Very Important Person Localization in Unconstrained Conditions: A New Benchmark

Xiao Wang[1], **Zheng Wang**[2,3†], **Toshihiko Yamasaki**[2,3], **Wenjun Zeng**[4]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology
[2]Research Institute for an Inclusive Society through Engineering (RIISE), The University of Tokyo
[3]Department of Information and Communication Engineering, The University of Tokyo
[4]Microsoft Research Asia

Abstract

This paper presents a new high-quality dataset for Very Important Person Localization (VIPLoc), named Unconstrained-7k. Generally, existing datasets are: 1) limited in scale; 2) built under simple and constrained conditions, where the number of disturbing non-VIPs is not large, the scene is relatively simple, and the face of VIP is always in frontal view and salient. To tackle these problems, the proposed Unconstrained-7k dataset is featured in two aspects. First, it contains over 7,000 annotated images, making it the largest VIPLoc dataset under unconstrained conditions to date. Second, our dataset is collected freely on the Internet, including multiple scenes, where images are in unconstrained conditions. VIPs in the new dataset are in different settings, *e.g.,* large view variation, varying sizes, occluded, and complex scenes. Meanwhile, each image has more persons (> 20), making the dataset more challenging.

As a minor contribution, motivated by the observation that VIPs are highly related to not only neighbors but also iconic objects, this paper proposes a Joint Social Relation and Individual Interaction Graph Neural Networks (JSRII-GNN) for VIPLoc. Experiments show that the JSRII-GNN yields competitive accuracy on NCAA (National Collegiate Athletic Association), MS (Multi-scene), and Unconstrained-7k datasets. https://github.com/xiaowang1516/VIPLoc.

1 Introduction

Very Important Person Localization (VIPLoc) aims to automatically locate core individuals that play important roles in given images. It has a guiding influence on the occurrence of current events, such as sports games, birthday parties, and speeches. Localizing important person potentially benefits person re-identification (Wang et al. 2020c; Yang et al. 2020; Xu et al. 2021), event recognition (Yao et al. 2020) and event detection (Bhardwaj, Yang, and Cudré-Mauroux 2020; Tong et al. 2020). In particular, analyzing the action of important person enables intelligent system to better understand what has happened and predict what will happen.

Our work is motivated by two aspects. First, most existing VIPLoc datasets, such as VIP (Mathialagan, Gallagher, and Batra 2015), NCAA (Li, Li, and Zheng 2018), and MS (Li, Hong, and Zheng 2019), are limited either in the dataset scale or data diversity. Specifically, the number of images is often confined in several hundreds, such as 200 in VIP dataset (Mathialagan, Gallagher, and Batra 2015). It is infeasible to test the robustness of algorithms under large-scale data. Moreover, in most cases, images are collected in fixed scenes, such as basketball games in NCAA dataset (Li, Li, and Zheng 2018), as shown in Figure 1(a). In constrained conditions, faces of VIPs are in the frontal view, big, and salient, such as VIPs in MS dataset (Li, Li, and Zheng 2018), as shown in Figure 1(b). But in reality, faces of VIPs might not be so salient, *e.g.,* with large view variation, varying small, or occluded (see Figure 1(c)). As a result, current methods may be biased toward ideal settings and their effectiveness may be impaired once the model is applied in reality. To address this problem, it is essential to introduce datasets that are closer to realistic settings.

Second, few methods addressed the VIPLoc task. As a representative, the inference-based methods attempt to infer the importance of person directly. (Ramanathan et al. 2016) employed an attention model to predict the importance from persons' action and appearance. (Mathialagan, Gallagher, and Batra 2015) used a regression model on spatial and saliency information to infer a relatively important person. These methods ignored some semantic information such as relationships among persons, which are essentially useful to evaluate the importance of persons. As a pioneering work, (Li, Hong, and Zheng 2019; Li, Li, and Zheng 2018) started to mine the relationship between persons. In reality, the VIP is not only related to other persons but also associated with some iconic objects in the current occasion. For example, a birthday star should have a strong interaction with the birthday cake, blowing out the candle; a football player may run with a ball, preparing to shoot; a speaker is giving a talk standing by the lectern. Here, the birthday star, the football player, and the speaker are the VIPs in each event image (please refer to Figure 3). The examples above tell us that individual interactions with objects provide useful information for inferring the VIPs, working as a complementary effect, if we can fuse the respective strengths of social relation and individual interaction factors.

