Hierarchical Online Instance Matching for Person Search

Di Chen,[1,4] **Shanshan Zhang,**[1,3]* **Wanli Ouyang,**[5] **Jian Yang,**[1,2]* **Bernt Schiele**[4]
[1]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministryof Education,
School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]Jiangsu Key Lab of Image and Video Understanding for Social Security
[3]Science and Technology on Parallel and Distributed Processing Laboratory (PDL)
[4]MPI Informatics, [5]The University of Sydney, SenseTime Computer Vision Research Group, Australia
{dichen, shanshan.zhang, csjyang}@njust.edu.cn, {dichen, schiele}@mpi-inf.mpg.de, wanli.ouyang@sydney.edu.au

Abstract

Person Search is a challenging task which requires to retrieve a person's image and the corresponding position from an image dataset. It consists of two sub-tasks: pedestrian detection and person re-identification (re-ID). One of the key challenges is to properly combine the two sub-tasks into a unified framework. Existing works usually adopt a straightforward strategy by concatenating a detector and a re-ID model directly, either into an integrated model or into separated models. We argue that simply concatenating detection and re-ID is a sub-optimal solution, and we propose a Hierarchical Online Instance Matching (HOIM) loss which exploits the hierarchical relationship between detection and re-ID to guide the learning of our network. Our novel HOIM loss function harmonizes the objectives of the two sub-tasks and encourages better feature learning. In addition, we improve the loss update policy by introducing Selective Memory Refreshment (SMR) for unlabeled persons, which takes advantage of the potential discrimination power of unlabeled data. From the experiments on two standard person search benchmarks, *i.e*. CUHK-SYSU and PRW, we achieve state-of-the-art performance, which justifies the effectiveness of our proposed HOIM loss on learning robust features.

Introduction

In video surveillance, pedestrian detection and person re-identification are two important tasks. They have attracted a lot of research attention in recent years individually. Pedestrian detection requires to detect the bounding box of persons in a panorama image regardless of their identity information, while person re-ID aims to match the reappearance of a probe person within pre-defined bounding boxes. Both tasks are important but they are not directly applicable to a surveillance system because of their limited functionality. Therefore, *person search* is introduced by (Xu et al. 2014), which consolidates the two sub-tasks into a unified system. However, the task of person search is more challenging since it gathers domain-specific difficulties from the two sub-tasks together, *e.g*. viewpoint and pose variance, occlusion, complex background, false alarms, misalignments, *etc*.

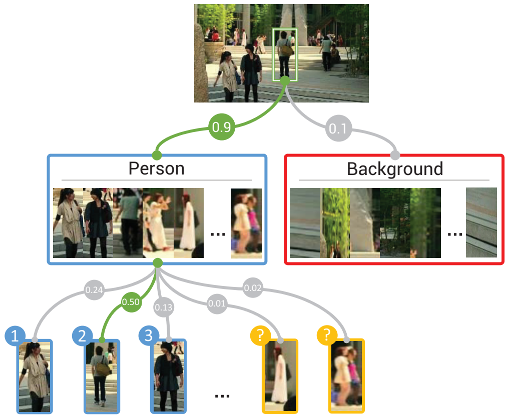

Figure 1: The hierarchical relationship between detection (the first layer) and re-identification (the second layer).

On the topic of person search, a few works have been proposed since the release of two large-scale person search datasets (Xiao et al. 2017b; Zheng et al. 2017). Generally, these works can be clustered into two categories: *one-stage* methods (Xiao et al. 2017b; 2017a; Liu et al. 2017a; Chang et al. 2018; Munjal et al. 2019; Yan et al. 2019) that solve detection and re-ID in an end-to-end model and *two-stage* methods (Zheng et al. 2017; Chen et al. 2018; Lan, Zhu, and Gong 2018) that split detection and re-ID into two independent models. One-stage methods aim to learn a mutual representation shared by the two sub-tasks, but they suffer from contradictory objectives of the two sub-tasks and conflicting focusing points as discussed in (Chen et al. 2018), since they are simple concatenations of a detector and a linear embedding layer without harmonizing the two sub-tasks. On the other hand, two-stage methods avoid this issue by separated modeling and they usually yield better performance, but neglect the inter-dependency between pedestrian detection and re-ID and incur a higher computation cost.

Motivated by the discussion above, we propose to explicitly model the relationship between pedestrian detection and re-ID, and further exploit it as a prior to guide the one-stage model learning. Specifically, detection and re-ID form

*Corresponding author.

a *hierarchical structure* with two layers, as illustrated by Fig. 1. The first layer describes the goal of pedestrian detection, which captures the human *commonness* and aims to distinguish person from background clutter given an arbitrary proposal. The second layer interprets the objective of re-ID, which is to classify the persons' identities according to their *uniqueness*. Based on the two-layered tree structure, we propose the Hierarchical Online Instance Matching (HOIM) loss, which formulates the hierarchical relationship with the law of total probability. By encoding the task relationship into the training stage, the tasks of detection and re-ID are deeply aggregated. hence the learned embeddings are more robust since they describe the human *commonness* for person/background separation and *uniqueness* for person identity discrimination simultaneously. Additionally, the detector performance, which is crucial to person search accuracy, is improved since the contradictory objectives of detection and re-ID is alleviated. Moreover, our method consumes less memory and computational resources compared to two-stage methods, since extracted features are shared between the two sub-tasks.

