Hierarchical Distillation Learning for Scalable Person Search

Wei Li[1], Shaogang Gong[1], Xiatian Zhu[1,]

[a] *School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK*
[b] *Vision Semantics Limited, London E1 4NS, UK*

Abstract

Existing person search methods typically focus on improving person detection accuracy. This ignores the model inference efficiency, which however is fundamentally significant for real-world applications. In this work, we address this limitation by investigating the scalability problem of person search involving both model accuracy and inference efficiency simultaneously. Specifically, we formulate a Hierarchical Distillation Learning (HDL) approach. With HDL, we aim to comprehensively distil the knowledge of a strong teacher model with strong learning capability to a lightweight student model with weak learning capability. To facilitate the HDL process, we design a simple and powerful teacher model for joint learning of person detection and person re-identification matching in unconstrained scene images. Extensive experiments show the modelling advantages and cost-effectiveness superiority of HDL over the state-of-the-art person search methods on three large person search benchmarks: CUHK-SYSU, PRW, and DukeMTMC-PS.

Keywords: Person search, Person re-identification, person detection, Knowledge distillation, Scalability, Model inference efficiency.

*Corresponding author

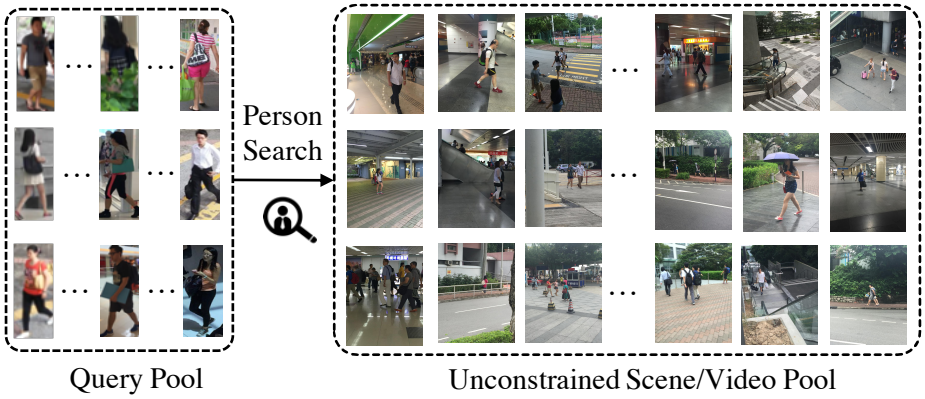

Figure 1: The significance of scalability in person search. Both sets of query persons and scene imagery are of large scale.

1. Introduction

Person search considers the problems of person detection and person re-identification (re-id) *simultaneously* [**?** **?**]. It is valid and necessary due to that the practical application of person re-id relies heavily on person detection. The detection quality of persons on the surveillance scene images affects the re-id performance largely. For example, missing detection causes the *inability* of person re-id on the corresponding person instance, and misalignment introduces *noise* or *information loss* to person re-id.

In addition to person matching accuracy, this task joining by person search also expands the scope for model efficiency considerations. Conventionally, person search efficiency is mostly considered in person re-id model design, since person bounding boxes are assumed already available. This breaks the connection between person re-id and person detection, therefore, losing their joint computing opportunity for improving model efficiency. This issue is naturally solved in the person search problem setting.

Model efficiency is fundamentally crucial for *scalable* person search, due to the intrinsic large scale search requirement in real-world deployment (Figure **??**). The efficiency problem was initially investigated in the introduction of

person search [**?**], followed by a few followup *joint learning* model designs [**?** **?** **?**]. However, all these existing methods are significantly outperformed by *independent learning* competitors [**?** **?**]. Moreover, some of the joint learning methods [**?** **?** **?**] are even *not necessarily* more efficient than independent learning, because of their query-specific search design nature. That is, the model needs to conduct an independent search process in *every* whole scene image for *every* query person, with the search cost proportional to the *quadratic pairwise combination* (i.e. multiplication) of the query and gallery samples. This implies potentially even *more inefficient* solutions than simpler independent learning [**?** **?**], totally opposite to their original efficiency objective.

In the literature, only the OIM method [**?**] makes an initial attempt for efficient person search. The key idea is that person detection and person re-id can share a large proportion of computing cost by jointly using the low-level feature network layers. This is analogous to the core idea of Faster R-CNN [**?**]. After the OIM model is trained, person detection and re-id feature extraction can be conducted jointly on the gallery data by a single network. It is a *one-off* process, independent to the size of query images therefore much more scalable than query-specific search models. However, the main focus of OIM is on how to exploit unlabelled person instances for improving re-id matching. This method does not fully investigate the significant model efficiency problem. This is partly due to that its performance is somewhat weak, e.g. significantly inferior to the current state-of-the-art methods [**?** **?**]. Overall, the *scalability* problem including both *model accuracy* and *inference efficiency* for person search remains largely under-studied, despite its significant practical importance.

