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

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

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

Abstract

Important people detection is to automatically detect the individuals who play the most important roles in a social event image, which requires the designed model to understand a high-level pattern. However, existing methods rely heavily on supervised learning using large quantities of annotated image samples, which are more costly to collect for important people detection than for individual entity recognition (e.g., object recognition). To overcome this problem, we propose learning important people detection on partially annotated images. Our approach iteratively learns to assign pseudo-labels to individuals in un-annotated images and learns to update the important people detection model based on data with both labels and pseudo-labels. To alleviate the pseudo-labelling imbalance problem, we introduce a ranking strategy for pseudo-label estimation, and also introduce two weighting strategies: one for weighting the confidence that individuals are important people to strengthen the learning on important people and the other for neglecting noisy unlabelled images (i.e., images without any important people). We have collected two large-scale datasets for evaluation. The extensive experimental results clearly confirm the efficacy of our method attained by leveraging unlabelled images for improving the performance of important people detection.

1. Introduction

The objective of important people detection is to automatically recognize the most important people who play the most important role in a social event image. Performing this task is natural and easy for a human. This topic has attracted increasing attention, as it has a wide range of realistic applications including event detection [21], activity/event recognition [24, 21], image captioning [23], an so on.

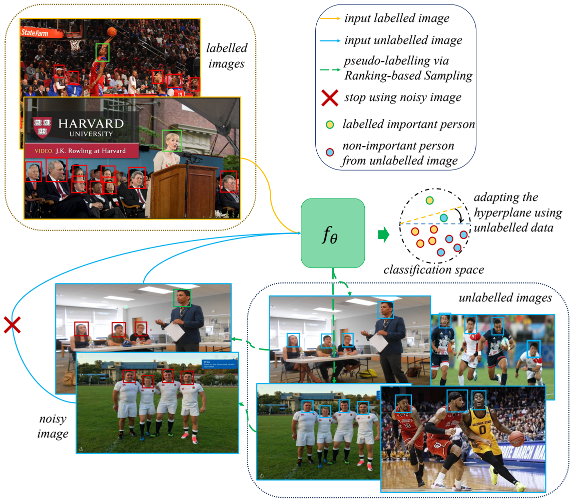

labelled images
input labelled image
input unlabelled image
pseudo-labelling via Ranking-based Sampling
stop using noisy image
labelled important person
non-important person from unlabelled image
f_θ
adapting the hyperplane using unlabelled data
classification space
unlabelled images
noisy image

Figure 1. Collecting large quantities of labelled data for important people detection is difficult and costly. Additionally, since there are always important people in social event images, we design a semi-supervised method that learns to automatically select available unlabelled images as well as prevent the noisy unlabelled images and detecting important people in unlabeled images to adapt the hyperplane of importance classification initialized by model trained with only labelled data.

Developing a model to detect important people in images remains challenging, as it requires the model to understand a higher-level pattern (*e.g.*, relation among people in an image) in each image compared with the information needed in other vision tasks (*e.g.*, object-level information in classification or object detection). Existing methods of important people detection require massive quantities of labelled data, which are difficult and very costly to collect for this task, as it requires humans to vote on the important people [18, 6]. Since there always are important people in social event im-

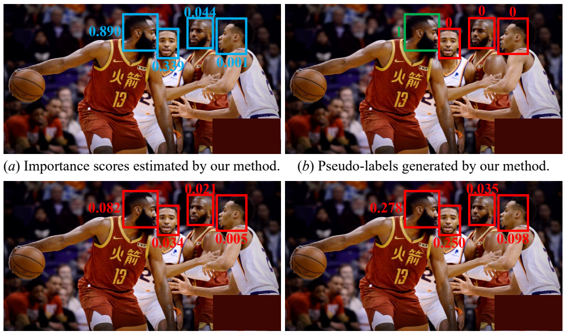
[*] Equal contribution.
[†] Corresponding author

ages, it is natural to ask if an important people detection model can be built to learn from partially annotated data, *i.e.*, a limited quantity of labelled data with a large number of unlabelled images. The question then arises regarding the design of a model that can learn from partially annotated data for important people detection if we augment the limited labelled training data with unlabelled images.

However, learning to detect important people in unlabelled images has its own challenging characteristics. First, it is not an individual entity (e.g., object) recognition task but is rather a certain classification problem [7], relying on the relation between people in an image. Second, as the statistics of two important people detection datasets shown in Figure 3, most images contain more than two people, resulting in a data imbalance problem that the number of important people is always much smaller than that of non-important people; this would yield a pseudo-labelling imbalance problem when pseudo-labels are assigned to unlabelled images, which will hamper the performance of semi-supervised learning because it is highly probable that all individuals will be regarded as "non-important" during pseudo-labelling (Figure 2(c)(d)). Third, not all unlabelled images contain important people; images without such people represent noisy unlabelled samples during learning.

To tackle the aforementioned challenges of semi-supervised important people detection, we develop an iterative learning procedure (Figure 1) that iteratively trains an important people detection model on data with labels or pseudo-labels and subsequently generates pseudo-labels again of all individuals in unlabelled images. In particular, we introduce a ranking-based sampling strategy for overcoming the imbalance problem in pseudo-label learning, where we rank all individuals in each unlabelled image in terms of the score of the important class and consider the individuals with relatively high score (*i.e.*, higher than a threshold) as important people (*i.e.*, with the pseudo-label of "important") while regarding the rest as non-important individuals (*i.e.*, with the pseudo-label of "non-important"). By using the proposed ranking-based sampling, we avoid the problem of classifying all individuals in an unlabelled image as "non-important" and thus the pseudo-labels of unlabelled image are more reliable (Figure 2(b)).