Another major difference between person search w.r.t. re-ID is that the datasets contain people with unknown identity, which is under-exploited by conventional re-ID loss functions. To take advantage of these unlabeled data, Online Instance Matching (OIM) loss (Xiao et al. 2017b) is proposed by treating them as negative samples, which adopts a first-in-first-out (FIFO) strategy to manage the embedding buffer. However, FIFO only focuses on the life cycle of the embeddings, ignoring the relative importance between the candidates and memorized entries. For instance, two persons with similar appearances in the same batch would *both* be pushed into the buffer by the 'first-in' rule and increase feature redundancy, while discriminative embeddings located at the tail of the buffer would be incorrectly popped out, following the 'first-out' rule. Therefore, we propose a Selective Memory Refreshment (SMR) strategy that assigns each embedding an importance weight, which jointly considers the hardness, diversity and existing time of each entry. A new embedding would be pushed into the buffer only if its importance weight is larger than the minimum value of the existing ones. At the same time, the embedding with the smallest importance would be popped out. The proposed SMR policy excels over the FIFO strategy of OIM in that it filters out easy and redundant samples, hence improving the feature learning efficiency.

In summary, our contribution is three-fold:

- A novel Hierarchical Online Instance Matching loss is proposed to learn better embeddings for one-stage person search models.
- We introduce a new Selective Memory Refreshment strategy to optimize the loss updating procedure of OIM and HOIM loss.
- Our proposed method achieves state-of-the-art performance with a smaller model size and faster running speed.

Related Work

Person Search Person Search has drawn a lot of attention recently. A straightforward solution is to use a pedestrian detector and a person encoder sequentially. Zheng et al. conduct a systematic evaluation on various detectors and re-ID descriptors, and provide novel insights on re-weighting the person matching similarity with detection confidence score. Lan, Zhu, and Gong point out the multi-scale matching problem and proposed the Cross-Level Semantic Alignment method for person search. Chen et al. first reveal that for the person search problem, two-stage methods are superior to one-stage ones without reconciling the objective contradiction. They also propose to exploit background information as a complementary cue for person matching.

Apart from two-stage methods, end-to-end models based on Faster R-CNN are also a popular choice. Early works concatenate an auxiliary linear layer upon the top convolutional layer of Faster R-CNN for re-ID embedding generation. The whole model is jointly trained under the supervision of standard Faster R-CNN losses and OIM loss (Xiao et al. 2017b) or Center Loss (Xiao et al. 2017a; Wen et al. 2016). Instead of generating a number of boxes on an image, Liu et al. propose to match the probe person directly on the panorama image by recursively shrinking the search area. Chang et al. adopt a similar idea and further build the first deep reinforcement-learning-based person search framework. Yan et al. propose a sophisticated matching method by building a graph using nearby persons as context information. Munjal et al. fuse the query information into a siamese network as guidance for feature learning, proposal generation and similarity calculation. All of the above methods ignore the association between detection and re-identification. However, in this paper, we show that encoding the task relationship as a prior into the network training procedure is beneficial for person search.

Person re-ID Early works on person re-ID mainly focus on designing hand-crafted features (Wang et al. 2007; Farenzena et al. 2010; Zhao, Ouyang, and Wang 2013; Liao et al. 2015) and effective distance metrics (Kostinger et al. 2012; Li et al. 2015; Zhang, Xiang, and Gong 2016). Recent progress is generally based on deep neural networks which can be grouped into two classes, *i.e. siamese models* which are trained with ranking losses such as contrastive loss (Yi et al. 2014; Li et al. 2014; Ahmed, Jones, and Marks 2015; Varior et al. 2016; Liu et al. 2017b; Xu et al. 2018) or triplet loss (Ding et al. 2015; Cheng et al. 2016), and *identification models* (Xiao et al. 2016; Zheng et al. 2016; 2017; Fan et al. 2018; Xiang et al. 2018) which adopt cross entropy loss for supervision. Fan et al. and Xiang et al. proposed to project both the embeddings and classification weight vectors onto a hypershpere. Our proposed HOIM loss has a similar formulation, which renormalizes the embeddings to have unit length. However, our method replaces the linear projection weight with embedding buffers, which eliminates the need for training by adaptively making use of the embedding characteristics. Moreover, our method is scalable to unlabeled persons and conducts implicit hard mining via Selective Memory Refreshment.