In this work, we investigate the scalability problem for person search. We explore the potential of knowledge distillation [**?**] by developing a *Hierarchical Attention Learning* (HDL) method. The core idea behind the HDL is to transfer the person search knowledge of a heavy teacher model that can be optimised more discriminatively with stronger learning capability into a lightweight student model with weaker learning capability. Whilst knowledge distillation has been previously studied mostly in single label image classification [**?** **?** **?**

?], it has not been explored for the more complex person search problem with two different tasks involved. To this end, we design a novel approach for distilling comprehensive knowledge in the teacher network hierarchy including feature representation, attention, and prediction. To facilitate distillation, we further develop a strong joint learning teacher model for ensuring the knowledge quality which is lacking in the literature, and a structurally consistent and computationally efficient student model.

We make three **contributions** in this work: **(1)** We investigate for the first time the scalability problem involved in person search. This is a fundamentally significant problem to be solved for scaling up the deep learning solutions to person search in the real-world applications. **(2)** We formulate a *Hierarchical Distillation Learning* approach for more discriminating knowledge transfer from a stronger teacher model into an efficient student model. **(3)** We design a simple and effective teacher model for joint learning of person search, which largely facilitates the knowledge distillation by avoiding knowledge transfer between structure inconsistent teacher and student models. Extensive experiments show the model cost-effectiveness and performance advantages of our HDL over the state-of-the-art alternative approaches on three person search benchmarks: CUHK-SYSU [?], PRW [?], and DukeMTMC-PS [?].

2. Related Work

Person Re-Identification. Person re-identification (re-id) [? ? ? ?] is part of person search. Typically, re-id assumes the availability of person bounding boxes across the supervised [? ? ? ? ? ? ? ?], unsupervised [? ? ? ? ? ?], and domain adaptation [? ? ? ? ? ? ?] settings. This overlooks the opportunity for interacting person re-id and person detection. From the system deployment viewpoint, this is an incomplete problem design. Moreover, the existing re-id studies often ignore the correlation between person detection and re-id matching. For example, missing detection can cause a deemed failure of person re-id therefore affecting the final search result. Poor person detection

may negatively affect the re-id matching accuracy. Such considerations however are totally missing in the current person re-id benchmark datasets [**?** **?** **?**]. The recent introduction of person search benchmarks [**?** **?**] aims to solve these issues by jointly considering person detection along with re-id matching in a single problem setting.

Person Search. Due to a more comprehensive problem formulation, person search has gained increasingly more attention and research efforts [**?** **?** **?** **?** **?**] since its establishment [**?** **?** **?**]. Existing methods are generally fallen into two groups: (1) *independent learning* (IL) [**?** **?**] and (2) *joint learning* (JL) [**?** **?** **?** **?** **?** **?**] based models.

Thus far, the *independent learning* based person search methods achieve the state-of-the-art performance [**?** **?** **?**]. They separate person detection and re-id matching by designing independent network models. Strong and computationally expensive CNN models [**?**] are often selected in such designs for maximising the search accuracy. One of the major disadvantages for these methods is costly deployment and slow execution. The model inference efficiency can be further reduced due to the addition of auxiliary components such as foreground segmentation and multi-branch fusion [**?**]. Although reaching good performance, this group of methods are less scalable computationally therefore unsuitable for large scale deployments typically required in real applications.

The *joint learning* based person search methods have been developed with one of the main objectives as solving the above efficiency limitation [**?** **?** **?** **?** **?** **?**]. The methods in [**?** **?**] improve the model inference speed by taking advantages of the Faster R-CNN design. The key idea is to make person detection and re-id tasks share the low-level feature computation. NPSM [**?**], RCAA [**?**] and QEEPS [**?**] suggest query-specific person search strategies. CGPS [**?**] learns contextual graph representations via coupling the targets and the background contexts. Opposite to their design objectives for efficiency gain, these existing models *all* suffer from another scalability limitation: every query-gallery pair needs to be processed independently. This means that the

detection cost is proportional to the combination of query and gallery samples. Instead, person detection on all gallery images is conducted *one-off* in [**?** **?**], therefore independent and scalable to any sizes of query tasks. The efficiency of NPSM [**?**] is also significantly limited by the need of generating region proposals, e.g. EdgeBox [**?**]. Besides, all these models are often less powerful than the IL counterparts.

In contrast to all the existing methods, we consider the scalability and cost-effectiveness problem of person search including both model accuracy and inference efficiency. None of the previous methods are designed to address this problem, lacking sufficient model generalisation and/or inference efficiency. To this end, we explore the idea of knowledge distillation [**?**]. We also develop a simple and strong joint learning model that reaches the performance of the state-of-the-art independent learning method. This layouts a very competitive baseline method and inspires novel ideas to the future works.