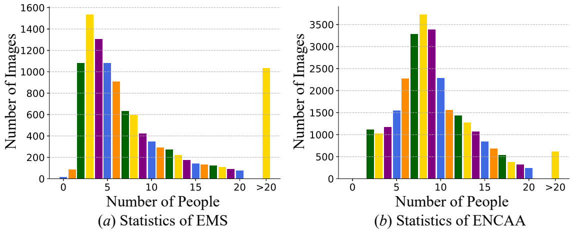
To further alleviate the pseudo-labelling imbalance problem of "non-important" data dominating the learning loss, we introduce an importance score weight to weight the confidence that individuals are important people in each unlabelled image while updating the important people detection model. Finally, to address the problem caused by noisy unlabelled images (without any important people in the images), we introduce an effectiveness weight, a continuous scalar ranging from 0 to 1 which indicates the confidence about an unlabelled image containing important people (*i.e.*, 0 for no important people in the image, as opposed to 1) to filter out those images (Figure 1). Here, two proposed weights are estimated in a feed-forward manner at each iteration and do not require any supervisions.



(*a*) Importance scores estimated by our method. (*b*) Pseudo-labels generated by our method.

(*c*) Pseudo-labels predicted by Mean Teacher. (*d*) Pseudo-labels estimated by Label Propagation.

Figure 2. Examples of our method's results and pseudo-labels estimated by different methods during training. Blue face boxes together with numbers in Figure (a) show the importance scores generated by our method. In Figure (b), (c), (d), pseudo-labels generated by our method and related approaches are shown in terms of "important" category's probabilistic numbers and face boxes in different colors. Here, individuals marked with red face boxes are assigned with "non-important' pseudo-labels and individuals marked with green face boxes are treated as "important" people.



(*a*) Statistics of EMS (*b*) Statistics of ENCAA

Figure 3. Statistics of Social Event Images in EMS and ENCAA datasets. Y-axis is the number of picture containing respective number of people (x-axis).

While there are no studies on learning important people detection from partially labelled data, we contribute two large datasets called Extended-MS (EMS) and Extended-NCAA (ENCAA) for evaluation of semi-supervised important people detection by augmenting existing datasets (*i.e.*, the MS and NCAA datasets [18]) with a large number of unlabelled images collected from the internet. Extensive experiments verify the efficacy of our proposed method attained by enabling the labelled and unlabelled data to interact with each other and leveraging the information of unlabelled images to assist in training of the entire important people detection model. We have conducted an ablation study to investigate the effect of each component of our method (*i.e.*, ranking-based sampling, importance score weighting and effectiveness weighting) on semi-supervised learning-based important people detection. Additionally, the results of our method incorporating with existing semi-supervised learning approaches (*e.g.*, mean-teacher (MT) [25] and label-propagation (LP) [9]) demonstrate that our proposed method is generic and stable for semi-supervised learning-based important people detection.

2. Related Work

2.1. Important people/object detection

Important people/object detection has been explored by prior work [1, 11, 13, 14, 18, 21, 23, 17, 6], but our research is more related to the studies of important people detection [21, 17, 6, 18, 23]. To facilitate the research of important people detection, the work [18, 6] has collected three small datasets, but it also indicates that it is difficult and costly to annotate a massive quantity of data for this task. These works mainly focused on developing fully supervised methods. In particular, Ghosh et al. [6] propose a coarse-to-fine strategy for important people detection; Li et al. [18] build a hybrid graph modelling the interaction among people in the image and develop a graph model called PersonRank to rank the individuals in terms of importance scores from the hybrid graph; In [17], Li et al. proposed an end-to-end network called the POINT that can automatically learn the relations among individuals to encourage the network to formulate a more effective feature for important people detection.

In contrast to the methods mentioned above, we mainly focus on designing a semi-supervised method to leverage information of massive unlabelled samples to assist in training a model on limited labelled data to perform important people detection.

2.2. Learning from partially labelled data

Learning from partially annotated data has recently become an important part of research in computer vision, as it enables the machine learning model (deep model) to learn from a large quantity of data without costly labelling. Recent work [25, 12, 5, 4, 9, 2, 8, 19, 10, 26] on semi-supervised learning mainly follows the well-known iterative bootstrapping method introduced in [27] (*i.e.*, the classifier trained on a current set of labelled samples is used to generate labels for unlabelled data in each iteration). Among these studies, Grandvalet et al. [8] proposed adding a loss term to minimize the entropy of the generated labels of unlabelled data based on the cluster assumption [3]. According to the latter, the work in [12] proposed a method called "Pseudo Label" that generates a label with the maximum probability for every unlabelled sample and uses it as a true label. Another well-known assumption [3] is about smoothness, whereby researchers base their methods on the consistency regularization strategy to enable the model to be invariant to the added noise. Miyato et al. [19] introduce a consistency loss based on the predictions for an unlabelled sample with and without a learned noise to encourage the model to be invariant to the learned noise. In [10], the authors propose the Π model to regularize the consistency with the models of the previous iterations by using temporal ensembles, while Tarvainen et al. [25] introduce a teacher network that represents the exponential average over each iteration's model (*i.e.*, student model) to tackle the limitation of using temporal ensembles noted in [10]. In contrast, Iscen et al. [9] proposed a method to regularize the consistency between the prediction of the unlabelled sample and the guessed label by using the label propagation technique. Following the consistency regularization strategy, recent methods MixMatch [2] and UDA [26] embed the idea of data augmentation techniques in consistency regularization, where they regularize the model to be consistent over two augmentations of the same unlabelled images. In addition to these, Li et al. [16] design a meta-learning framework that learns to impute unlabelled data such that the performance on the validation data of the model trained on these imputed data can be improved.

