aim to design a relation module that can encode the effective importance feature by aggregating the relation feature and the person feature. More specifically, we aggregate r relation features encoded by r parallel **relation submodules**[3] and concatenate them into one relation feature vector. Then, we employ the residual connection to merge the relation feature and the person feature, yielding the *importance feature* for each person $\mathbf{p}_i$:

$$\mathbf{f}_i^I = \mathbf{f}_i^O + Concat[\mathbf{f}_i^{R^1}, \cdots, \mathbf{f}_i^{R^r}], (i = 1, \cdots N), \quad (2)$$

where $\mathbf{f}_i^{R^1}$ is a relation feature of person $\mathbf{p}_i$ computed by the first relation submodule. We use this parallel structure

---

[2]In this work, C is trained on the validation set. Details of extracting the exterior patch and C are reported in the Supplementary Material

[3]The structure of these relation submodules are the same while the parameters are NOT shared, which enables POINT to automatically learn various types of relations.
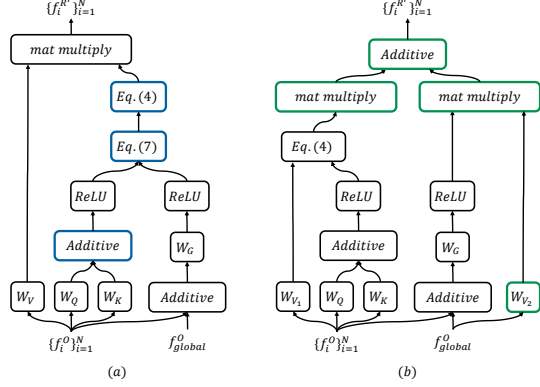
Figure 5. Figure (a) and (b) illustrate two methods introduced in this work to embed global information into the person-person interaction (they are the illustrations of the relation submodule as well). Figure (a) is the method using Eq. (3) and Eq. (7) while Figure (b) is the method using Eq. (8). The blue rectangle boxes show the difference between our method and the attention model [18] and the relation module in [4, 22] while the green boxes illustrate the difference between the two methods we proposed. (Better viewed in color).

because it it allows POINT to automatically model various types of people relations and has been shown to be more effective in our work and others [4, 18]

Relations Modeling in the Relation Submodule. We now describe our importance relation computation in the ℓth ($\ell = 1, ..., r$) relation submodule. For each given image with N detected persons, we obtain a feature set $\{\mathbf{f}_1^O, ..., \mathbf{f}_N^O, \mathbf{f}_{global}^O\}$, and then the relation feature $\mathbf{f}_i^R$ with respect to the ith person is computed by

$$\mathbf{f}_i^R = \sum_{j=1}^{N} \mathcal{E}_{ji} \cdot (\mathbf{W}_V \mathbf{f}_j^O). \tag{3}$$

Here, we remove the superscript of $\mathbf{f}_i^{R^\ell}$ and use $\mathbf{f}_i^R$ for description convenience. The output of Eq. (3) aggregates the feature from the others by a weighted sum of the person features from the other people, and is linearly transformed by $\mathbf{W}_V$. We formulate $\mathcal{E}_{ji}$, the *importance relation* indicating the impact from the other people by:

$$\mathcal{E}_{ji} = \frac{exp(\hat{\mathcal{E}}_{ji}^p)}{\sum_{k=1}^{N} exp(\hat{\mathcal{E}}_{jk}^p)}, \tag{4}$$

where $\hat{\mathcal{E}}_{ji}^p$ is the *importance interaction* among persons and introduced in the following, and it is estimated from both the person-person interaction graph and the event-person interaction graph. Here, we compute the importance relation from person $\mathbf{p}_j$ to person $\mathbf{p}_i$ as the importance interaction from person $\mathbf{p}_j$ to person $\mathbf{p}_i$ scaled by the summation of the output importance interactions of person $\mathbf{p}_j$. Inspired by the PageRank algorithm [9], our model reflects the fact that

an importance interaction from a node that has too many importance interaction outputs is less important, and this weakens the effect of the importance interaction on the importance relation (Figure 4).