Considering the above two issues, this paper makes two contributions. The main contribution is the collection of a new VIPLoc dataset, named Unconstrained-7k (Fig-

†Corresponding Author:wangz@hal.t.u-tokyo.ac.jp

(a) NCAA (b) MS (c) Unconstrained-7k

Figure 1: Typical samples in (a) NCAA, (b) MS, and (c) Unconstrained-7k datasets. NCAA dataset only focuses on the scene of the basketball game. Its VIP is often shooting. There is a significant difference between VIP and other players. The number of persons in MS dataset is small, and it is limited to the controlled condition that the face of VIP is always in frontal view and salient. The difference of Unconstrained-7k is not only in the crowd scene but also in the multi-view of the persons.

ure 1(c)). It contains 7,250 annotated images collected freely from the Internet. To our knowledge, Unconstrained-7k is the largest VIPLoc dataset under unconstrained conditions. It differs from existing datasets in several aspects: large scale (number of labelled images), more then ten scenes, diverse views, and densities of persons (varying size, occlusion). Moreover, each image has more 20 persons, making the dataset more challenging. Thus, the Unconstrained-7k provides a more realistic benchmark.

As a minor contribution, motivated by the findings that VIPs are highly related to not only neighbors but also iconic objects, a Joint Social Relation and Individual Interaction Graph Neural Networks (JSRII-GNN) is proposed. JSRII-GNN takes advantage of both social relation and individual interaction to infer the probability of key relation between persons and key interaction between person-object pairs. After detecting the persons and objects in images, the graph of person-person relations and person-object interactions are constructed. The dual GNNs are exploited to extract features of the graphs. Finally, two kinds of feature embeddings are concatenated to judge the importance of each person in the images.

In the following sections, we first introduce the related datasets and methods. Then, we present our new dataset Unconstrained-7k, and highlight its strength by a comparison with other datasets. Next, we uncover the correlations of person-person and person-objects in the image, and detail our proposed framework JSRII-GNN for VIPLoc. At last, we conduct experiments on public datasets and ours, to demonstrate the effectiveness of our method.

2 Related Works

In this section, we briefly survey the issue of VIPLoc, including related datasets and methods.

2.1 VIPLoc Related Datasets

Previous research work analyzed the concept of individual importance in group photographs and started the research of VIPloc. This research work is based on a VIP dataset (Mathialagan, Gallagher, and Batra 2015), which consists of only 200 images (not available yet). Recent research (Li, Li, and Zheng 2018) has released two related datasets, including the NCAA basketball dataset and MS dataset. (Hong, Li, and Zheng 2020) extended NCAA and MS datasets to a semi-supervised application, accompanied by two extended versions, *i.e.,* ENCAA, and EMS. The scene of the NCAA and ENCAA datasets are relatively simple, including only indoor basketball sports scenes. MS and EMS datasets contain images mainly from six types of scenes under constrained conditions, where VIPs are in frontal view and salient. These datasets are biased toward simple settings, and their effectiveness may be impaired once meeting reality. In contrast, we collected a new Unconstrained-7k dataset, which featured by 7,250 annotated images, ten scenes, unconstrained condition (diverse views, varying size, occlusion, and more than 20 persons per image).

2.2 VIPLoc Related Methods

Important object detection has been explored by prior works (Berg et al. 2012; Lee, Ghosh, and Grauman 2012; Lee and Grauman 2015; Li, Li, and Zheng 2018; Ramanathan et al. 2016; Mathialagan, Gallagher, and Batra 2015; Li, Hong, and Zheng 2019; Ghosh and Dhall 2018). Our research is more related to the studies of important person detection (Ramanathan et al. 2016; Li, Li, and Zheng 2018; Ghosh and Dhall 2018; Mathialagan, Gallagher, and Batra 2015). To facilitate the research of VIPLoc, (Ghosh and Dhall 2018) proposed a coarse-to-fine strategy for important person detection; (Li, Li, and Zheng 2018) built a hybrid person relationship and developed a PersonRank model to rank the individuals in terms of importance scores; (Li, Hong, and Zheng 2019) also aggregated relation features (Zhang et al. 2019; Liu et al. 2019) and the person's feature to form the importance feature. In contrast to these methods, we mainly focus on designing a dual relation method to leverage the interaction information, which not only contains person-person relations but also person-object relations to perform VIPLoc.

(a) body (b) face (c) head

Figure 2: Three types of detection results for annotations.

3 Unconstrained-7k Dataset

3.1 Existing Datasets

There are two typical datasets for the VIPLoc task. 1) The NCAA Basketball Image Dataset is formed by extracting frames covering basketball match events. 2) The MS dataset contains 2310 images from more than six types of scenes. This dataset includes training and testing subsets. The bounding boxes of detected faces and importance labels are provided.

3.2 Description

In this paper, a new VIPLoc dataset, Unconstrained-7k, is introduced. During the dataset collection, we retrieved 50,000 images from the Internet through key words queries, such as "speech", "demonstration", "interview", "sports", "military", "meeting", *etc*. We manually identified ten kinds of scenes. Most images are crowded scenes, so as to reflect the VIP, which is more in line with this task. In actual scenes, persons and camera shooters are more casual, resulting in an unstable imaging perspective. VIPs are imaged not always with frontal face. When shooting, there may be other views, such as viewpoint change, bowing the head, turning around, and back view. In order to better reflect the real scene, the pictures we collected cover all the above scenes.