Unlike the above methods mainly proposed for standard image classification, in this work, we mainly focus on developing a semi-supervised approach for important people detection, where those methods are unsuitable. In particular, the importance of a person in an image is related to that of other people in the same image. In contrast, current semi-supervised approaches treat all unlabelled samples in an unlabelled image as independent samples and ignore the relations among them. In this paper, we design a method that can automatically exploit the pattern in unlabelled images and use a limited quantity of labelled data to assist in the overall training of an important people detection model.

3. Methodology

Developing a deep learning approach for important people detection requires a large quantity of labelled training data, which is difficult and costly to collect. To solve this problem, we aim to leverage the information from unlabelled data to assist in training a model for important people detection on partially annotated data. An illustration of our method is shown in Figure 4 and is detailed in the following.

3.1. Semi-supervised Pipeline

Consider a labelled important people image dataset that contains $|\mathcal{T}|$ labelled images $\mathcal{T} = \{\mathbf{I}_i^{\mathcal{T}}\}_{i=1}^{|\mathcal{T}|}$, where for image $\mathbf{I}_i^{\mathcal{T}} = \{\boldsymbol{x}_j^{\mathcal{T}}, y_j^{\mathcal{T}}\}_{j=1}^{N_i}$ there are N_i detected persons $\boldsymbol{x}_j^{\mathcal{T}}$ and the respective importance labels $y_j^{\mathcal{T}}$, such that $y_j^{\mathcal{T}} = 0$ for the "non-important" class and $y_j^{\mathcal{T}} = 1$ for the "important" class. We also have a set of unlabelled images $\mathcal{U} = \{\mathbf{I}_i^{\mathcal{U}}\}_{i=1}^{|\mathcal{U}|}$, where for each image $\mathbf{I}_i^{\mathcal{U}}$ there are N_i detected individuals without any importance annotations $\mathbf{I}_i^{\mathcal{U}} = \{\boldsymbol{x}_j^{\mathcal{U}}\}_{j=1}^{N_i}$. In this work, our goal is to design a method that can learn from partially annotated data. In other words, we aim at developing a model $\hat{\boldsymbol{y}} = f_\theta(\{\boldsymbol{x}_j\}_{\boldsymbol{x}_j \in \mathbf{I}})$ to learn from the augmented training set (*i.e.*, $\mathcal{T} \cup \mathcal{U}$). Here, the model f_θ parameterized by θ takes as input all detected individuals $\{\boldsymbol{x}_j\}_{\boldsymbol{x}_j \in \mathbf{I}}$ in a given image $\mathbf{I}$ and predicts the importance labels for all input persons.

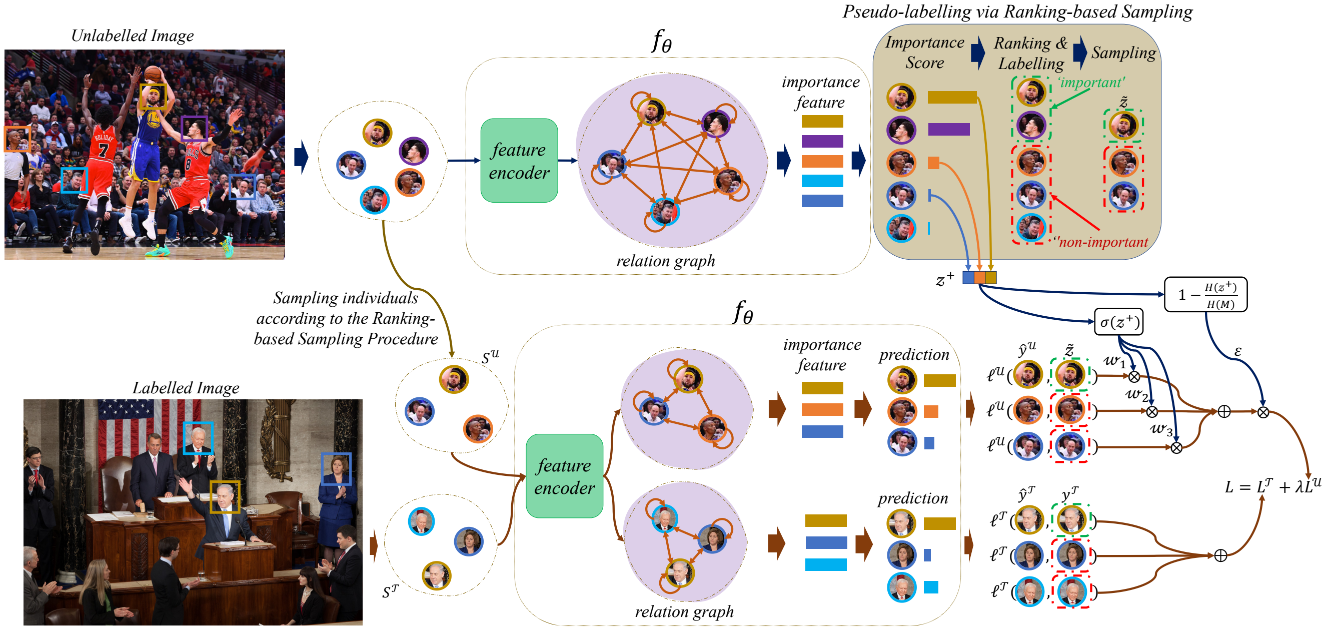


Figure 4. Illustration of our proposed framework. We feed all detected persons into f_θ to estimate the pseudo-labels by Ranking-based Sampling (RankS) according to the ranking and a threshold. We sample a fixed number of individuals in each labelled image or unlabelled image according to labels or pseudo-labels for training. During RankS, we also estimate the importance score weights $\boldsymbol{w}$ as well as the effectiveness weight ε, which indicates the confidence that the unlabelled image features important people (*i.e.*, $\varepsilon = 1$ means that there are important people in the image, while $\varepsilon = 0$ signifies the opposite), to prevent adding too many "non-important" individuals.