Pedestrian Detection Hand-crafted features and linear classifiers are commonly adopted by early works on pedestrian detection, among which DPM (Felzenszwalb et al. 2009), ACF (Dollar et al. 2014) and ICF (Dollar et al. 2009;

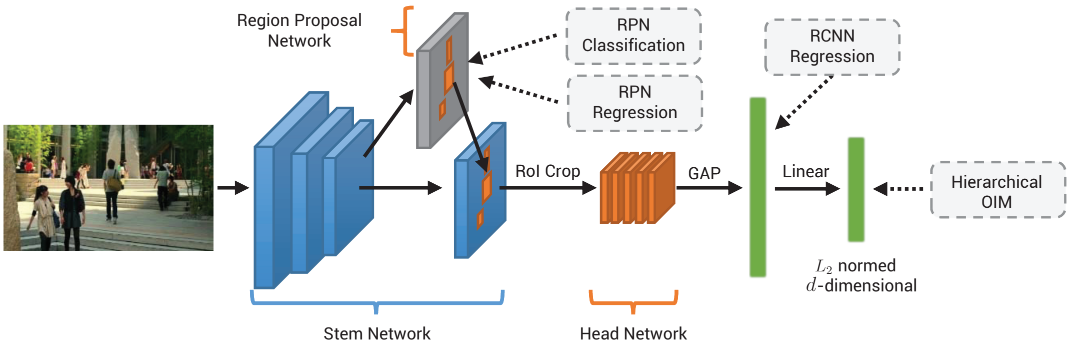

Figure 2: Overview of our framework. Our network is based on Faster R-CNN (Ren et al. 2017) with a ResNet-50 (He et al. 2016) backbone. An extra L_2-normalized linear projection layer is appended for person embedding generation. We remove the second-stage classification loss of the original Faster R-CNN and integrate its functionality into our HOIM loss.

Zhang, Bauckhage, and Cremers 2014; Zhang, Benenson, and Schiele 2015) are three classical methods. Deep neural networks have been popular in recent years (Zhang et al. 2016; 2018; Ouyang and Wang 2013; 2012) for its exceptional feature learning capacity. Some of the CNN-based works utilize the R-CNN architecture, which employs weak pedestrian detectors such as ICF (Zhang et al. 2016; 2018) as proposals and CNNs for refinement. Faster R-CNN (Ren et al. 2017) takes one step further by incorporating proposal generation and refinement into an end-to-end network. It achieves leading performance on pedestrian detection by applying proper adjustments (Zhang, Benenson, and Schiele 2017; Zhang, Yang, and Schiele 2018). Our work is also based on the adjusted Faster R-CNN, except that the classification loss on the head network is merged into our proposed HOIM loss.

Method

In this section, we first present an overview of the whole structure of our model. Then, we revisit the design philosophy of OIM loss. Next, we describe the details of the hierarchical relationship between pedestrian detection and re-ID, which aims to augment the vanilla OIM loss. Subsequently, we describe the SMR strategy to enhance HOIM discrimination power. Finally, a useful add-on component to improve the performance is introduced.

Overview

As shown in Fig. 2, our model is based on Faster R-CNN (Ren et al. 2017), which is composed of a stem network for spatial feature learning, a region proposal network (RPN) for proposal sampling, and a head network (R-CNN) for feature fine-tuning and classification. An extra L_2-normalized linear layer is added upon the top of the head network to generate the feature embeddings for re-ID.

We follow the network configurations in (Xiao et al. 2017b) except that we replace the second-stage classification loss and OIM loss with our HOIM loss. Together with RPN classification loss, RPN and R-CNN bounding box regression losses, we train the whole model jointly using Stochastic Gradient Descent (SGD).

During inference, our model takes in a scene image and produces several bounding boxes as query candidates. Each box also includes a detection confidence score and a d-dimensional embedding. To find a specific probe person, we first extract his/her embedding by freezing RPN and using the provided box for RoI pooling directly, then calculate the cosine similarities with the candidate embeddings. Finally, a ranking list of the candidates is formed by sorting the similarity scores in descending order.

Revisiting Online Instance Matching

OIM loss is the basis of our proposed HOIM loss. It is specially designed for the person search task, which has two major differences to re-ID. Firstly, person search datasets include individuals with unknown identities. The widely used loss functions in re-ID task, *i.e.* Cross Entropy and Triplet loss, are not scalable to these data, while OIM loss makes use of them by enacting them as negative samples to the labeled data, as is shown in the first and second row in Fig. 3. Secondly, the training iterations of an end-to-end model for simultaneous detection and re-ID are more ill-conditioned than vanilla re-ID models, since the batch size is usually much smaller, which in result reduces the class and instance diversity. Therefore, Cross Entropy and Triplet loss would suffer from unstable gradients and the consequent convergence failure. Whereas the non-parametric property and the tailored update policy of OIM loss ensure effective training.

Suppose there are N different identities in the training dataset. OIM loss constructs a look-up table with size $N \times d$ to memorize the class centroid embeddings. Another circular queue with size $M \times d$ is built up to store diverse embeddings of unlabeled persons. Together the look-up table and circular queue forms a projection matrix $\mathbf{W} \in \mathbb{R}^{(N+M)\times d}$. Different from the projection matrix in a vanilla softmax layer, $\mathbf{W}$ here is considered as an *external buffer*, which is updated separately from the model parameters.