Existing works mainly exploited the results of body or face detection to make annotations. The NCAA dataset uses the body, while the MS dataset uses the face. We have tried both annotation strategies. But it has not achieved very satisfactory results. Our dataset focuses on crowded scenarios and occluded scene. It is difficult to annotate persons with body annotation, which is often accompanied by other person's interference, as shown in Figure 2(a). In our dataset, viewpoint changes significantly, and persons' faces may not be captured. VIPs may not appear with frontal face in the image, as shown in Figure 2(b). In the multi-person scene, the head occlusion is not as serious as the body occlusion. Therefore, we use the head to annotate the VIPLoc dataset, as shown in Figure 2(c).

In our annotation process, we applied the head detector (Vora and Chilaka 2018) to obtain the position of person heads in the collected images. Then, the bounding boxes of detected heads are provided to the annotators. The annotators make their annotations by selecting the bounding boxes shown on the images. In order to reduce the bias, six annotators were asked to vote the VIPs. The person that gets the largest number of votes is then selected as the ground-truth VIP. It should be mentioned that not all images contain VIPs. We removed the images without annotation. Finally, we split the annotated data into train and test partitions according to the ratio of 1:1.

Datasets	NCAA	MS	Unconstrained-7k
Mean	0.027	0.014	0.010
Variance	0.012	0.013	0.025

Table 1: A comparison of the ratio of VIP size and image size.

Dataset	NCAA	MS	Unconstrained-7k
Max-Face	31.4	35.7	**23.8**
Max-Pedestrian	24.7	30.7	**21.8**
Max-Saliency	26.4	40.3	**22.6**
Most-Center	30.0	50.9	**26.8**
Max-Scale	31.8	73.9	**25.5**
SVR-Person	64.5	75.9	**46.7**

Table 2: Preliminary study on NCAA, MS, and Unconstrained-7k. mAP (%) values of six basic methods are reported.

3.3 Featured Properties

Specifically, the Unconstrained-7k dataset contains 7250 images from more than ten types of scenes. Compared with the existing datasets, our dataset has two advantages in terms of the average number of persons per image and the data diversity. The average number of persons per image in MS dataset is 8.59, while that in Unconstrained-7k dataset is 21.5, more than twice of that in MS dataset. We also make statistics of the ratio of VIP size and image size on NCAA, MS and Unconstrained-7k. The results in Table 1 show that the ratios in Unconstrained-7k have the smallest mean and the largest variance. It means that the VIPs in Unconstrained-7k have relatively small sizes and large variations. Furthermore, the Unconstrained-7k dataset is close to real life applications. Existing datasets were constructed for fixed events, such as basketball matches, or in constrained environments, where persons' faces are always in the frontal view and have large size. Figure 1 shows some typical samples.

3.4 Preliminary Study

We also performed a preliminary study to investigate the challenge of our dataset. We used six basic methods to conduct experiments on NCAA, MS, and Unconstrained-7k datasets, and the mAP values are evaluated. The hand-craft methods include 1) Max-Face, which flags a VIP to the person that gets the highest confidence score of face/head detection; 2) Max-Pedestrian, which flags a VIP to the person that gets the highest confidence score of face/head detection; 3) Max-Saliency, which flags a VIP to the person that has the largest appearance discrepancies with other persons in the image; 4) Most-Center, which flags a VIP to the person that is closest to the centre of the image; 5) Max-Scale, which flags a VIP to the person whose face/body size is the largest; 6) SVR-Person (Mathialagan, Gallagher, and Batra

(a) Social Relation

(b) Individual Interaction

Figure 3: (a) Very important persons are often associated with most other persons. We define this factor for VIPLoc as social relationship. (b) Persons have different interactions with the key object. The important person contains a more close interaction in the current event, such as birthday star and cake, speakers and lectern, and player and football in the examples. We define this factor for VIPLoc as individual interaction.

2015), which predicts VIP based on the concatenation of the four types of features (spatial, action, appearance, and attention feature). Table 2 shows the experimental results. It is obvious that the Unconstrained-7k dataset is the most challenging one, since all methods get the worst results on this dataset.

4 Proposed Method

4.1 Motivation

VIPLoc is a more challenging task than person detection (Wang et al. 2017, 2019, 2020b) as it requires extracting higher semantic information than other detection tasks. We find two relationships that help locate the VIP: the person-person relation and the person-object interaction. For the former, the relationship response between the VIPs and others will be stronger than that between others, as shown in Figure 3(a). For the latter, the VIPs have a certain relationship with the object in the scene, as shown in Figure 3(b). In a birthday party event, a birthday star should have a strong interaction with the birthday cake, blowing out the candle. In ball games, the important person is the star who contributes the most to the game at the time, and the key object is the ball. The relationship between the star and the ball is the key interaction in the current scene.