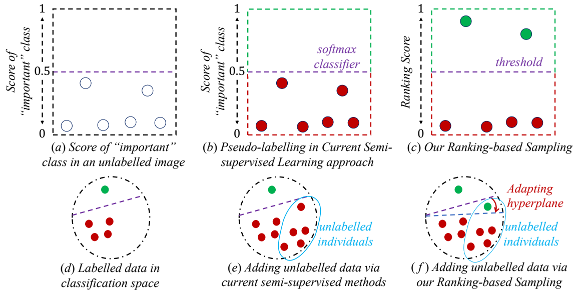

Score of "important" class
Score of "important" class
Ranking Score
1
0.5
0
softmax classifier
threshold
(a) Score of "important" class in an unlabelled image
(b) Pseudo-labelling in Current Semi-supervised Learning approach
(c) Our Ranking-based Sampling
Adapting hyperplane
unlabelled individuals
unlabelled individuals
(d) Labelled data in classification space
(e) Adding unlabelled data via current semi-supervised methods
(f) Adding unlabelled data via our Ranking-based Sampling

Figure 5. Comparison of the pseudo-labelling procedure between our method (RankS) and current semi-supervised approaches. Here, dots are individuals in images: red dots indicate "non-important" individuals and the green ones are "important" people. Existing semi-supervised methods suffer the imbalanced psuedo-labelling problem that all persons in an image are always assigned as "non-important" class (Figure (a)-(b)) and adding those pseudo-labels doesn't help the training (Figure (d)-(e)). In contrast, our method alleviates this problem by the ranking strategy that assigns pseudo-labels according to the ranking score and threshold (Figure (c) and (f)).

To train the model f_θ on $\mathcal{T} \cup \mathcal{U}$, we adopt a fully supervised model POINT [17] and formulate an iterative learning method to update this model. More specifically, we first train the model on one batch of labelled data (for one iteration) to minimize the classification loss of K sampled individuals (K is 8 in this work as in [17]) in each labelled image. For each labelled image, we pick the ground-truth important people and randomly select $K-1$ non-important people, forming the sampled individuals set $\mathcal{S}_i^{\mathcal{T}} = \{\boldsymbol{x}_j\}_{j=1}^{K}$ as in [17]. The trained model is then used to generate the pseudo-labels for K sampled individuals identified in each un-annotated image of an unlabelled images batch by $\boldsymbol{z} = g(f_\theta, \{\boldsymbol{x}_j^{\mathcal{U}}\}_{\boldsymbol{x}_j^{\mathcal{U}} \in \mathcal{S}_i^{\mathcal{U}}})$, where $z_j \in \boldsymbol{z}$ is the pseudo-label of $\boldsymbol{x}_j^{\mathcal{U}}$ estimated by pseudo-label estimation function $g(\cdot)$, and $\mathcal{S}_i^{\mathcal{U}}$ is a set of randomly sampled individuals in $\mathbf{I}_i^{\mathcal{U}}$. For instantiating $g(\cdot)$, we simply apply the softmax operator on the prediction $f_\theta(\{\boldsymbol{x}_j^{\mathcal{U}}\}_{\boldsymbol{x}_j^{\mathcal{U}} \in \mathcal{S}_i^{\mathcal{U}}})$. Finally, we input the unlabelled persons and their pseudo-labels into model f_θ for training.

In this way, we can unify the entire training procedure as optimizing θ to minimize the following loss:

$$\begin{aligned} L &= L^{\mathcal{T}} + \lambda L^{\mathcal{U}} \\ &= \frac{1}{|\mathcal{T}|K}\sum_{i=1}^{|\mathcal{T}|}\sum_{\boldsymbol{x}_j^{\mathcal{T}} \in \mathcal{S}_i^{\mathcal{T}}} \ell^{\mathcal{T}}(\hat{y}_j^{\mathcal{T}}, y_j^{\mathcal{T}}) + \frac{\lambda}{|\mathcal{U}|K}\sum_{i=1}^{|\mathcal{U}|}\sum_{\boldsymbol{x}_j^{\mathcal{U}} \in \mathcal{S}_i^{\mathcal{U}}} \ell^{\mathcal{U}}(\hat{y}_j^{\mathcal{U}}, z_j), \end{aligned} \tag{1}$$

where $\ell^{\mathcal{T}}(\cdot)$ and $\ell^{\mathcal{U}}(\cdot)$ are a classification function (*i.e.*, cross entropy) for labelled data and a loss function for unlabelled data (*e.g.*, mean square error) as in [25, 2, 26], respectively. Additionally, $\hat{y}_j^{\mathcal{T}}$ and $\hat{y}_j^{\mathcal{U}}$ are the predictions of a sampled individual $\boldsymbol{x}_j^{\mathcal{T}}$ and $\boldsymbol{x}_j^{\mathcal{U}}$ in a labelled image $\mathbf{I}_i^{\mathcal{T}}$ and an unlabelled image $\mathbf{I}_i^{\mathcal{U}}$, respectively. λ is the weight of the unlabelled data loss; it is initialized to 0 and linearly increases to its maximum over a fixed number of epochs, which is the well-known linear schedule [2, 26, 20]. Accordingly, the interaction between labelled and unlabelled data is initialized to 0 and is strengthened gradually (*i.e.*, during training the model is increasingly confident and the pseudo-labels of unlabelled images become more reliable).