Constructing Interaction Graphs. In order to estimate the importance interaction $\hat{\mathcal{E}}_{ji}^p$, we first create the *person-person interaction graph* and *event-person interaction graph*, which are defined as $\mathcal{H}^p = (\mathcal{V}^p, \mathcal{E}^p)$ and $\mathcal{H}^g = (\mathcal{V}^g, \mathcal{E}^g)$, respectively. Here, $\mathcal{V}^p = \{\mathcal{V}_i^p\}_{i=1}^N$ are nodes representing persons and $\mathcal{V}^g = \{\mathcal{V}_i^p\}_{i=1}^N \cup \{\mathcal{V}^e\}$ are nodes in $\mathcal{H}^g$, where $\mathcal{V}^e$ is a node representing the event occurring in the image. In addition, each element $\mathcal{E}_{ji}^p$ in $\mathcal{E}^p$ models the person-person interaction from $\mathbf{p}_j$ to $\mathbf{p}_i$ indicating how $\mathbf{p}_j$ is interacting with $\mathbf{p}_i$, and each element $\mathcal{E}_i^g$ in $\mathcal{E}^g$ represents the event-person interaction indicating the probability of a person being involved in the event.

In the person-person interaction graph $\mathcal{H}^p$, the interaction between pairwise persons is computed by the **person-person interaction module**, which is an additive attention function [3, 1][4]:

$$\mathcal{E}_{ji}^p = max\{0, \mathbf{w}_P \cdot (\mathbf{W}_Q \mathbf{f}_i^O + \mathbf{W}_K \mathbf{f}_j^O)\}, \tag{5}$$

where both $\mathbf{W}_Q$ and $\mathbf{W}_K$ are matrices that project the person features $\mathbf{f}_i^O$ and $\mathbf{f}_j^O$ into subspaces and the vector $\mathbf{w}_P$ is applied to measure how $\mathbf{p}_j$ is interacting with $\mathbf{p}_i$ in the subspace. Additionally, the $max\{\cdot\}$ function is employed to trim the person-person interaction at zero if the person is not interacting with the other person.

In the meantime, we estimate the event-person interaction by the **event-person interaction module** [5]:

$$\mathcal{E}_i^g = max\{0, \mathbf{w}_G \cdot (\mathbf{f}_i^O + \mathbf{f}_{global}^O)\}, \tag{6}$$

where $\mathbf{f}_i^O + \mathbf{f}_{global}^O$ is transformed into a scalar weight by $\mathbf{w}_G$ to indicate the probability of the person ($\mathbf{p}_i$) being involved in the event. The event-person interaction is trimmed at 0, acting as a ReLU nonlinearity. The zero trimming operation restricts the event-person interactions only of the people being not related to the event.

Estimating Importance Interaction from Both Graphs. Since we have two interaction graphs, the person-person interaction graph and the event-person interaction graph, the method for estimating the importance interaction $\hat{\mathcal{E}}_{ji}^p$ from

---

[4]There are two commonly used attention functions/mechanisms: the additive attention function [1] and the less expensive scale dot product function [12, 18]. While the two are similar in theoretical complexity, the additive operation slightly and consistently outperforms the scale dot product operation [3]. This outcome is also verified in our experiments, so we use the additive attention function for the person-person interaction modeling.

[5]Eq. 6 is different from Eq. 5 as the event-person interaction differs from the person-person interaction (asymmetric) presenting how a person is interacting with another people: the event-person interaction should be equal to the person-event interaction (symmetric) and is estimated to find whether a person is involved in the event.
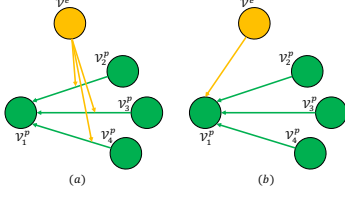
Figure 6. We introduce two methods to integrate the event-person interaction graph with the person-person interaction graph. First, we treat the event-person interaction as the prior importance acting as the regulator to adjust the weight of the person-person interactions (Figure (a)). Second, we treat the event-person interaction as an extra input link for each person (Figure (b)).

both graphs can significantly affect the importance relation computation and then impact the final results. In this work, we introduce two methods (Figure 6) to estimate the importance interaction from multiple graphs. Intuitively, we treat the event-person interaction as a prior importance and estimate the importance interaction $\hat{\mathcal{E}}_{ji}^p$ as:

$$\hat{\mathcal{E}}_{ji}^p = \mathcal{E}_{ji}^p \cdot \mathcal{E}_j^g. \qquad (7)$$

The advantage of this strategy is that the prior importance $\mathcal{E}_j^g$ acts as a regulator to adjust the effect of the person-person interaction $\mathcal{E}_{ji}^p$ on aggregating the relation features by enhancing the effect when the prior importance is large and reducing the impact in the opposite case.