Given a person embedding $\mathbf{x} \in \mathbb{R}^d$, the cosine similarities between $\mathbf{x}$ and the memorized embeddings are calculated by a linear projection, as is comprehensively illustrated in the

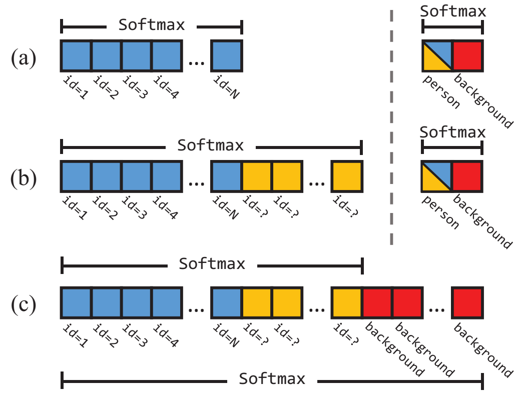

Figure 3: Comparison between vanilla softmax loss (a), OIM loss (b) and HOIM loss (c). Blue, yellow and red squares represent logits/cosine similarities of labeled persons, unlabeled persons and background-clutters between the current sample respectively.

second row of Fig. 3.

$$\mathbf{s} = \mathbf{W}\mathbf{x} \in \mathbb{R}^{N+M}, \qquad (1)$$
$$\text{where } \mathbf{s} = [s_1, s_2, \dots, s_N, s_{N+1}, \dots, s_{N+M}]$$

Then the probability of $\mathbf{x}$ belonging to identity t (denoted as $\text{id} = t$), given the fact that $\mathbf{x}$ also represents a person (denoted as Λ), could be produced by a softmax function:

$$p(\text{id} = t, \Lambda) = \frac{e^{s_t/\tau}}{\sum_{j=1}^{N} e^{s_j/\tau} + \sum_{j=N+1}^{N+M} e^{s_j/\tau}} \qquad (2)$$

where τ is a temperature factor to control the softness of the probability distribution. The final objective is to minimize the negative log-likelihood:

$$\mathcal{L}_{\text{OIM}} = -\mathbb{E}_{\mathbf{x}}[\log p(\text{id} = t, \Lambda)], \quad t = 1, 2, \dots, N. \qquad (3)$$

During training, the look-up table is incrementally updated with a momentum of η:

$$\mathbf{w}_t \leftarrow \eta \mathbf{w}_t + (1 - \eta)\mathbf{x}, \quad \text{if } \mathbf{x} \text{ belongs to class } t. \qquad (4)$$

The circular queue pops out old embeddings and pushes in new ones to preserve a fixed size.

Hierarchical Relationship between Detection & Re-ID

OIM loss deals only with person embeddings and leaves background to another Cross Entropy loss. Our proposed HOIM loss is meant to integrate the hierarchical structure of pedestrian detection and person re-ID into OIM loss explicitly, in order to jointly process person and background embeddings. The projection matrix $\mathbf{W}$ is firstly expanded to a size of $(N+M+B) \times d$, denoted as $\mathbf{W}'$, by concatenating another circular queue with size $B \times d$ to memorize a number of background embeddings. For an arbitrary embedding $\mathbf{x}$ w.r.t. a valid proposal, we first calculate the augmented probability distribution by projection and softmax normalization:

$$\mathbf{s}' = \mathbf{W}'\mathbf{x} \in \mathbb{R}^{N+M+B}, \qquad (5)$$

$$p_i = \frac{e^{s_i/\tau}}{\sum_{j=1}^{N+M+B} e^{s_j/\tau}}, \quad i = 1, 2, \dots, N+M+B. \qquad (6)$$

Then the hierarchical structure could be described using the law of total probability, as the probability of $\mathbf{x}$ being a person is the sum over the probabilities of $\mathbf{x}$ being all the persons memorized by $\mathbf{W}'$:

$$p(\Lambda) = \sum_{j=1}^{N+M} p(\text{id} = j, \Lambda) = \sum_{i=1}^{N+M} p_i \qquad (7)$$

The probability of $\mathbf{x}$ being background could be formulated in the same manner:

$$p(\Phi) = \sum_{i=N+M+1}^{N+M+B} p_i \qquad (8)$$

The first level of HOIM loss for detection is formulated as a binary cross entropy loss:

$$\mathcal{L}_{\text{det}} = -y \log(p(\Lambda)) - (1 - y) \log(p(\Phi)) \qquad (9)$$

where y is a binary label indicating whether $\mathbf{x}$ is a person or not.

For the second level, we slice $\mathbf{s}'$ and ignore the background part. Thus the re-ID sub-loss shares the same formulation to OIM loss described in the last subsection. The calculation process is illustrated in the third row of Fig. 3. The formal composition of our proposed HOIM loss is the linear combination of the two-level losses:

$$\mathcal{L}_{\text{HOIM}} = \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{OIM}}, \quad \text{where} \quad \lambda = 2p(\Lambda)^2 \qquad (10)$$

where λ is the loss weight for $\mathcal{L}_{\text{OIM}}$. It is dynamically changed according to the detection confidence digit $p(\Lambda)$. Intuitively, if the detection confidence is high, more weight could be put on distinguishing the detected person; otherwise, the model should focus on the detection task.

By expanding the projection matrix of OIM for background memorizing, our HOIM loss aggregates pedestrian detection and re-ID into a unified layer. Embeddings learned by HOIM loss is not only able to identify different persons, but also disassociates background clutter from persons. Therefore, the embeddings are more robust for person matching, especially when the detected boxes have low quality, *i.e.* false alarms and misalignments. Moreover, adding background descriptors increases the feature diversity of the projection matrix, which alleviates the risk of over-fitting.