3.2. Pseudo-labelling by Ranking-based Sampling

An intuitive way to generate pseudo-label (either "important" people or "non-important" people) of images is to learn a classifier using limited labelled data and classify the unlabelled persons in each images. However, the number of important people and of non-important people are always imbalanced in an image, where the former is much smaller than the latter; this would yield an pseudo-labelling imbalance problem that it is highly probable that all individuals will be regarded as "non-important".

To solve this problem, we design a ranking-based sampling (RankS) strategy to predict the pseudo-labels of the individuals. Intuitively, if there are important people in an unlabelled image, some people must be more important than others, which forms a ranking list on the detected persons in an image as shown in Figure 5. Therefore, we introduce the label-guessing function as

$$\mathcal{S}_i^{\mathcal{U}}, \tilde{\boldsymbol{z}} = RankS(f_\theta, \{\boldsymbol{x}_j^{\mathcal{U}}\}_{\boldsymbol{x}_j^{\mathcal{U}} \in \mathbf{I}_i^{\mathcal{U}}}, \alpha, K), \tag{2}$$

where $RankS(\cdot)$ is the ranking-based sampling procedure used to generate pseudo-labels $\tilde{\boldsymbol{z}}$ based on the importance score of all individuals in $\mathbf{I}_i^{\mathcal{U}}$ by using the previous iteration's trained model f_θ, K is the number of sampled individuals in each unlabelled images as illustrated in Sec. 3.1 and α is a hyper-parameter for assigning hard pseudo-labels to K sampled individuals according to the ranking.

To be detailed, we use the relation network in [17] as the backbone (*i.e.*, f_θ), which takes as input a set of people per image to build relation graph and encode features for importance classification from the relation graph. We first apply the model trained at the previous iteration and the softmax operator to compute the probability of the "important" category as the importance score, ranking all individuals detected in the same unlabelled image in the descending order of importance score by scaling the importance score with the maximum score, resulting in a ranking score. We then assign "important" pseudo-label to those individuals whose ranking score is higher than a threshold α and the rest are assigned as "non-important". We pick the top-1 "important" individual and randomly select $K-1$ "non-important" people to form $\mathcal{S}_i^{\mathcal{U}}$, which is used for training coupled with the pseudo-labels of sampled individuals.

Therefore, replacing the pseudo-labels with those generated by RankS allows the unlabelled loss term in Eq. 1 to be rewritten as

$$L^{\mathcal{U}} = \frac{1}{|\mathcal{U}|K} \sum_{i=1}^{|\mathcal{U}|} \sum_{\boldsymbol{x}_j^{\mathcal{U}} \in \mathcal{S}_i^{\mathcal{U}}} \ell^{\mathcal{U}}(\hat{y}_j^{\mathcal{U}}, \tilde{z}_j). \tag{3}$$

In this way, we regularize the consistency between the sampled individuals' pseudo-labels estimated from the full relation graph (*i.e.*, relation graph of all individuals in $\mathbf{I}_i^{\mathcal{U}}$) and the prediction from the sub-graph (*i.e.*, the relation graph of $\mathcal{S}_i^{\mathcal{U}}$) as illustrated in Figure 4. That is, we force a constraint that the importance of the person estimated based on a subset of persons selected by our Ranking sampling should be close to the one estimated from all person detected in an image. Thanks to the ranking and labeling in RankS, the pseudo-labels $\hat{z}$ alleviate the imbalance problem to some extent, as this approach avoids the problem of assigning all people in an image the "non-important" label during pseudo-labelling (Figure 5 and 2).

3.3. Balancing Loss via Importance Score Weighting

Still, we treat the respective people in unlabelled images equally, while there are much more "non-important" samples than "important" samples in unlabelled images, which could make the pseudo-labelling imbalance problem remain.

To further alleviate the pseudo-labelling imbalance problem, instead of assigning each person in each unlabelled image the same weight (*i.e.*, $\frac{1}{K}$ for K sampled individuals in an image), we introduce a person-specific weight w, called importance score weight (ISW), into the unlabelled data loss term in Eq. 3 so that the contribution of the "important" person will be strengthened, and those of "non-important" ones will be weakened. For this purpose, we can rewrite Eq. 3 as

$$L^{\mathcal{U}} = \frac{1}{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \sum_{\boldsymbol{x}_j^{\mathcal{U}} \in \mathcal{S}_i^{\mathcal{U}}} w_j \ell^{\mathcal{U}}(\hat{y}_j^{\mathcal{U}}, \tilde{z}_j), \text{ s.t. } \sum_{j=1}^{K} w_j = 1, w_j > 0. \tag{4}$$

To estimate the weight w_j of person $\boldsymbol{x}_j^{\mathcal{U}}$, we first consider the probability of the "important" class z_j^+ for $\boldsymbol{x}_j^{\mathcal{U}}$ and treat z_j^+ as the importance score. As we mentioned in Sec. 3.2, given an unlabelled image, K persons are sampled, and their importance scores form an importance score vector $\boldsymbol{z}^+$. We then apply the normalization function to $\boldsymbol{z}^+$, which results in normalized importance score weights $\boldsymbol{w} = \sigma(\boldsymbol{z}^+) = (w_1, w_2, \cdots, w_K)$, where $\sigma(\cdot)$ is a normalization function applied to $\boldsymbol{z}^+$, so that the constraint in Eq. 4 is satisfied. In this work, instead of using a hard importance score weight (*i.e.*, $w_j \in \{0, 1\}$), we use softmax to obtain a soft importance score weight (*i.e.*, $w_i \in [0, 1]$), so that our model has a better tolerance on the bias of computing importance scores. Here, we do not apply importance score weighting to labelled data for couple reasons. First, the number of unlabelled images is much larger than that of labelled data and the imbalance problem has been largely mitigated by importance score weighting. Second, as "non-important" individuals in labelled data have ground-truth annotations, we consider using this more reliable information and weakening the effect of unlabelled "non-important" individuals.