We aim to design JSRII-GNN, which consists of a social relation and an individual interaction GNN. It learns to build the relationships and combines the relation modeling with feature learning for VIPLoc. The overall framework (see Figure 4) mainly includes person/object feature extraction, social relation and individual interaction graphs building, and GNN features integration.

4.2 Person/Objects Feature Extraction

The detector in this section is built upon the state-of-the-art object detection framework, *i.e.,* Yolov4 (Bochkovskiy, Wang, and Liao 2020), pre-trained on COCO dataset (Caesar, Uijlings, and Ferrari 2018) for objects and pre-trained on head data (Shao et al. 2018) for person heads, so that reasonable regions for heads and objects can be produced in an automatic manner for VIPLoc.

We use the ResNet-50 to extract features from each head patch because it has demonstrated its superiority in terms of instance feature. For the head in an image, we feed the patches into two separated Resnet-50s, transforming the head patch into a $7 \times 7 \times 2048$ feature. While the coordinate is a four dimensional vector, we produce a heatmap, which is of size 224×224, where one or several cells corresponding to the person's coordinate are assigned 1 and the others as zero. We apply convolutional kernels to this heatmap to produce a $7 \times 7 \times 256$ feature. Then, we concatenate the head and heatmap features, and employ two convolutional layers with one fully-connected (fc) layer to transform this concatenated feature into a 1024 dimensional vector, called the person feature.

Similar to heads, detected objects are fed to the network to generate a 1024 dimensional vector. We call these the object features.

4.3 Social Relation and Individual Interaction Graphs

Social Relation Graph. To capture the relationships between different persons in each image, we build the social relation graph by estimating the distances of persons. For the adjacent matrix of the graph, $A_S \in R^{N_p \times N_p}$, we directly set $A_S(p_i, p_j) = 1$, p_i and p_j are two persons in one image, where $sim(p_i, p_j)$ is the cosine similarity of p_i and p_j.

$$A_S(p_i, p_j) = \begin{cases} 1 & sim(p_i, p_j) \leq \tau \\ 0 & otherwise \end{cases}, \quad (1)$$

where τ is a threshold[1] to determine the key social relations.

Individual Interaction Graph. The contextual objects in the scene are vital information for important person localization. The relationship between persons and objects is a many-to-many association problem, which makes it difficult to capture the key interactions between persons and contextual objects through still image. Therefore, different from social relation graph, the individual interaction graph is designed to model the co-existence of persons and contextual objects. The adjacent matrix of the individual interaction graph, $A_I \in R^{N_p \times N_o}$, represents the interaction between persons and the objects that exist in the image. Therefore, we set $A_I(p_k, o_l) = 1$, if p_k and o_l are the nodes of the key individual interaction, following the same constraints as in Equation (1).

Graph Neural Network. The dual graphs are built to represent the persons of the image, *i.e.,* the interactions between different persons, and the co-existence of persons and objects (Liu et al. 2020). Traditional Neural Networks (Gallicchio and Micheli 2020) usually apply 2-D filters on images to abstract visual features from low-level space to high-level space. In contrast, GNN (Huang, Liu, and Lin 2018) performs relational reasoning by performing message propagation from nodes to its neighbors in the graphs (Chen et al.

[1] τ is set to 0.3 in this paper.

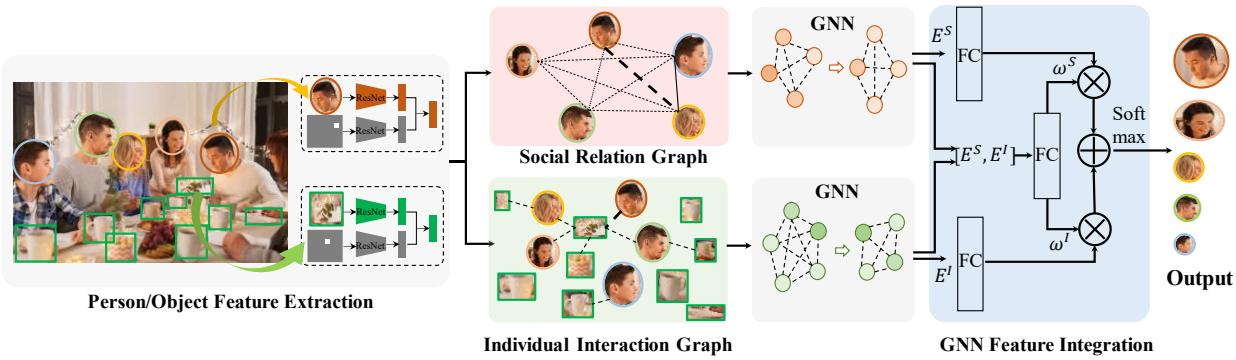


Figure 4: An illustration of our framework. We exploit the detection and feature representation module to extract person features and object features from the whole image. Using these features, we construct two relationship graphs (social relation and individual interaction). The JSRII-GNN learns relation features and then features are concatenated into a relation feature vector. The relation feature and the person feature are added together, resulting in the importance feature, which is employed to infer the importance point of person.