Selective Memory Refreshment for Embedding Buffer

In HOIM loss, the two circular queues in the projection matrix $\mathbf{W}'$, denoted as $\mathbf{Q}$ and $\tilde{\mathbf{Q}}$, are used to store unlabeled person and background-clutters respectively. During training, they are updated in a FIFO manner (Xiao et al.

2017b). However, it ignores the underlying manifold structure of all the embeddings in the current batch and the circular queues, *i.e.* similar embeddings would be pushed into the queue while discriminative memories located at the tail of the queue would be popped out, which in consequent increases feature redundancy and triviality.

To alleviate this issue, we propose to rank the embeddings by assigning each of them an importance factor ω. Embeddings in the current batch will be evaluated according to their ω, and would only replace the existing ones with low importance if pushed into the queue. Specifically, the design of ω should be able to describe three properties of the candidate embeddings:

1. **Hardness:** For an unlabeled person, it is considered *hard* to discriminate if it has high similarities to the embeddings in the look-up table. Hard samples are more favorable than *easy* ones since they could provide stronger magnitude of gradients for effective training.
2. **Diversity:** Embeddings in the queue are supposed to be different from each other to act as non-trivial negative samples. Thus instances with high similarities to the existing queue should be assigned with low importance.
3. **Mortality:** As the training iterations proceed, the old embeddings become antiquated accordingly and should be forgotten eventually. Hence the importance factors are ought to decay alongside every iteration.

Based on the intuitions above, the importance factor for an unlabeled person $\mathbf{x}$ is formulated as follows:

$$\omega = \frac{\max(\mathbf{Lx})}{\max(\mathbf{Qx}) + \epsilon} \cdot k^{\iota} \tag{11}$$

where $\mathbf{L}$ denotes the look-up table, k is a decay factor within the range of $(0, 1)$ and ι indicates the iteration number. ϵ is a small constant to ensure numerical stability. Similarly, we define the importance factor of a background clutter as:

$$\tilde{\omega} = \frac{\max([\mathbf{Lx}, \mathbf{Qx}])}{\max(\tilde{\mathbf{Q}}\mathbf{x}) + \epsilon} \cdot k^{\iota} \tag{12}$$

where $[\cdot, \cdot]$ represents the concatenation operation.

For an arbitrary embedding $\mathbf{x}$, either from an unlabeled person or a background clutter, its ω is calculated first. Then we find the memory slot with the smallest importance factor ω_{min} from the corresponding queue and compare it with the candidate importance factor. $\mathbf{x}$ would be pushed into the memory slot *only if* ω is larger than ω_{min}. Otherwise, it would be ignored. Compared to the FIFO strategy, our proposed method chooses negative samples more intelligently.

Add-on: Focal Loss

The SMR strategy introduced above only deals with unlabeled samples, *i.e.* persons without identity and background-clutters. The labeled persons are left without weighting and have a potential consequence of overwhelming the loss and gradients. To solve this issue, we propose to add Focal Loss (Lin et al. 2017) into the HOIM formulation, which down-weights easy examples to reduce their contributions to the total loss and focuses training on hard samples. Specifically, a density factor is multiplied to the log-likelihood within $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{OIM}}$:

$$\log(p) \leftarrow \alpha(1-p)^{\gamma} \log(p) \tag{13}$$

where α and γ are tun-able parameters.

Experiments

In this section, we first introduce the datasets and evaluation protocols, describe the implementation details, followed by ablation studies on the efficacy of each component and model inspections. Finally, we compare the performance of our model with state-of-the-art methods.

Datasets and Evaluation Protocol

CUHK-SYSU (Xiao et al. 2017b) is a large-scale dataset that contains 18,184 images and 96,143 person bounding boxes. The images are collected from hand-held cameras and various movie screenshots, with diverse appearances on illumination, viewpoint, gesture, background and occlusion. We follow the standard train/test split pre-defined by the dataset, where the training set consists of $11{,}206$ images and $5{,}532$ identities, whilst the testing set contains $2{,}900$ probe persons and $6{,}978$ gallery images. In addition, each probe image is assigned with several gallery subsets with different sizes. We use the default gallery size 100 if not specified.

PRW (Zheng et al. 2017) is sampled from videos shot by 6 stationary cameras on a university campus. The dataset holds $11{,}816$ frames with $43{,}110$ annotated bounding boxes, where $34{,}304$ of them are tagged with 932 identities. It splits $5{,}704$ images with 482 different IDs for training and $2{,}057$ probe persons w.r.t. $6{,}112$ gallery images for testing. Different from CUHK-SYSU, each probe image requires matching from the whole gallery set instead of pre-defined subsets.

Evaluation Protocol Mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC top-K) are used as the performance metrics. The mAP metric reflects both precision and recall at every position in the ranked sequence for each probe person. CMC top-K shows the percentage of all the rank sequences where correct matches appear at rank $\leq$ K. A candidate among the top-K results is considered correct if its IoU to the ground truth is larger than 0.5.

Implementation Details[1]

Our implementation is based on PyTorch (Paszke et al. 2017). The ResNet-50 backbone is initialized with ImageNet-pretrained weights. The first convolutional block is frozen during training. The appended embedding projection layers are followed by batch normalization (Ioffe and Szegedy 2015) layers.