3.4. Detecting Noisy Unlabelled Images

Apart from the imbalance problem, it is essential that the model should be able to detect and neglect noisy images with no detected important people. For unlabelled images, it is not guaranteed that all images contain important people. To solve this problem, we further estimate an effectiveness weight (EW) ε, a continuous varying value between 0 and 1, reflecting the confidence that an unlabelled image features important people (*i.e.*, $\varepsilon = 1$ means that there are important people in the image, while $\varepsilon = 0$ signifies the opposite). We apply this weight into Eq. 4 as follows:

$$\begin{aligned} L^{\mathcal{U}} &= \frac{1}{|\mathcal{U}|}\sum_{i=1}^{|\mathcal{U}|}\varepsilon_i \sum_{\boldsymbol{x}_j^{\mathcal{U}}\in\mathcal{S}_i^{\mathcal{U}}} w_j \ell^{\mathcal{U}}(\hat{y}_j^{\mathcal{U}}, \tilde{z}_j) \\ &\text{s.t.} \quad w_j \in \boldsymbol{w} = \sigma(\boldsymbol{z}^+), 0 \le \varepsilon_i \le 1, \end{aligned} \tag{5}$$

where ε_i acts as a gate to enable the model to choose or neglect the $i-th$ unlabelled image. Inspired by [9], we consider specifying ε using the entropy of the importance score $\boldsymbol{z}^+$. In particular, if there are important persons, those persons' importance scores will be high, and the other people's importance scores will remain low (*i.e.*, the entropy will be low). In contrast, if there are no important people, the importance scores of all persons in the respective unlabelled image will be almost uniform (*i.e.*, the entropy will be high). To constrain ε between 0 and 1, we specify ε as

$$\varepsilon = 1 - \frac{H(\boldsymbol{z}^+)}{H(M)}, \tag{6}$$

where $H(\cdot)$ is the entropy function. Additionally, M is a vector with the same dimension as $\boldsymbol{z}^+$, and all elements of M are equal. Vector M in Eq. 6 simulates the possible case of no important people, and thus, $H(M)$ is the maximum possible entropy of each unlabelled image. In this equation, if there are no important people in an unlabelled image, $H(\boldsymbol{z}^+)$ will be equal to $H(M)$, resulting in $\varepsilon = 0$, *i.e.*, noisy unlabelled images will be neglected, or their effect will be weakened.

Therefore, replacing the unlabelled data loss term in Eq. 1 with Eq. 5, we formulate our complete method as

$$\begin{aligned} L &= L^{\mathcal{T}} + \lambda L^{\mathcal{U}} \\ &= \frac{1}{|\mathcal{T}|K}\sum_{i=1}^{|\mathcal{T}|}\sum_{\boldsymbol{x}_j^{\mathcal{T}}\in\mathcal{S}_i^{\mathcal{T}}} \ell^{\mathcal{T}}(\hat{y}_j^{\mathcal{T}}, y_j^{\mathcal{T}}) + \frac{\lambda}{|\mathcal{U}|}\sum_{i=1}^{|\mathcal{U}|}\varepsilon_i \sum_{\boldsymbol{x}_j^{U}\in\mathcal{S}_i^{\mathcal{U}}} w_j \ell^{\mathcal{U}}(\hat{y}_j^{\mathcal{U}}, \tilde{z}_j) \\ &\text{s.t.} \quad w_j \in \boldsymbol{w} = \sigma(\boldsymbol{z}^+), \varepsilon_i = 1 - \frac{H(\boldsymbol{z}^+)}{H(M)}. \end{aligned} \tag{7}$$

By introducing three strategies (*i.e.*, ranking-based sampling, importance score weighting and effectiveness weighting) to the basic semi-supervised learning pipeline, our proposed method enables the collaboration between labelled data and unlabelled data to benefit the overall model training.

4. Experiments

In this work, we conduct extensive experiments on two large datasets collected in the course of this research to investigate the effect of the use of unlabelled images on important people detection and evaluate our proposed semi-supervised important people detection method. More detailed information about datasets, essential network (*i.e.*, POINT) and additional experiments results are reported and analyzed in the Supplementary Material.

4.1. Datasets

Due to the lack of datasets for semi-supervised learning-based important people detection, we augment two datasets, namely, the MS and NCAA datasets from [18], by collecting a large number of unlabelled images from the internet, and forming the Extended-MS (EMS) and Extended-NCAA (ENCAA) datasets.

The EMS Dataset contains $10,687$ images featuring more than six types of scenes, of which 2310 images are from the MS dataset, and 8377 images were obtained by directly crawling the web [1]. For both labelled data and unlabelled data, a face detector [15] is used to detect all possible persons, and bounding boxes are provided. Similar to [17], the EMS dataset is split into three parts: a training set (8607 images, consisting of 690 labelled samples and 8377 unlabelled samples), a validation set (230 labelled samples), and a testing set (1390 labelled samples).

The ENCAA Dataset. Based on 9736 labelled images from the NCAA dataset, we collect $19,062$ images from the internet by extracting frames from numerous basketball videos and filtering out images that do not feature multiple players. Similar to the construction of the EMS dataset, we also divide the ENCAA dataset into three parts: 2825 labelled samples picked randomly and all unlabelled samples form the training set; 941 randomly selected labelled samples are used as a validation set; and the remaining labelled samples (*i.e.*, 5970 images) constitute the testing set. Each person's bounding box is generated by an existing object detector, namely, the YOLOv3 detector [22].

4.2. Baselines

In this work, we use the state-of-the-art fully supervised method as the baseline to evaluate our methods. In addition, we adapt three recent semi-supervised methods, namely, Pseudo Label (PL) [12], Mean Teacher (MT) [25] and Label Propagation (LP) [9], to important people detection.