An alternative strategy is to treat the event-person interaction as an additional graph to the person-person interaction graph. In other words, we define the importance interaction as the person-person interaction (i.e., $\hat{\mathcal{E}}_{ji}^p = \mathcal{E}_{ji}^p$) and the relation feature is aggregated as:

$$\mathbf{f}_i^R = \sum_{j=1}^N \mathcal{E}_{ji} \cdot (\mathbf{W}_{V_1}\mathbf{f}_j^O) + \mathcal{E}_i^g \cdot (\mathbf{W}_{V_2}\mathbf{f}_{global}), \qquad (8)$$

where $\mathcal{E}_{ji}$ is computed by Eq. (4). Here, the relation feature aggregates the feature from the others by a weighted sum of person features from the other people, linearly transformed by $\mathbf{W}_{V_1}$ and the global feature transformed by $\mathbf{W}_{V_2}$. In this way, the global information can be considered during encoding the importance features without affecting the effect of the person-person interaction.

The above two strategies are verified to be effective for combining both person-person interactions and event-person interactions, and they have comparable results.

Parameters of the Relation Module. The relation module Eq. (2) is summarized in Figure 2. It is easy to implement using basic operators, as illustrated in Figure 5. As the dimension of the output feature is the same as the input feature, we can stack more than one relation module (N_r relation modules) to refine the importance feature. In Eq. (2), since we have r relation submodules in one relation module, the parameters are $5 \times r$ projections: $\theta^R =$

$\{\mathbf{W}_Q^\ell \in \mathcal{R}^{d_f \times d_k}, \mathbf{W}_K^\ell \in \mathcal{R}^{d_f \times d_k}, \mathbf{W}_V^\ell \in \mathcal{R}^{d_f \times d_v}, \mathbf{w}_P^\ell \in \mathcal{R}^{d_f}, \mathbf{w}_G^\ell \in \mathcal{R}^{d_f}\}_{\ell=1}^r$, where $d_f = 1024$ is the dimension of the person feature and $d_k = d_v = \frac{d_f}{r}$ Due to the reduced dimension of each relation submodule, the total computational cost is similar to that of the single relation submodule with full dimensionality.

3.4. Classification Module for End-to-End Learning

After we obtain the importance feature for each person in an image, we utilize two fully connected layers (i.e., the classification module $f^S(\mathbf{f}_i^I|\theta^S)$) to transform the feature into two scalar values indicating the probability of the person belonging to the important people or non-important people classes. During training, the commonly used cross-entropy loss is employed to penalize the model, and the SGD is used to optimize the model for backward computation. During testing, the probability of the important people class is used as the importance point for each people. In each image, and the people with the highest importance point will be selected as the most important people.

4. Experiments

In this section, we conducted extensive experiments on two publicly available image-based important people detection datasets. We followed the standard evaluation protocol in the dataset [10]. The mean average precision (mAP) and some visual comparisons are reported. The CMC curve, other visual results and the classification accuracy of all people tested are reported and analyzed in the Supplementary Material.

4.1. Datasets

For evaluation on important people detection in still images, there are two publicly available datasets [10]: 1) The Multi-scene Important People Image Dataset (MS Dataset) and 2) the NCAA Basketball Image Dataset (NCAA Dataset).

1) The MS Dataset. The MS Dataset contains 2310 images from more than six types of scenes. This dataset includes three subsets: a training set (924 images), a validation set (232 images), and a testing set (1154 images). The detected face bounding box and importance labels are provided.

2) The NCAA Dataset. The NCAA Dataset is formed by extracting 9,736 frames of an event detection video dataset [13] covering 10 different types of events. The person bounding box and the importance annotations are provided as well.

4.2. Comparison with Other Methods

We first compared our method with existing important people detection models: 1) the VIP model [14], 2) Ramanathan's model [13] and 3) the PersonRank (PR)

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**

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

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

Table 3. The mAP (%) for Evaluating our Methods of Integrating Global Information on both Datasets.