The momentum η, softmax temperature τ and importance decay factor k of HOIM are set to 0.5, $1/30$ and 0.99 respectively. Sizes of the embedding buffers, *i.e.* N, M and B, are set individually for different datasets. For CUHK-SYSU, they are $5{,}532$, $5{,}000$ and $5{,}000$; for PRW, N is set to 482, while M and B are both reduced to 500 to balance the probability distribution.

We train our model jointly with a batch size of 5 on a single NVIDIA Tesla P40 GPU. The training data are resized to

[1] https://github.com/DeanChan/HOIM-PyTorch

Table 1: Ablation study results on CUHK-SYSU.

Method	mAP(%)	top-1(%)	Δ(%)
OIM	75.5	78.7	
OIM-base	83.6	87.4	
+ Focal Loss	85.1	87.6	(+1.5, +0.2)
+ SMR	85.5	88.2	(+0.4, +0.6)
HOIM	**89.7**	**90.8**	(+4.2, +2.6)

have at least 900 pixels on the short side and at most 1,500 on the long side. Crop and zero-padding are used (if necessary) to fit images with different resolutions into a batch. The target learning rate is set to 0.003, which is gradually warmed-up at the first epoch and decayed by a factor of 0.1 at the 16th epoch. The training process converges at epoch 22. During inference, it takes around 280 ms to detect the persons in an image and extract the corresponding embeddings simultaneously.

Ablation Study

We present analytical experiments on CUHK-SYSU with a gallery size of 100 to evaluate the efficacy of each component in the proposed model. We start from a baseline, which is refined from the successful OIM model (Xiao et al. 2017b) by adding batch normalization, large-batch-size training and learning rate warm-up. It shares the same network structure with our proposed model except that it separates detection and re-ID supervisions into two independent losses, *i.e.* R-CNN classification loss and OIM loss. Experiment results are recorded in Table 1, from which we can see that the baseline model, denoted as 'OIM-base', achieves 83.6% and 87.4% in mAP and top-1 accuracy respectively, providing a relatively good start point.

Next up, we add Focal Loss to both R-CNN classification loss and OIM loss. We follow the suggested configurations in (Lin et al. 2017) by setting the weighting factor α and focusing parameter γ to 0.25 and 2.0 respectively. Additionally, we find that increasing the α for R-CNN classification from 0.25 to 1.0 yields slightly better performance. Therefore, we stick to this parameter setting throughout the paper. Comparing to the baseline, adding focal loss brings a performance gain by 1.5 pp. and 0.2 pp. w.r.t. mAP and top-1 accuracy, showing that weighting samples according to their relative hardness is beneficial to the task of person search.

Furthermore, we replace the FIFO rule of the OIM circular queue with the SMR strategy. From Table 1 we find that the mAP and top-1 accuracy are slightly boosted by 0.4 and 0.6 pp. correspondingly. Hence we conclude that the proposed SMR policy is superior to the FIFO rule of the original OIM loss.

Finally, we put all the components together and present our HOIM model, which takes the place of both R-CNN classification and OIM loss. The mAP and top-1 accuracy are lifted by 4.2 and 2.6 pp. The exceptional performance substantially verifies our concept of modeling pedestrian detection and re-ID by their hierarchical relationship.

Table 2: Detection performance of a vanilla detector and its OIM/HOIM extensions

Method	RPN		Faster R-CNN	
	Recall(%)	AP(%)	Recall(%)	AP(%)
detector	89.27	69.07	93.12	87.02
OIM-base	-9.1	-21.69	-12.01	-11.18
HOIM	-0.73	-12.66	-1.36	-1.35

Model Inspection

To further understand why HOIM performs better than our OIM baseline, we provide analysis on the detection performance and embedding discriminability, which are the most fundamental factors for person search accuracy. The analysis is based on models from Ablation Study section and we have the following findings:

HOIM has better detection accuracy than OIM. We list the detection performance of a vanilla Faster R-CNN and its extensions with OIM/HOIM loss heads in Tab. 2. All three models are trained and tested under the same protocol. We can see that adding re-ID losses to the vanilla Faster R-CNN would harm the detection performance because of the contradictory objectives (Chen et al. 2018). However, adding HOIM *harms less* to the detector performance than the OIM loss.

HOIM embeddings are more discriminative under false detections. False detections commonly exist in the gallery, some are misaligned boxes and others are misclassified background patches. To be more specific, 1400/2900 probe-galleries contain at least one false detections generated by HOIM, and the portion of our OIM baseline is 1760/2900. When matching to a probe person, our method would rank down those false detections by producing lower similarities, yielding a better ranking list. An intuitive example is shown in Fig 4. When the bbox shifts from person to background, HOIM embedding of misaligned bbox (marked in green) has lower similarity to the target person (marked in red) compared to OIM, especially when the box has an IoU < 0.5, which would be labeled as a false alarm.