[1] We collected unlabelled images from the internet by searching for various social event topics such as "graduation ceremony".

POINT. We adopt the POINT [17] method, a state-of-the-art method of important people detection, as the baseline, which we train only on labelled data using a fully supervised learning approach.

Pseudo Label is a simple yet efficient semi-supervised learning approach for ordinary classification tasks, which chooses the class with the maximum predicted probability as the true label for each unlabelled sample.

Mean Teacher maintains two models: student and teacher. Given unlabelled samples, the outputs of the teacher model are used as pseudo-labels. The consistency loss is determined over the predictions of unlabelled images predicted by student model and the pseudo-labels generated by the teacher model such that the learned model can be invariant to stochastic noise between student and teacher models.

Label Propagation infers the pseudo-labels of unlabelled samples from the nearest neighbour graph, which is constructed based on the embeddings of both labelled and unlabelled samples.

4.3. Implementation Details

We implement all methods in PyTorch. For a fair comparison, we adopt POINT (we have detailed it in Supplementary Material) as the essential network with SGD used as the optimizer in our method as well as other semi-supervised baselines (*i.e.*, PL, MT and LP). We run all methods for 200 epochs and use the same hyper-parameters for all methods. The hyper-parameter α is learned on the validation data and is set to 0.99 for all the experiments. The weight decay is 0.0005 and the momentum is 0.9 in all experiments. The learning rate is initialized to 0.001, and we follow the learning rate update strategy of [17], *i.e.*, the learning rate is scaled by a factor of 0.5 every 20 epochs. We adopt the commonly used linear schedule to update weight λ, *i.e.*, we increase λ linearly from 0 to its maximum (*i.e.*, 1) over 35 epochs. We follow the standard evaluation metric in [17], *i.e.*, the mean average precision is reported to measure the performance of all methods.

4.4. Comparisons with Related Methods

We first compare our method with current semi-supervised learning methods adapted for important people detection and the fully supervised learning baseline. From Table 1, it is worth noting that the recent semi-supervised learning approaches attain comparable results (*e.g.*, the results of LP vs. those of POINT are 88.61 % vs. 88.21 % on the ENCAA dataset if 66 % of labelled images are used) but sometimes underperform the fully supervised baseline (*e.g.*, the results of LP vs. those of POINT are 86.66 % vs. 88.48 % on the EMS dataset if all labelled images are used). In contrast, our method achieves a significant and consistent improvement over the baseline; *e.g.*, After adding unlaballed images, our method outperforms the fully supervised baseline by 4.45 % and 4.15 % on the EMS and ENCAA datasets, respectively, in the regime with fewer labels (33 %). These results of PL, LP and MT clearly demonstrate that treating each person independently are unable to leverage valuable information from unlabelled images to help training. On the contrary, the results of our method indicate that three proposed strategies enable our method to effectively leverage the information of unlabelled images to assist in training on a limited quantity of labelled data and significantly boost performance.

Dataset	EMS			ENCAA		
#labelled images	33 %	66 %	100 %	33 %	66 %	100 %
POINT (fully supervised)	83.36	85.97	88.48	84.60	88.21	89.75
Pseudo Label (PL)	83.37	85.35	88.57	85.70	88.43	90.56
Label Propagation (LP)	82.34	86.33	86.66	85.36	88.61	90.18
Mean Teacher (MT)	84.50	86.29	87.55	83.33	84.66	87.55
Ours	**87.81**	**88.44**	**89.79**	**88.75**	**90.86**	**92.03**

Table 1. Comparison with related methods on both datasets.

4.5. Effect of the Proportion of Labelled Images

To further understand the factors that affect the performance of semi-supervised important people detection, we evaluate our method using different portions of labelled images. We randomly select 33 %, 66 % and 100 % of labelled images, and the remaining labelled images together with unlabelled images are used WITHOUT labels. We report the results in Table 1, Table 2 and Table 3. It is clearly observed that using more labelled data can boost the overall performance of important people detection, which also enables the semi-supervised model to estimate more accurate pseudo-labels for unlabeled images and further boost performance. It also indicates that developing a semi-supervised model that can correctly predict pseudo-labels and combine them with the labelled training set is necessary. From another point of view, the results shown in Table 2 imply that our method can consistently outperform the fully supervised approach as well as related baselines and clearly demonstrate the consistent efficacy of the three proposed strategies.

4.6. Ablation Study

We conduct ablation study to investigate the effect of three proposed strategies (*i.e.*, ranking-based sampling (RankS), importance score weighting (ISW) and effectiveness weighting (EW)) on important people detection and shown the results in Table 2, where "Ours$_{\text{w/o ISW and EW}}$" indicates our method using RankS only.

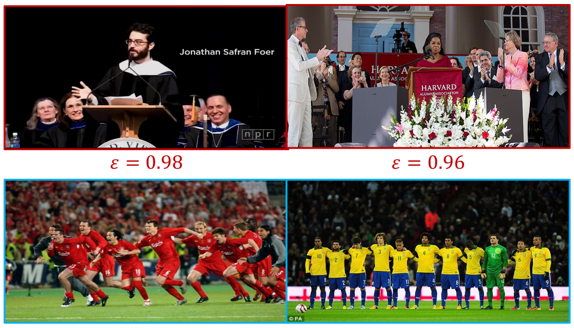
In Table 2, it is evident that all strategies can improve performance in most label regimes, and the ranking-based sampling strategy attains the greatest improvement; for instance, on the ENCAA dataset, if 33 % of labelled images are used, the method "Ours$_{\text{w/o ISW and EW}}$" outperforms the "Ours$_{\text{w/o RankS, ISW and EW}}$" by 2.78 %. This result clearly shows that the ranking-based sampling enables that the relatively high score should be labelled as "important" and