3. Hierarchical Distillation Learning

For model training, we often collect m training scene images $\mathcal{I} = \{\boldsymbol{I}_i\}_{i=1}^{m}$ captured from multiple camera views. The annotation includes person bounding boxes $\mathcal{Y}_{box}$ and identity labels $\mathcal{Y}_{id} = \{y_i\}_{i=1}^{n}$ on a total of n_{id} training people, i.e. $y_i \in \{1, \cdots, n_{\text{id}}\}$. A single unconstrained scene image may contain multiple (varying) person instances. The objective is to learn an efficient person search model for simultaneous person detection and re-id matching. To this end, we formulate a ***Hierarchical Distillation Learning*** (HDL) approach featured with comprehensive knowledge distillation and joint learning of person detection and person re-id (i.e. person search) in unconstrained surveillance scene imagery data. An architectural overview of the proposed HDL method is depicted in Figure **??**.

3.1. HDL Overview

In design, the proposed HDL model takes the advantages of knowledge distillation [**?**]. Specifically, HDL consists of three components: (1) A *teacher* model

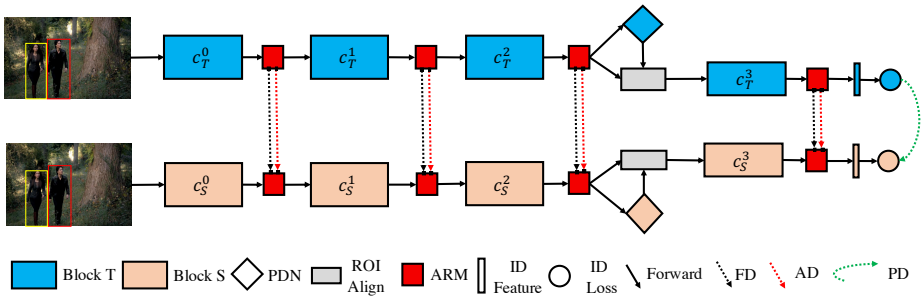

Figure 2: Overview of the proposed Hierarchical Distillation Learning (HDL) approach. The HDL process consists of two steps: (1) We first train the heavy teacher model (Sec **??**). See the details in Sec **??**. (2) We then train the lightweight student model (Sec **??**) by knowledge distillation from the teacher model. In test, we deploy the efficient student model for scalable person search. The symbols c_T^j and c_S^j ($j \in \{0, 1, 2, 3\}$) denote channel dimensions in the corresponding j-th block of the teacher and student models. The first layers and standard detection loss functions in both teacher and student models are omitted for simplicity. **T**: Teacher; **S**: Student; **PDN**: Person Detection Network; **ARM**: Attention Residual Module; **FD**: Feature Distillation; **AD**: Attention Distillation; **PD**: Prediction Distillation.

with a large size and great learning capability, designed to realise a strong person search network (Section **??**). (2) A *student* model with a small size and inferior learning capability, developed for superior inference efficiency in deployment (Section **??**). (3) A *hierarchical distillation* learning strategy, formulated for comprehensive knowledge transfer from the stronger teacher model to the student model (Section **??**). This addresses the hard-to-learn problem in training the small student model.

By deploying the student network as the final model in test time, we are able to achieve both superior model generalisation capacity and model inference speed, i.e. higher cost-effectiveness during deployment.

3.2. A Strong Joint Learning Teacher Model

By the means of knowledge distillation, the performance of the final (student) model relies heavily on the strength of the teacher model. That is, weaker teacher, weaker student. It is therefore critical and necessary to formulate a strong teacher model. To ease the distillation of person search knowledge, it is

also desired that the teacher model can share a similar structure of the student model *functionally* with the ability to jointly conduct both person detection and re-id matching. This avoids distilling the knowledge from two separate teacher networks (one for person detection, one for person re-id) to a single joint leaning student network, which is much more difficult.

Nonetheless, the only existing joint learning teacher model, OIM [**?**], is significantly inferior to the two-stage followup models [**? ?**]. Therefore, using the OIM model as the teacher model will lead to a similarly weak student model. To address this issue, we formulate a *stronger yet simpler* joint learning teacher model.

Teacher Model Architecture. Our teacher model is based on the design idea of Faster R-CNN [**?**] with person search specific modification. Specifically, it consists of three parts: (i) feature subnet, (ii) person detection subnet, and (iii) person re-id subnet. For design flexibility, any standard deep convolutional networks [**? ?**] can be used as the stem network. In the follows, we detail the three parts.

(I) Feature Subnet. To build the feature subnet, we use the lower part of the stem network starting from the first layer to the intermediate layer with $\frac{1}{r}$ down-sampling ratio. This subnet takes as input the scene image $\boldsymbol{I} \in \mathcal{R}^{H \times W \times 3}$ (H and W as image height and width), and outputs the image-level features $\boldsymbol{X}_f \in \mathcal{R}^{\frac{H}{r} \times \frac{W}{r} \times c_f}$ (c_f feature channels). The output features are for both person detection and re-identification tasks simultaneously. This model structure sharing reduces the overall computational costs with only a single unified forward pass needed.