In addition, we make a visualized comparison between the embeddings learned by the baseline and our proposed HOIM model. Embeddings for both labeled and unlabeled persons are extracted directly from each row of the projection matrix. Background-clutters generated by HOIM also come from the projection matrix. As for the baseline, we randomly select 250 images from the dataset and crop 20 different background regions each image to generate the embeddings. Principle Component Analysis is adopted to project the 256-dimensional embeddings into a 2D subspace. The scatter diagrams for the dimension-reducted embeddings are gathered in Fig. 5, where the left column belongs to the OIM baseline, and the right column belongs to our proposed HOIM model. Generally, embeddings learned by HOIM are better at separating background and person, which is more robust when matching false detections.

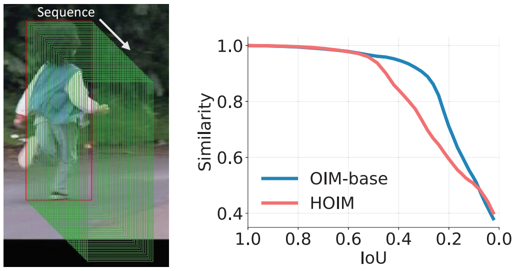

Figure 4: Detection sensitivity analysis. **(Left)** Sequence of boxes (marked in green) for calculating the similarities to the perfectly-aligned first box (marked in red). **(Right)** Similarity vs. IoU for the box sequence. Our method is better at assigning lower similarities to false detections (IoU< 0.5).

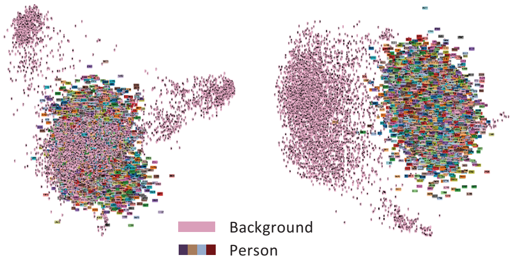

Figure 5: Embedding visualizations on CUHK-SYSU. The left/right column shows the distribution of background-clutters and persons encoded by our OIM baseline/HOIM model respectively.

Comparison with the State-of-the-Arts

Table 3 reports the person search results on CUHK-SYSU and PRW. All the candidate methods can be clustered into two styles, namely two-stage methods (the upper block) and one-stage ones (the lower block). 'CNN' in the table represents a Faster R-CNN detector. Our proposed HOIM model reaches a performance of 89.7% and 90.8% w.r.t. the mAP and top-1 metrics, which consistently outperforms all the counterparts. We also evaluate the performance consistency with varying gallery sizes, which is pre-defined in (Xiao et al. 2017b). We can observe from the left column of Fig. 6 that all methods undergo a performance degeneration as the gallery size extends. This indicates that person search becomes more challenging at larger search scales since more hard samples are included in the matching domain. Our method also ranks the best at all gallery sizes.

On PRW, our model achieves the best performance in both mAP and top-1 accuracy among one-stage and two-stage methods. It is also worth to notice that our model is more resource-friendly. Unlike two-stage methods MGTS (Chen et al. 2018), CLSA (Lan, Zhu, and Gong 2018) and one-stage methods QEEPS (Munjal et al. 2019), CTXGraph (Yan et al. 2019) which requires complex forward pass, our method forwards an image directly through one network, which reduces the model size and computational operations. A comparison on Floating Point Operations (FLOPs) is shown in the right column of Fig. 6. We can see that our proposed HOIM model excels both on speed and accuracy, which could be more substantial in practical use.

Table 3: Performance Comparison with State-of-the-arts.

Method	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
DPM + IDE w. CWS (Zheng et al. 2017)	-	-	20.5	48.3
CNN + MGTS (Chen et al. 2018)	83.0	83.7	32.6	72.1
CNN + CLSA (Lan, Zhu, and Gong 2018)	87.2	88.5	38.7	65.0
OIM (Xiao et al. 2017b)	75.5	78.7	21.3	49.9
IAN (Xiao et al. 2017a)	76.3	80.1	23.0	61.9
NPSM (Liu et al. 2017a)	77.9	81.2	24.2	53.1
RCAA (Chang et al. 2018)	79.3	81.3	-	-
CTXGraph (Yan et al. 2019)	84.1	86.5	33.4	73.6
QEEPS (Munjal et al. 2019)	88.9	89.1	37.1	76.7
Ours	**89.7**	**90.8**	**39.8**	**80.4**

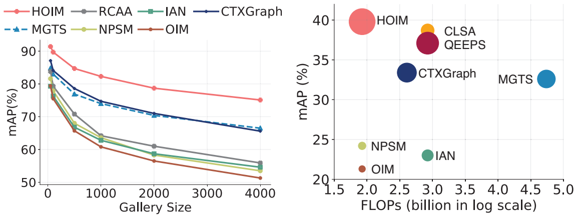

Figure 6: **(Left)** Performance comparison on CUHK-SYSU with varying gallery sizes. Dashed lines represent two-stage methods while solid lines denote one-stage models. **(Right)** mAP/computation trade-offs on PRW. Floating Point Operations (FLOPs) are estimated by their network backbones. Sizes of the circles reflect the relative top-1 magnitudes.

Conclusion

In this paper, we make the first attempt to model the inter-dependency between pedestrian detection and re-identification explicitly, and further propose a novel Hierarchical Online Instance Matching loss which exploits the hierarchical relationship of the two tasks to guide the end-to-end training. We also propose a Selective Memory Refreshment strategy to evaluate the contribution of the unlabeled persons for the classification loss. Extensive experiments on two standard person search benchmarks demonstrate the effectiveness of our method.