MS Dataset		NCAA Dataset	
Method	mAP	Method	mAP
POINT$^{\mathcal{H}^p}$	91.2	POINT$^{\mathcal{H}^p}$	96.0
POINT$^{Eq. (8)}$	91.3	POINT$^{Eq. (8)}$	96.7
POINT$^{Eq. (3)+Eq. (7)}$	**92.0**	POINT$^{Eq. (3)+Eq. (7)}$	**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**

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

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

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

model [10] as well as all baselines (i.e., max-face, max-pedestrian, max-saliency, most-center, max-scale and SVR-person) provided in [10]. The experimental results are shown in Table 1[6]. From the table, it is clear that our POINT obtains state-of-the-art results. It is noteworthy that our POINT achieves a significant improvement of 23.2 % on the NCAA Dataset over the PersonRank method that achieved the best performance previously (i.e., 74.1 %). This verifies 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. This also indicates the effectiveness of incorporate the relation modeling with feature learning for important people detection. Interestingly, the improvement on the MS Dataset is significantly less than that on the NCAA Dataset (i.e., 3.4% vs 23.2%, respectively). The reason is that there are limited numbers of images (i.e., 2310 images in total), which limited the training of our deep model, even though the data augmentation of the training data (such as RandomCrop) has been used on the MS Dataset.

4.3. Evaluation of Our POINT

Evaluating Different Components of POINT. Since there is a lack of end-to-end trainable deep learning models for important people detection, we form a baseline that only comprises the feature representation module and the impor-

---

[6]On the MS Dataset, we did not compare Ramanathan's model [13] as it uses temporal information, which is not provided in the MS Dataset. All the results of other methods are from [10]

tance classification module. This approach predicts the importance of persons without considering their relations with others and the event-person relations. It is defined as:

$$\mathbf{s}_i^{Baseline} = f^O(\mathbf{p}_i|\theta^O) \circ f^S(\mathbf{f}_i^O|\theta^S). \quad (9)$$

It is formed to evaluate the effect of the relation module (i.e., our POINT) and different components of the feature (i.e., the interior feature, the location feature and the exterior/contextual feature). The results are reported in Table 2 where the Base^{Inter} indicates the baseline using only the interior feature and POINT$^{Inter + Loca + Exter}$ is our full model. The POINT, using the feature comprising all features, is described in Section 3.2.

From Table 2, it is noteworthy that our POINT consistently obtains better mAP values than the baseline using different types of features (e.g., 92.0% vs 89.2%, respectively, on the MS Dataset using three types of cues). This result indicates that embedding the relation module introduced in this paper can significantly aid in extracting more discriminant, higher level semantic information, which dramatically increases the performance. Additionally, we can see that both the baseline and POINT improve the mAP on

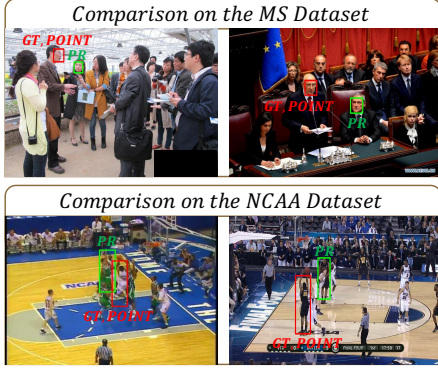*Comparison on the MS Dataset*

Comparison on the NCAA Dataset

Figure 7. Visual results of detecting important people and comparison with related work (i.e., PersonRank (PR)) on Both Datasets.

important people detection by using more cues compared to those using less information or a single type of information (e.g., the Base$^{\text{Indi+Cont+Loca}}$ has an improvement of 16.6% mAP over the Base$^{\text{Indi}}$, which obtains 72.6% mAP on the MS Dataset).

Integrating the Additional Global Information and Estimating the Importance Relation. In this work, we introduce two methods to integrate the event-person interaction graph with the person-person interaction graph. Table 3 presents the results of our POINT for detecting important people without global information (i.e., POINT$^{\mathcal{H}^p}$), our POINT using the global information in different ways (i.e., POINT$^{\text{Eq. (3)+Eq. (7)}}$ and POINT$^{\text{Eq. (8)}}$). It is clearly shown that both methods successfully integrate the global information into the importance feature and improve the performance. In general, the improvement when using the global information as a prior importance is higher than that of treating the event-interaction graph as an additional graph (e.g., 1.3% vs 0.7%, respectively, on the NCAA Dataset).