2020a). Therefore, we can apply GNNs on the dual graphs to infer the VIPs.

Given a graph with N nodes in which each node has a d-dimensional feature vector, the operation of one graph neural layer can be formulated as:

$$X^{(k+1)} = \delta\left(\frac{1}{\sqrt{D}} M \frac{1}{\sqrt{D}} X^{(k)} W^{(k)}\right). \tag{2}$$

where $M \in R^{N\times N}$ is the adjacent matrix of the graph, $D \in R^{N\times N}$ is the degree matrix of M. $X^{(k)} \in R^{N\times d}$ is the output of the $(l-1)$-th layer, $W^{(k)} \in R^{d\times d}$ is the learned parameters, and $\delta(\cdot)$ is a non-linear activation function. In our framework, the adjacent matrices of the dual graphs are A_S and A_I. $X^{(0)} = [f(x_1), f(x_2), \ldots, f(x_N)]^T$ is the initial feature matrix, where $f(x_i)$ is the column vector extracted from the nodes x_i, including persons or objects. The final outputs of the GNNs are updated features of nodes $X^{(K)}$, in the graphs, which can be aggregated into person feature vector for important person prediction. At last, the scores of social relation GNN and individual interaction GNN are combined by weighted fusion for the final prediction.

4.4 GNN Feature Integration

Although both social relation and individual interaction models are useful for VIPLoc, direct concatenation does not produce satisfactory results under complex conditions. In multi-person scenarios, social relation graph network can locate the VIP, while it is difficult to accurately locate the VIP in the few-person scenarios. The individual interaction graph network can determine VIP in scenarios with obvious iconic objects, such as cake in a birthday party, ball in the ball games, microphone in the speech, while it is not feasible to determine the VIP in the outdoor scenarios with no obvious objects. These interference factors would lead to a degradation in performance if used directly for this tasks. The integrated mechanism between the social relation and individual interaction models is particularly important. The integrated architecture needs to encourage models that contribute to determine VIP and suppress models that are obstacles to determining VIP.

Among the diverse facets of the human multisensory system (Chen et al. 2020b), selective attention (Corbetta and Shulman 2002) allows the human brain to focus on salient information, even for complex sensory inputs. The attention mechanism (Chen et al. 2020c) dynamically brings salient features (Wang et al. 2020a) to the forefront as needed without transferring the entire information into the abstraction. Inspired by the human multisensory capability, we consider the contribution of each graph network in the integration mechanism. Motivated by the above observation, we leverage the integration network to consider pairwise information when predicting graph weights. The proposed integration network mainly consists of three fully connected layers and one softmax layer.

In the training stage, for image I, $\{(E_i, y_i) \in I | i = 1, 2, \cdots, N\}$ represents the information of persons, where E_i denotes embedding information from the social relation graph or the individual interaction graph and the label y_i is 1 if the i-th person is VIP and 0 otherwise.

The predicted key confidence of the person can be represented as follows:

$$y_i^* = \sum_{r\in\{S,I\}} \omega^r f(E_i^r). \tag{3}$$

The loss function for VIPLoc is the cross-entropy loss

$$\mathcal{L}_{VIPLoc} = -\left(y_i \cdot \log y_i^* + (1-y_i)\cdot \log\left(1-y_i^*\right)\right). \tag{4}$$

For each training step, we use 64 positive and negative pairs, representing a total of 128 pairs. The learned weights control the contributions of both models in the integration of the social relation graph and individual interaction graph. In our framework, the integration networks spontaneously learn to implicitly assess the quality of the given multi-modality data. During training, the commonly used cross-entropy loss is employed to penalize the model, and the SGD (stochastic gradient descent) is used to optimize the model

Datasets	VIP	PR	POINT	JSRII-GNN
NCAA	53.2	74.1	97.3	**97.6**
MS	76.1	88.6	92.0	**92.5**
Unconstrained-7k	42.6	59.5	68.9	**73.8**

Table 3: Comparison in terms of the mAP (%) with existing methods on NCAA, MS, and unconstrained-7k datasets.

for backward computation. During testing, the probability of the important person class is used as the importance point for each person. In each image, the person with the highest importance score will be selected as the VIP.