Dataset	EMS			ENCAA		
#labelled images	33 %	66 %	100 %	33 %	66 %	100 %
Ours$_{\text{w/o Ranks, ISW and EW}}$	83.70	86.81	87.67	84.35	87.66	89.93
Ours$_{\text{w/o ISW and EW}}$	85.55	87.25	88.53	87.13	90.53	91.49
Ours$_{\text{w/o EW}}$	86.34	87.45	89.67	87.68	90.60	92.00
Ours	87.81	88.44	89.79	88.75	90.86	92.03

Table 2. Ablation study on both datasets. RankS represents ranking-based sampling while ISW and EW indicate importance score weighting and effectiveness weighting, respectively. Ours$_{\text{w/o ISW and EW}}$ means our model without using ISW and EW.

Dataset	EMS			ENCAA		
#labelled images	33 %	66 %	100 %	33 %	66 %	100 %
Ours$_{\text{LP}}$	87.51	88.10	89.65	88.95	91.06	91.98
Ours$_{\text{MT}}$	87.23	88.56	90.72	88.97	90.93	91.62
Ours	87.81	88.44	89.79	88.75	90.86	92.03

Table 3. Evaluation of different techniques (*i.e.*, LP and MT) when used for instantiating pseudo-label estimation function (*i.e.*, $g(\cdot)$) instead of using Softmax function.



$\varepsilon = 0.98$ $\varepsilon = 0.96$

$\varepsilon = 0.13$ $\varepsilon = 0.34$

Figure 6. Examples of unlabelled images and their effectiveness weights estimated automatically by our method.

the rest remain "non-important" when predicting pseudo-labels within each unlabelled image, preventing assigning all "non-important" or all "important" pseudo-labels during label guessing in an image. This is also verified by Figure 2, where our method correctly predicts pseudo-labels for all individuals (Figure 2(b)) during training and estimate accurate importance scores at the end (*e.g.*, Figure 2(a)) while current semi-supervised learning approaches (*i.e.*, LP and MT) assign all individuals as "non-important" samples

From Table 2, we also observe that adding importance score weighting (ISW) can consistently albeit slightly boost the performance (*e.g.*, the results of "Ours$_{\text{w/o EW}}$" vs. those of "Ours$_{\text{w/o ISW and EW}}$" are 89.67 % vs. 88.53 % on the EMS if all labelled images are used). This indicates that ISW is able to alleviate the problem of data imbalance and ultimately benefits the training of important people detection.

In addition, comparing the full model and our model using both RankS and ISW, we clearly observe that the estimated effectiveness weight (EW, defined in Eq. 6) improves the performance (*e.g.*, "Ours" improves the performance of "Ours$_{\text{w/o EW}}$" from 86.34 % to 87.81 % on EMS if 33 % of labelled images are used). This implies that our effectiveness weighting strategy is able to detect and neglect noisy unlabelled images with no important people, and this benefits important people detection. To further better understand how the effectiveness weight works, we visualize EW of several unlabelled images and present them in Figure 6. We clearly observe that if there are no important people in the unlabelled image, EW is small (as shown in the second row in Figure 6), while if important people are present, EW is nearly 1 (as shown in the first row in Figure 6). This result again clearly demonstrates the efficacy of our proposed EW on detecting noisy images and neglecting noisy samples during training.

Additionally, we also evaluate the effect of different techniques (*i.e.*, LP and MT) used to estimate importance score in our method during pseudo-labelling in Table 3, where "Ours$_{\text{LP}}$" implys our method using Label Propagation technique for importance score estimation during pseudo-labelling. It is clearly shown in Table 3 that the variants of method using different techniques for importance score estimation yield similar results, which demonstrates the stableness of our methods.

5. Conclusion

In this work, we study semi-supervised learning in the context of important people detection and propose a semi-supervised learning method for this task. Compared with recent semi-supervised learning approaches, our method is shown to be able to effectively leverage the information of unlabelled images to assist in model training. We also conduct extensive experiments on important people detection by a semi-supervised learning method, and the results confirm that 1) the pseudo-labels of individuals in a given unlabelled image should have the special pattern in important people detection (*i.e.*, the relatively high score should be labelled as "important" and the rest remains "non-important"), and our proposed ranking-based sampling is able to achieve this; 2) our importance score weighting can alleviate the imbalance problem and boost performance; and 3) enabling the model to neglect the noisy unlabelled images with no important people is important during semi-supervised learning. By our learning, we are able to avoid costly labelling on important people detection and achieve satisfactory performance.

6. Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), NSFC(U1911401,U1811461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), Guangdong NSF Project (No. 2018B030312002), Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03).

References

[1] 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.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

[3] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *Transactions on Neural Networks*, 20(3):542–542, 2009.

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In *European Conference on Computer Vision*, 2018.

[5] WeiWang Dong-DongChen and Zhi-HuaZhou WeiGao. Tri-net for semi-supervised deep learning. In *International Joint Conferences on Artificial Intelligence*, 2018.

[6] Shreya Ghosh and Abhinav Dhall. Role of group level affect to find the most influential person in images. In *European Conference on Computer Vision*, 2018.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[8] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, 2005.

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

[10] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

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

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

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

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

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

[16] Wei-Hong Li, Chuan-Sheng Foo, and Hakan Bilen. Learning to impute: A general framework for semi-supervised learning. *arXiv preprint arXiv:1912.10364*, 2019.

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

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

[19] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[20] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 2018.

[21] 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. In *Computer Vision and Pattern Recognition*, 2016.

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

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

[24] 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.

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

[26] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*, 2019.

[27] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, 1995.