We also compared our method of estimating the importance relation with that of the attention weight [18] (i.e., the relation module in other vision tasks [4, 22]), and the results on both datasets are reported in Table 4. While the whole relation network is completely different from [18, 4, 22] due to different tasks, it is clear that our relation module is more effective than the relation model used in [18, 4, 22], as we have consistent improvement (e.g., 92.0% vs 90.0%, respectively, on the MS Dataset). This result verifies the efficacy of Eq. (4).

Visual Results and Comparisons. In this section, selected visual results and comparisons are reported in Figure 7 to further evaluate our POINT. As shown in Figure 7, it is clear that our POINT 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.

Effect of r and N_r on Important People Detection. The number of relation submodule r and the number of stacked relation module N_r can slightly affect our POINT. To evaluate the effect of both parameters, we report the results of our POINT using r ranging from 1 to 32 and keeping $N_r = 1$ in Table 5. Then, we select $r = 4$ as it yields the best result and set N_r ranging from 1 to 6. The evaluation results of the effect of N_r are reported in Table 6. The results shows that using the $r > 1$ relation submodule in a relation module enables our POINT to obtain better results because using multiple relation submodules allows our POINT to model various types of relations. In addition, we find that when we set $N_r > 1$, the POINT obtains slightly better results (e.g., setting $N_r = 2$ on the MS dataset and $N_r = 4$ on the NCAA dataset are the best) because the added relation modules can aid in refining the importance features.

Evaluation of the Attention Functions. Currently, there are two commonly used attention functions for modeling interaction between any pairs of entities, the additive and the scaled dot product attention functions. Similar to [3], we find the additive attention function works slightly but consistently better than the scaled dot product function from Table 7 (e.g., 97.3% vs 96.2%, respectively, on the NCAA Dataset).

Running time. We implement our model using PyTorch on a machine with CPU E5 2686 2.3 GHz, GTX 1080 Ti and 256 GB RAM. The running time of our POINT for processing an image is sensitive to the number of persons in the image. On average, POINT can process 10 frames per second (fps), which is significantly faster than the PersonRank (0.2 fps) and the VIP (0.06 fps). This result indicates that our POINT largely improves the speed of the important people detection model.

5. Conclusion

We have proposed a deep importance relation network to investigate deep learning for exploring and encoding the relation features and exploiting them for important people detection. More importantly, we have shown that POINT successfully integrate the relation modeling with feature learning to learn the feature for relation modeling. In addition, POINT can learn to encode and exploit the relation feature for important people detection. It was clearly shown that our proposed POINT could obtain state-of-the-art performance on two public datasets.

6. Acknowledgement

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *International Conference on Machine Learning*, 2014.

[2] Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. Understanding and predicting importance in images. In *Computer Vision and Pattern Recognition*, 2012.

[3] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. Massive exploration of neural machine translation architectures. In *Conference on Empirical Methods in Natural Language Processing*, 2017.

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

[5] Sung Ju Hwang and Kristen Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International journal of computer vision*, 100(2):134–153, 2012.

[6] Duy-Dinh Le, Shin'ichi Satoh, Michael E Houle, and Dat Phuoc Tat Nguyen. Finding important people in large news video databases using multimodal and clustering analysis. In *International Conference on Data Engineering*, 2007.

[7] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition*, 2012.

[8] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015.

[9] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2014.

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

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, 2016.

[12] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *Conference on Empirical Methods in Natural Language Processing*, 2015.

[13] 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.

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

[15] Merrielle Spain and Pietro Perona. Measuring and predicting object importance. *International Journal of Computer Vision*, 91(1):59–76, 2011.

[16] Christian Szegedy, Scott Reed, Dumitru Erhan, and Dragomir Anguelov. Scalable, high-quality object detection. *arXiv*, 2014.

[17] Yongyi Tang, Peizhen Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Latent embeddings for collective activity recognition. In *Advanced Video and Signal Based Surveillance*, 2017.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[19] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Computer Vision and Pattern Recognition*, 2018.

[20] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2015.

[21] Flood Sung Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Computer Vision and Pattern Recognition*, 2018.

[22] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. Relational deep reinforcement learning. *arXiv*, 2018.