(II) Person Detection Subnet. We subsequently build a person detection subnet (e.g. region proposal net) on top of the output features for detecting candidates in a given scene image. The details are as follows. With a $512 \times 3 \times 3$ conv layer, we first make the features discriminating for person appearance. The followed is the anchor layer for per-feature-location person detection. To make it more effective for person class specifically rather than generic object classes,

we use eleven different anchor-box scales and only one aspect ratio ($\frac{h}{w} = 2.44$) as [**?**]. Finally, we remove the redundant detections by applying a Non-Maximum Suppression process. In training, the person detection is optimised jointly by a softmax cross-entropy classification loss and a spatial location regression loss.

(III) Person Re-Id Subnet. We utilise the rest layers of the stem network to build person re-id subnet. It is based on the outputs of both feature and detection subnets. Specifically, we first use RoIAlign [**?**] at a spatial scale of 7×7 to crop the detection regions from the output of the feature subnet. This yields the detection-level features $\boldsymbol{X}_p \in \mathcal{R}^{7\times7\times c_f}$. $\boldsymbol{X}_p$ is first processed by batch normalisation, then used as the input of the person re-id subnet to produce the identity discriminative features $\boldsymbol{X}_p^{'} \in \mathcal{R}^{3\times3\times c_p}$, where c_p is the feature dimensions of the last layer in the stem network. To obtain the re-id feature $\boldsymbol{x}_p \in \mathcal{R}^{c_p}$, we globally pool $\boldsymbol{X}_p^{'}$ followed by batch normalisation. In training, we introduce a softmax cross-entropy identity classification loss function for re-id discriminative learning defined as:

$$\mathcal{L}_{\text{ID}} = -\frac{1}{N_{\text{p}}} \sum_{i=1}^{N_{\text{p}}} \log(\bar{\boldsymbol{p}}^i), \tag{1}$$

where N_{p} specifies the number of persons detected in the current mini-batch training data. $\bar{\boldsymbol{p}}^i$ is the posterior probability of the i-th training person instance on the ground-truth identity class. Specifically, it is written as:

$$\bar{\boldsymbol{p}}^i = \frac{\exp\left(\boldsymbol{p}^i\right)}{\sum_{i\in\mathcal{Y}_{id}} \exp\left(\boldsymbol{p}^i\right)}, \tag{2}$$

where $\boldsymbol{p}^i$ is the identity class logits predicted by the identity classification layer.

In person search on unconstrained scene images, person detection is often imperfect with inevitable false alarms and misalignment [**?**]. To mitigate this issue, we further impose a detection refinement loss same as the person detection subnet, in conjugate with the above re-id loss function. This refines the person localisation and suppresses the wrong detections.

Remarks. In this study, we aim for a simple but powerful teacher model. This is in contrast to most existing models that often become more complex

making the model analysis and comparison increasingly difficult. For instance, comparison between different models is mostly at the system level therefore less informative. By this simple teacher model we attempt to discourage this trend and answer a question that *how well a simple person search method can perform in a proper design*, which is *unfortunately* lacking in the literature. Interestingly, the proposed teacher model is surprisingly effective although being simple. In comparison, our method has a couple of significant merits: (1) More training friendly; (2) Potentially inspire new research ideas for developing novel joint learning person search models.

3.3. An Efficient Joint Learning Student Model

One major weakness of using the standard CNN architecture in the teacher model (e.g. ResNet-50) is the high cost of model inference cost. Whilst facilitating to learn the discriminating feature representations, this is not desirable for large scale deployments. There is hence a need for developing a computationally more efficient student model.

To that end, we design a lightweight building block based on depth-wise separable convolutions, inspired by efficient CNN models such as MobileNets [**? ?**]. The details of the student's building block are shown in Fig **??**. To build the entire student network, we just simply replace all levels of blocks of the teacher network by the proposed efficient blocks. This means that the student model adopts the teacher's overall structure.

Remarks. An important advantage of such a design is that, the teacher and student models are *structurally* consistent. This brings significant convenience for knowledge distillation, as described in the follows.

3.4. Hierarchical Distillation Learning

Smaller networks are typically inferior for discriminating training. To facilitate the learning of our student model, we propose a hierarchical distillation learning (HDL) strategy that can transfer comprehensively the teacher's knowledge for helping the student's training.