5 Experiments

We conduct experiments on two publicly available datasets, NCAA, MS, and our Unconstrained-7k dataset. We follow the standard evaluation protocol in the NCAA and MS datasets. To quantify the performance of different methods on VIPLoc, the mean Average Precision (mAP), which is widely used in object detection, is used to assess the correctness of detected VIP and reported in this paper. In addition, the cumulative matching characteristics (CMC) curve is plotted to show the results of top k-rank VIP.

5.1 Comparison with State-of-the-Arts

We first compare our method with existing VIPLoc models: 1) the VIP model (Mathialagan, Gallagher, and Batra 2015), which learns importance of persons in images and investigates the correlation between importance and visual saliency. 2) PersonRank (PR) (Li, Li, and Zheng 2018), which is a modified pagerank algorithm to rank the persons in image, and 3) the imPOrtance relatIon NeTwork (POINT) (Li, Hong, and Zheng 2019), which uses the deep learning for exploring and encoding the relation features and exploiting persons for VIPLoc.

Performance on NCAA Dataset. Table 3 shows that POINT achieves 97.3% mAP, and our method achieves 97.6%. In addition, we report the CMC curve in Figure 5(a). Our approach obtaines 72.4% rank-1 matching rate.

Performance on MS Dataset. Table 3 shows that POINT achieves 92.0% mAP, and our method achieves 92.5%. In addition, we report the CMC curve in Figure 5(b). Our approach obtaines 88.4% rank-1 matching rate.

Performance on Unconstrained-7k Dataset. Table 3 shows that the performances on the NCAA and MS datasets are higher than that on the Unconstrained-7k dataset, which verifies that our dataset is challenging. Our method achieves the top performance of 73.8%. In addition, we report the CMC curve in Figure 5(c). Our approach obtaines 46.3% rank-1 matching rate. Although the new dataset is more challenging, this result shows that JSRII-GNN is much better than previous methods. It verifies the effectiveness of our method.

Results on the three datasets show that our proposed method is capable of better inferring the importance of persons and localizing the most important one in an image.

Social Relation	Individual Interaction	mAP
✓		66.3
	✓	60.1
✓	✓	73.8

Table 4: Ablation study. The mAP (%) results on the Unconstrained-7k Dataset are reported.

Methods	mAP
Direct Concatenation	68.2
Adaptive Fusion	73.8

Table 5: The mAP (%) for Evaluating Different Integrated Features of our JSRII-GNN on Unconstrained-7k Dataset.

5.2 Investigation on JSRII-GNN

Ablation Study. Table 4 shows the ablation study results on the Unconstrained-7k dataset. JSRII-GNN w/ social relation gets 66.3% mAP value, w/ Individual Interaction obtains 60.1% mAP value, and the mAP result reaches 73.8% w/ both. It confirms the effectiveness of each relationship part, and their complementary effects.

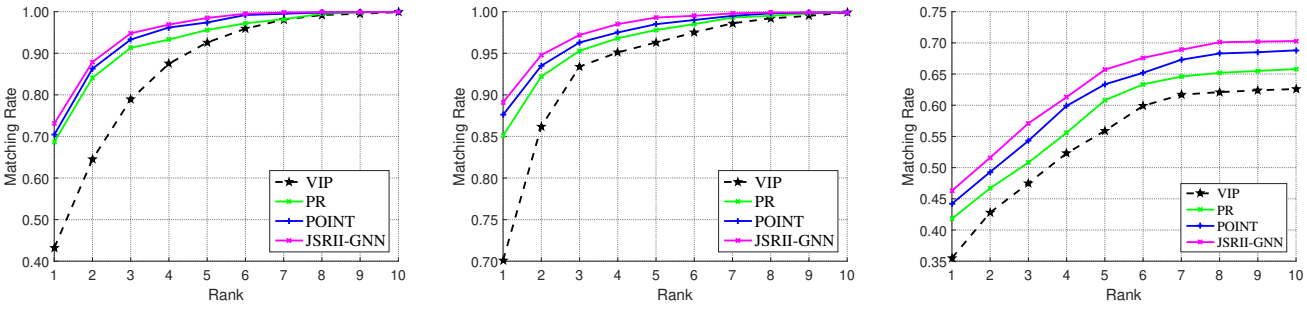
Investigation on the feature integration. There are two strategies to connect the person features of social relation graph and individual interaction graph: direct concatenation and adaptive fusion. The direct combination fuses the features of the two parts using the same weight. The adaptive fusion uses an attention network to learn which part contributes more to the selection of VIP and give it a greater weight. The comparison between the two strategies is shown in Table 5. The mAP of the direct concatenation strategy is 68.2%, while the mAP of the adaptive fusion is 73.8%. The adaptive fusion shows better performance, which verifies the effectiveness of the integrated features in our scheme.