Acknowledgement

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by National Science Fund of China under Grant No. U1713208, National Natural Science Foundation of China under Grant No. 61702262, Funds for International Cooperation and Exchange of the National Natural Science Foundation of China under Grant No. 61861136011, Natural Science Foundation of Jiangsu Province, China under Grant No. BK20181299, Young Elite Scientists Sponsorship Program by CAST (2018QNRC001), Program for Changjiang Scholars, and "111" Program AH92005.

References

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR*.

Chang, X.; Huang, P.-Y.; Shen, Y.-D.; Liang, X.; Yang, Y.; and Hauptmann, A. G. 2018. Rcaa: Relational context-aware agents for person search. In *ECCV*.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Tai, Y. 2018. Person search via a mask-guided two-stream cnn model. In *ECCV*.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *CVPR*.

Ding, S.; Lin, L.; Wang, G.; and Chao, H. 2015. Deep feature learning with relative distance comparison for person re-identification. *PR* 48(10):2993–3003.

Dollar, P.; Tu, Z.; Perona, P.; and Belongie, S. 2009. Integral channel features. In *BMVC*.

Dollar, P.; Appel, R.; Belongie, S.; and Perona, P. 2014. Fast feature pyramids for object detection. *TPAMI* 36(8):1532–1545.

Fan, X.; Jiang, W.; Luo, H.; and Fei, M. 2018. Spherereid: Deep hypersphere manifold embedding for person re-identification. *arXiv preprint arXiv:1807.00537*.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*.

Felzenszwalb, P. F.; Girshick, R. B.; Mcallester, D.; and Ramanan, D. 2009. Object detection with discriminatively trained part based models. *TPAMI* 32(9):1627–1645.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Kostinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*.

Lan, X.; Zhu, X.; and Gong, S. 2018. Person search by multi-scale matching. In *ECCV*.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. DeepReID: Deep filter pairing neural network for person re-identification. In *CVPR*.

Li, X.; Zheng, W. S.; Wang, X.; Xiang, T.; and Gong, S. 2015. Multi-scale learning for low-resolution person re-identification. In *ICCV*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal loss for dense object detection. In *ICCV*.

Liu, H.; Feng, J.; Jie, Z.; Jayashree, K.; Zhao, B.; Qi, M.; Jiang, J.; and Yan, S. 2017a. Neural person search machines. In *ICCV*.

Liu, H.; Feng, J.; Qi, M.; Jiang, J.; and Yan, S. 2017b. End-to-end comparative attention networks for person re-identification. *TIP* 26(7):3492–3506.

Munjal, B.; Amin, S.; Tombari, F.; and Galasso, F. 2019. Query-guided end-to-end person search. In *CVPR*.

Ouyang, W., and Wang, X. 2012. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*.

Ouyang, W., and Wang, X. 2013. Joint deep learning for pedestrian detection. In *ICCV*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *TPAMI* 39(6):1137–1149.

Varior, R. R.; Shuai, B.; Lu, J.; Xu, D.; and Wang, G. 2016. A siamese long short-term memory architecture for human re-identification. In *ECCV*.

Wang, X.; Doretto, G.; Sebastian, T.; Rittscher, J.; and Tu, P. 2007. Shape and appearance context modeling. In *ICCV*.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*.

Xiang, W.; Huang, J.; Qi, X.; Hua, X.-S.; and Zhang, L. 2018. Homocentric hypersphere feature embedding for person re-identification. *arXiv preprint arXiv:1804.08866*.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*.

Xiao, J.; Xie, Y.; Tillo, T.; Huang, K.; Wei, Y.; and Feng, J. 2017a. Ian: The individual aggregation network for person search. *arXiv preprint arXiv:1705.05552*.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017b. Joint detection and identification feature learning for person search. In *CVPR*.

Xu, Y.; Ma, B.; Huang, R.; and Lin, L. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *ACM'MM*.

Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; and Ouyang, W. 2018. Attention-aware compositional network for person re-identification. In *CVPR*.

Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; and Yang, X. 2019. Learning context graph for person search. In *CVPR*.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *ICPR*.

Zhang, S.; Bauckhage, C.; and Cremers, A. B. 2014. Informed haar-like features improve pedestrian detection. In *CVPR*.

Zhang, S.; Benenson, R.; and Schiele, B. 2015. Filtered channel features for pedestrian detection. In *CVPR*.

Zhang, S.; Benenson, R.; and Schiele, B. 2017. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*.

Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; and Schiele, B. 2016. How far are we from solving pedestrian detection? In *CVPR*.

Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; and Schiele, B. 2018. Towards reaching human performance in pedestrian detection. *TPAMI* 40(4):973–986.

Zhang, L.; Xiang, T.; and Gong, S. 2016. Learning a discriminative null space for person re-identification. In *CVPR*.

Zhang, S.; Yang, J.; and Schiele, B. 2018. Occluded pedestrian detection through guided attention in cnns. In *CVPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2013. Unsupervised salience learning for person re-identification. In *CVPR*.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *ECCV*.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *CVPR*.