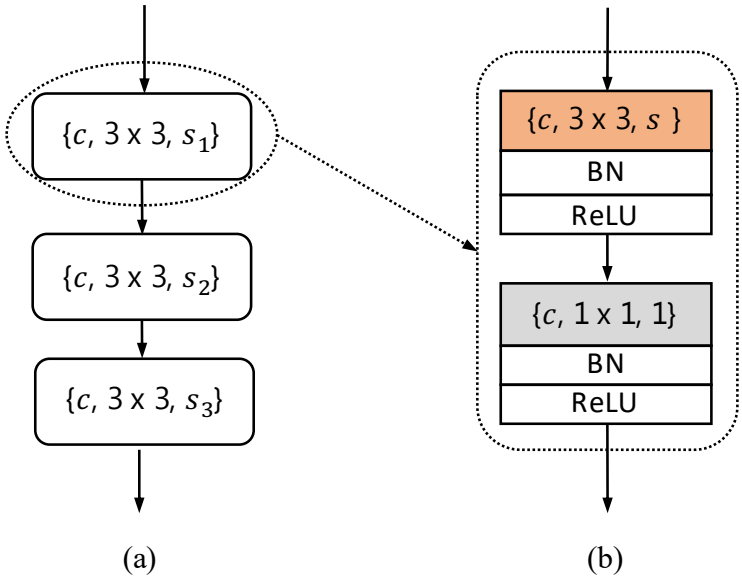

Figure 3: **(a)** A student's building block contains three modules. Each module **(b)** in such a block consists of two conv layers. Layer type is indicated by background colour: **grey** for *normal conv*, and **orange** for *depthwise separable conv* layers. The three items in the bracket of a conv layer are: filter number, filter shape, and stride. BN: Batch Normalisation. ReLU: Rectified Linear Unit.

Specifically, our HDL method considers three levels of knowledge during distillation: feature, attention, and prediction. For enabling attention distillation, we need an attention learning mechanism for both the teacher and student models. In order to learn and transfer richer attention knowledge distributed across different layers, we consider a *module-wise* attention design. That is, multiple selected building blocks can be attended in a pyramid structure (Fig **??**). As a side benefit, this may also assist the feature representation learning of both models concurrently.

Attention Residual Module. Formally, the input to an attention module is a 3-D tensor $\boldsymbol{X}^j \in \mathcal{R}^{h \times w \times c}$ where h, w, and c denote the height, width, and channel dimensions, respectively; And j indicates the block level of this module in the entire network. The essence of attention learning is to estimate a salience weight

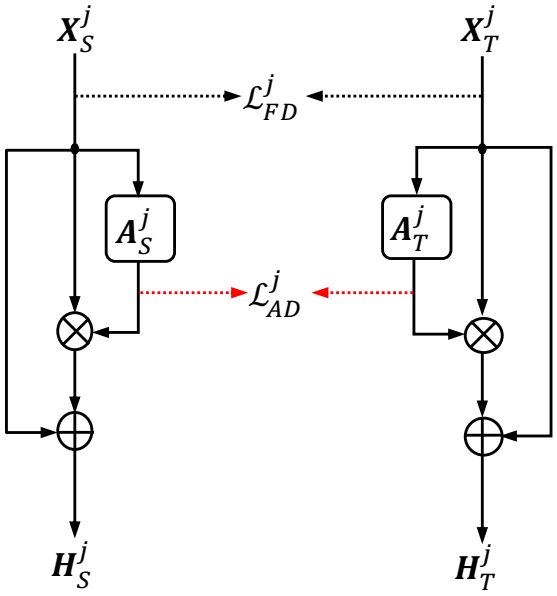

Figure 4: Attention residual module in knowledge distillation.

map $\boldsymbol{A}^j \in \mathcal{R}^{h\times w\times c}$ of the same size as $\boldsymbol{X}^j$. In this work, we adopt the Attention Residual Module (ARM) design [**?**] due to its superior learning capability. It is formulated (see Fig **??**) as:

$$\boldsymbol{H}^j = (1 + \boldsymbol{A}^j) * \boldsymbol{X}^j, \tag{3}$$

where $\boldsymbol{H}^j \in \mathcal{R}^{h\times w\times c}$ and $\boldsymbol{X}^j \in \mathcal{R}^{h\times w\times c}$ represent the modulated and original features, respectively. To further improve cost-effectiveness, we separate the spatial and channel attention learning as [**?** **?**].