Running time evaluation. We implement JSRII-GNN using PyTorch on a machine with CPU i7, GeForce GTX Titan X and 256 GB RAM. On average, JSRII-GNN can process 12 frames per second (fps), which is significantly faster than PersonRank (0.2 fps), and VIP (0.06 fps), and similar to POINT (11 fps). This result indicates that the efficiency of JSRII-GNN is considerable.

5.3 Visualization and Analysis

Figure 6 shows the effect of the individual interaction on the Unconstrained-7k Dataset. It shows that individual interaction plays a significant role in locating the VIP in the image. The social relation selects the VIP by analyzing each person on the screen to choose a more prominent one using persons' appearance attributes. However, note that this kind of interaction may not play a useful role when there are no critical objects in the picture.

We also visualize some better results and failure cases of our method in Figure 7. From the failure cases, we consider that the environment illumination and the direction of the face can be taken as the factor to improve the VIPLoc performance. (1) Low-illumination conditions cause a series of



(a) NCAA (b) MS (c) Unconstrained-7k

Figure 5: The CMC curves of different methods evaluated on NCAA, MS, and Unconstrained-7k datasets, respectively.

Figure 6: The visualization results of our JSRII-GNN on the Unconstrained-7k dataset. In each image, the yellow bounding box indicates the result that is obtained by JSRII-GNN w/o Individual Interaction. The green bounding box indicates the result that is obtained by JSRII-GNN w/ Individual Interaction (it is also the GT). Due to the assistance of significant objects, JSRII-GNN w/ Individual Interaction shows a better performance.

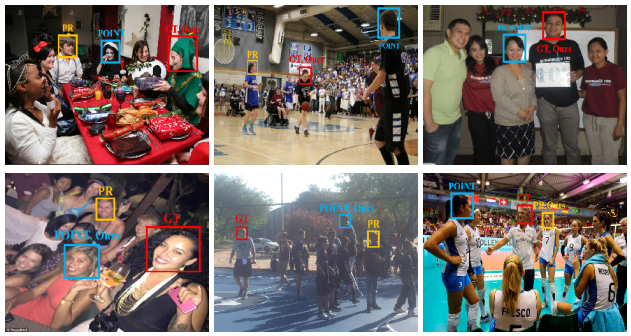


Figure 7: The visualization results of different methods on the Unconstrained-7k Dataset. Above each bounding box, we indicate the methods (PR, POINT, Ours) and ground truth (Ours). The top row shows the samples that our method gets the GTs. The botton row shows the failure cases.

visibility degradation and even sometimes destroy the color or content of the VIPs. (2) VIP with positive regular face is easier to recognize, others are more difficult to recognize. The main reason is that the positive face information is rich and the discrimination information is more. In other conditions, the head discrimination information is less, we can use landmarks point correction or more detailed semantic information to assist the VIPLoc.

6 Conclusions

This paper introduces a large-scale dataset for the VIPLoc task, Unconstrained-7k. Our new dataset is expected to benefit the research community. A JSRII-GNN is proposed in an attempt to bridge the person-person relation and the person-object relation. We have shown that JSRII-GNN successfully integrates the relation modeling with feature learning to learn the feature for social relation and individual interaction. In the future, we will investigate the problem of multiple VIPs with importances.

Clarity of privacy. We should clarify that there is no privacy issue in this dataset. This dataset was collected from news images on the Internet, and can be only used for academic research, not for commercial purposes.

Acknowledgments

This work is supported by National Nature Science Foundation of China (No. 61801335, U1803262).

References

Berg, A. C.; Berg, T. L.; III, H. D.; Dodge, J.; Goyal, A.; Han, X.; Mensch, A. C.; Mitchell, M.; Sood, A.; Stratos, K.; and Yamaguchi, K. 2012. Understanding and predicting importance in images. In *CVPR*.

Bhardwaj, A.; Yang, J.; and Cudré-Mauroux, P. 2020. A Human-AI Loop Approach for Joint Keyword Discovery and Expectation Estimation in Micropost Event Detection. In *AAAI*.

Bochkovskiy, A.; Wang, C.; and Liao, H. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR* abs/2004.10934.

Caesar, H.; Uijlings, J. R. R.; and Ferrari, V. 2018. COCO-Stuff: Thing and Stuff Classes in Context. In *CVPR*.

Chen, D.; Lin, Y.; Li, W.; Li, P.; Zhou, J.; and Sun, X. 2020a. Measuring and Relieving the Over-Smoothing Problem for Graph Neural Networks from the Topological View. In *AAAI*.

Chen, L.; Lv, B.; Wang, C.; Zhu, S.; Tan, B.; and Yu, K. 2020b. Schema-Guided Multi-Domain Dialogue State Tracking with Graph Attention Neural Networks. In *AAAI*.