(I) Feature Distillation. Feature distillation encourages the student to imitate the teacher's representation knowledge. Formally, we denote $\boldsymbol{X}_{S/T}^j$ as the feature maps at the j-th block level of the teacher ($\boldsymbol{X}_T^j$) or student ($\boldsymbol{X}_S^j$) network. For efficiency gain, the student network often has fewer feature channels. As a result, $\boldsymbol{X}_S^j$ and $\boldsymbol{X}_T^j$ are not aligned in channel dimension, which disables channel-to-channel distillation. To address this issue, we consider a 2-D spatial collective distillation scheme by discarding the channel dimension. Specifically,

we first accumulate the feature tensor along the channel dimension as:

$$f(\boldsymbol{X}_{S/T}^{j}) = \sum_{i} |\boldsymbol{X}_{S/T}^{j}(\cdot,\cdot,i)|^2, \tag{4}$$

where $\boldsymbol{X}_{S/T}^{j}(\cdot,\cdot,i)$ is the i-th feature channel of $\boldsymbol{X}_{S/T}^{j}$. We then obtain feature vectors by vectorisation:

$$\boldsymbol{x}_{S/T}^{j} = vec(f(\boldsymbol{X}_{S/T}^{j}))$$

We finally design the *feature distillation loss* as:

$$\mathcal{L}_{FD}(\boldsymbol{\Theta}_S) = \frac{1}{2}\sum_{j\in\mathcal{J}} \| \frac{\boldsymbol{x}_S^j}{\|\boldsymbol{x}_S^j\|_2} - \frac{\boldsymbol{x}_T^j}{\|\boldsymbol{x}_T^j\|_2} \|_2 \tag{5}$$

where $\boldsymbol{\Theta}_S$ denotes the parameters of the student model, and $\mathcal{J}$ the set of all block levels involved.

(II) Attention Distillation. Attention distillation aims for salience knowledge transfer. Specifically, we have the 3-D attention maps $\boldsymbol{A}_S^j$ and $\boldsymbol{A}_T^j$ from the student and teacher models at the j-th level. Similar to feature distillation, we first perform a channel-dimensional accumulation and vectorisation by computing $\boldsymbol{a}_{S/T}^{j} = vec(f(\boldsymbol{A}_{S/T}^{j}))$, then formulate the *attention distillation loss* as:

$$\mathcal{L}_{AD}(\boldsymbol{\Theta}_S) = \frac{1}{2}\sum_{j\in\mathcal{J}} \| \frac{\boldsymbol{a}_S^j}{\|\boldsymbol{a}_S^j\|_2} - \frac{\boldsymbol{a}_T^j}{\|\boldsymbol{a}_T^j\|_2} \|_2, \tag{6}$$

This essentially constrains the student model to mimic the attending behaviour optimised by the teacher model.

(III) Prediction Distillation. By prediction distillation, the student model attempts to simulate the high-level classification actions of the teacher model. Since the class space is the same for both models, their predictions are structurally consistent therefore allowing element-wise alignment. Formally, we design the *prediction distillation loss* as:

$$\mathcal{L}_{PD}(\boldsymbol{\Theta}_S) = t^2 \sum_{i\in\mathcal{Y}_{id}} \tilde{\boldsymbol{p}}_S^i \log \frac{\tilde{\boldsymbol{p}}_S^i}{\tilde{\boldsymbol{p}}_T^i} \tag{7}$$

which minimises the Kullback-Leibler divergence between the softened per-identity predictions $\tilde{\boldsymbol{p}}_S^i$ (by student) and $\tilde{\boldsymbol{p}}_T^i$ (by teacher). The temperature

parameter t controls the softening degree as:

$$\tilde{\boldsymbol{p}}^i_{S/T} = \frac{\exp\left(\boldsymbol{p}^i_{S/T}/t\right)}{\sum_{i\in\mathcal{Y}_{id}} \exp\left(\boldsymbol{p}^i_{S/T}/t\right)} \tag{8}$$

where $\boldsymbol{p}^i_{S/T}$ is the identity class logits predicted by the student or teacher model. As the gradient magnitudes produced by the soft targets $\tilde{\boldsymbol{p}}^i_{S/T}$ are scaled by $\frac{1}{t^2}$, we multiply this loss term by a factor t^2. This is to ensure that the relative contributions of the ground-truth and teacher probability distributions remain approximately unchanged.

Remarks. The proposed HDL method is based on existing distillation techniques that have been explored in varying context and problems [**?** **?** **?** **?**]. However, they are rarely jointly modelled in a unified model. Therefore, their complementary effects remain largely unknown. Moreover, the efficiency issue in person search is under-studied significantly, let alone exploiting the knowledge distillation notion. One main reason is that existing joint learning person search models [**?**] are dramatically inferior, therefore lacking a strong teacher model to enable the knowledge distillation. We overcome this obstacle to person search and further explore the potential of three fundamental distillation algorithms jointly for addressing the ignored and realistically significant scalability issue.

3.5. Model Training

As the conventional knowledge distillation, we start with training the teacher model, followed by student training using the proposed HDL algorithm.

Teacher Model. By joint learning person search, the loss function for the teacher network $\boldsymbol{\Theta}_T$ is formulated as:

$$\mathcal{L}(\boldsymbol{\Theta}_T) = \mathcal{L}_{ID}(\boldsymbol{\Theta}_T) + \mathcal{L}_{DET}(\boldsymbol{\Theta}_T), \tag{9}$$

where $\mathcal{L}_{ID}()$ is the cross-entropy loss for person identity classification, and $\mathcal{L}_{DET}()$ the person detection loss including box regression and binary-class classification.