Chen, S.; Andrejczuk, E.; Cao, Z.; and Zhang, J. 2020c. AATEAM: Achieving the Ad Hoc Teamwork by Employing the Attention Mechanism. In *AAAI*.

Corbetta, M.; and Shulman, G. L. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* 3(3): 201–210.

Gallicchio, C.; and Micheli, A. 2020. Fast and Deep Graph Neural Networks. In *AAAI*.

Ghosh, S.; and Dhall, A. 2018. Role of Group Level Affect to Find the Most Influential Person in Images. In *ECCV*.

Hong, F.; Li, W.; and Zheng, W. 2020. Learning to Detect Important People in Unlabelled Images for Semi-Supervised Important People Detection. In *CVPR*.

Huang, Q.; Liu, W.; and Lin, D. 2018. Person Search in Videos with One Portrait Through Visual and Temporal Links. In *ECCV*.

Lee, Y. J.; Ghosh, J.; and Grauman, K. 2012. Discovering important people and objects for egocentric video summarization. In *CVPR*.

Lee, Y. J.; and Grauman, K. 2015. Predicting Important Objects for Egocentric Video Summarization. *Int. J. Comput. Vis.* 114(1): 38–55.

Li, W.; Hong, F.; and Zheng, W. 2019. Learning to Learn Relation for Important People Detection in Still Images. In *CVPR*.

Li, W.-H.; Li, B.; and Zheng, W.-S. 2018. Personrank: Detecting important people in images. In *FG*.

Liu, X.; Liu, W.; Zhang, M.; Chen, J.; Gao, L.; Yan, C.; and Mei, T. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In *CVPR*.

Liu, X.; Liu, W.; Zheng, J.; Yan, C.; and Mei, T. 2020. Beyond the Parts: Learning Multi-view Cross-part Correlation for Vehicle Re-identification. In *ACM MM*.

Mathialagan, C. S.; Gallagher, A. C.; and Batra, D. 2015. VIP: Finding important people in images. In *CVPR*.

Ramanathan, V.; Huang, J.; Abu-El-Haija, S.; Gorban, A. N.; Murphy, K.; and Fei-Fei, L. 2016. Detecting Events and Key Actors in Multi-person Videos. In *CVPR*.

Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; and Sun, J. 2018. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *CoRR* abs/1805.00123.

Tong, M.; Wang, S.; Cao, Y.; Xu, B.; Li, J.; Hou, L.; and Chua, T. 2020. Image Enhanced Event Detection in News Articles. In *AAAI*.

Vora, A.; and Chilaka, V. 2018. FCHD: Fast and accurate head detection in crowded scenes. *arXiv preprint arXiv:1809.08766* .

Wang, Q.; Liu, X.; Liu, W.; Liu, A.; Liu, W.; and Mei, T. 2020a. MetaSearch: Incremental Product Search via Deep Meta-Learning. *IEEE Trans. Image Process.* 29: 7549–7564.

Wang, X.; Chen, J.; Liang, C.; Chen, C.; Wang, Z.; and Hu, R. 2017. Low-resolution pedestrian detection via a novel resolution-score discriminative surface. In *ICME*.

Wang, X.; Chen, J.; Wang, Z.; Liu, W.; Satoh, S.; Liang, C.; and Lin, C.-W. 2020b. When Pedestrian Detection Meets Nighttime Surveillance: A New Benchmark. In *IJCAI*.

Wang, X.; Liang, C.; Chen, C.; Chen, J.; Wang, Z.; Han, Z.; and Xiao, C. 2019. S3d: Scalable pedestrian detection via score scale surface discrimination. *IEEE Trans circ. Syst. video Techn.* .

Wang, Z.; Wang, Z.; Zheng, Y.; Wu, Y.; Zeng, W.; and Satoh, S. 2020c. Beyond intra-modality: A survey of heterogeneous person re-identification. In *IJCAI*.

Xu, X.; Liu, L.; Xiaolong, Z.; Weili, G.; and Ruimin, H. 2021. Rethinking data collection for person re-identification: active redundancy reduction. *Pattern Recognition* .

Yang, F.; Wang, Z.; Xiao, J.; and Satoh, S. 2020. Mining on Heterogeneous Manifolds for Zero-Shot Cross-Modal Image Retrieval. In *AAAI*.

Yao, W.; Zhang, C.; Saravanan, S.; Huang, R.; and Mostafavi, A. 2020. Weakly-Supervised Fine-Grained Event Recognition on Social Media Texts for Disaster Management. In *AAAI*.

Zhang, M.; Liu, X.; Liu, W.; Zhou, A.; Ma, H.; and Mei, T. 2019. Multi-Granularity Reasoning for Social Relation Recognition From Images. In *ICME*.