Student Model. To train the student model, we also exploit the HDL loss functions in addition to the joint learning person search loss that same as Eq

(**??**). This aims to transfer the already-trained teacher's knowledge. Formally, the loss function of the student model $\mathbf{\Theta}_S$ is designed as:

$$\begin{aligned}\mathcal{L}(\mathbf{\Theta}_S) =&(1-\lambda_0)*\mathcal{L}_{ID}(\mathbf{\Theta}_S)+\lambda_0*\mathcal{L}_{PD}(\mathbf{\Theta}_S)+\\ &\lambda_1*\mathcal{L}_{AD}(\mathbf{\Theta}_S)+\lambda_2*\mathcal{L}_{FD}(\mathbf{\Theta}_S)+\\ &\mathcal{L}_{DET}(\mathbf{\Theta}_S),\end{aligned} \tag{10}$$

where $\lambda_{0/1/2}$ are three loss weighing hyper-parameters, estimated by cross-validation.

3.6. Network Architecture Details

In this section , we provide the details of HDL network architecture.

Teacher Model. We adopt a ResNet50 [**?**] as the stem network for the teacher model. It consists of four blocks (named conv2_x to conv5_x) each containing 3, 4, 6, 3 residual units. In particular, we choose the first layer (conv1_x, i.e. $64 \times 7 \times 7$ conv layer) to the third block (conv4_x) as feature sub-network, and conv5_x as person re-id sub-network. The person detection sub-network is built on conv4_x. The channel dimensions for the four blocks (Fig **??**) are $c_T^0 = 256$, $c_T^1 = 512$, $c_T^2 = 1,024$, and $c_T^3 = 2,048$, respectively.

Student Model. For the student model, we use a $32 \times 3 \times 3$ conv layer with stride 2 as the input layer. To achieve a good balance between efficiency and accuracy, we construct the corresponding four blocks by setting $c_S^0 = 128$, $c_S^1 = 256$, $c_S^2 = 384$, and $c_S^3 = 512$. In each building block (Figure **??**), we set the strides as $s_1 = s_2 = 1$ and $s_3 = 2$.

Attention Module. For both teacher and student models, we introduce a ARM unit at the end of each block (Fig **??**). This forms an attention pyramid for richer salience learning.

4. Experiments

Datasets. To evaluate the proposed HDL model, we used three person search benchmarks: CUHK-SYSU [**?**], PRW [**?**], and DukeMTMC-PS which is newly

Figure 5: Example query and unconstrained scene images from (a) CUHK-SYSU [**?**], (b) PRW [**?**], and (c) DukeMTMC-PS [**?**].

Table 1: Data statistics of person search datasets.

Dataset	IDs	Images	ID Split		Image Split	
			Train	Test	Train	Test
CUHK-SYSU	8,432	18,184	5,532	2,900	11,206	6,978
PRW	932	11,816	482	450	5,704	6,112
DukeMTMC-PS	1,404	35,543	702	702	16,362	17,350

introduced based on the DukeMTMC tracking dataset [**?**]. Example images are shown in Fig **??**. We adopted the standard evaluation setting of *CUHK-SYSU* and *PRW* (Table **??**). We re-purposed the DukeMTMC data into a person search benchmark *DukeMTMC-PS*. The train/test ID split follows the person re-id counterpart [**?**]. This dataset provides much more training and test scene images than CUHK-SYSU and PRW, representing a more realistic and more challenging person search scenario. We will publicly release the DukeMTMC-PS dataset.

Performance Metrics. For person detection, a bounding box was considered as correct if the overlap with the ground truth is over 50% [**?** **?**]. For person re-id, we used the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP). To evaluate the model inference efficiency, we adopted the common measurement of floating point operations (FLOPs) consumed by processing one typical scene image and one person bounding box.

Competitors. For model performance comparisons, we considered six state-of-the-art deep learning person search methods, including four *joint learning* model (OIM [**?**], RCAA [**?**], IAN [**?**], NPSM [**?**]) and two *independent learning* models (MGTS [**?**], CLSA [**?**]). We did not include other significantly inferior hand-crafted feature based alternative approaches in terms of

both model performance and inference efficiency.

Implementation Details. We conducted the experiments in the PyTorch framework. For model training, we adopted the SGD algorithm with the momentum set to 0.9, the weight decay to 0.0005. We set batch size to 8 for CUHK-SYSU with input size of 800×800 and 4 for PRW and DukeMTMC-PS with input size of 1920×1080. Mean value padding was used for organising images into batches. For teacher model training, we set the epoch number to 60 and initialised the learning rate at 0.005, with a decay factor of 10 at 50-th epoch. For student model training, we set the epoch number to 150 and initialised the learning rate at 0.005, with a decay factor of 5 every 50 epochs. We set the weights $\lambda_0 = 0.9$, $\lambda_1 = 2 \times 10^4$, $\lambda_2 = 2 \times 10^3$ (Eq (**??**)), and the temperature $t = 4$ (Eq (**??**)) by cross-validation for all the experiments. The L_2 normalisation was applied before computing the pairwise cosine similarity for re-id matching.

4.1. Comparisons to State-of-the-Art Methods

Evaluation on CUHK-SYSU. We reported the person search performance on CUHK-SYSU with the standard gallery size of 100 scene images in Table **??**. We made the following observations: **(1)** Our teacher model HDL(T) achieves the second best rank-1 rate and mAP among all competitors. In particular, the margin of HDL(T) over all existing joint learning competitors are consistently significant. This suggests that the joint learning strategy is *not necessarily* inferior to independent learning, even without adopting sophisticated techniques like attention inference [**?**] and reinforcement learning [**?**]. **(2)** By the proposed distillation method, our student model HDL(S) can achieve very competitive performance, e.g. matching the state-of-the-art CGPS [**?**] and surpassing all other existing joint learning methods and one independent learning model MGTS [**?**]. This indicates the efficacy of the proposed distilling method in transferring the teacher's knowledge. **(3)** The proposed HDL(S) reaches the best model inference efficiency, i.e. the superior cost-effectiveness benefits over all the alternative solutions. Note, we do not evaluate the model inference cost

for NPSM [**?**], RCAA [**?**] and QEEPS [**?**] due to their query-specific search design, a less scalable strategy than query-independent search by all the other methods. **(4)** HDL(S) is over one order of magnitude more efficient than all existing methods, which facilitates large scale and cost-effective deployments.

Table 2: Performance evaluation on **CUHK-SYSU**. The gallery size is 100. IL: Independent Learning; JL: Joint Learning; T: Teacher; S: Student; R101: ResNet-101; G: GFLOPs (1×10^9); M: MFLOPs (1×10^6).

Type	Metric (%)	Rank-1	mAP	Cost (scene/person)
IL	MGTS [**?**]	83.7	83.0	>1725.6G/52.8G
	CLSA [**?**]	**88.5**	**87.2**	>410.7G/26.4G
JL	OIM [**?**]	78.7	75.5	410.7G/2.0G
	IAN(R101) [**?**]	80.5	77.2	1146.2G/2.0G
	NPSM [**?**]	81.2	77.9	-
	RCAA [**?**]	81.3	79.3	-
	QEEPS [**?**]	84.4	84.4	-
	CGPS [**?**]	86.5	84.1	410.7G/2.0G
	HDL(T)	87.3	86.0	427.5G/2.1G
	HDL(S)	86.2	84.6	**37.5G/76.4M**

We further tested the model performance with the full gallery size at 6,978. This allows to evaluate larger scale search performance. Following the previous works, we compared the mAP results. Figure **??** shows similar observations as in Table **??**, suggesting that the model performance advantages of HDL generalise to large scale search.

Evaluation on PRW. We compared the model performance on the PRW benchmark. Overall, we obtained similar comparison observations that our teacher model HDL(T) achieves the second best performance in both rank-1 and mAP rates. HDL(S) similarly approaches the accuracy levels of HDL(T) whilst significantly outperforming all existing joint learning competitors in addition to a great model efficiency advantage. This consistently indicates the cost-

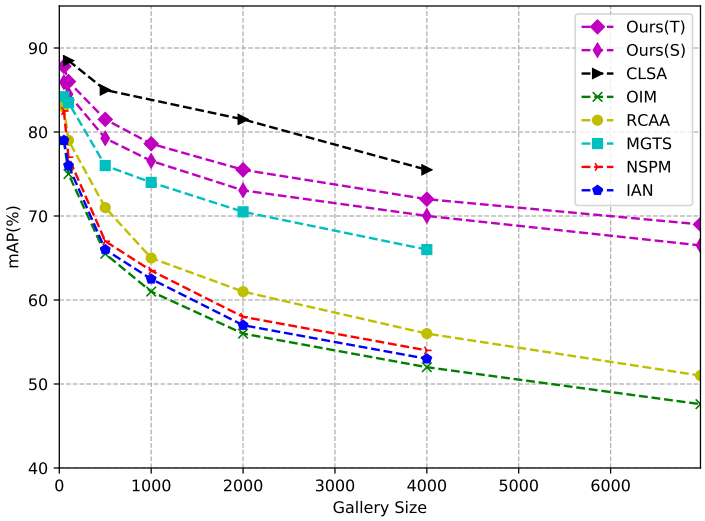

Figure 6: Test mAP of varying gallery sizes on CUHK-SYSU.

effectiveness and scalability superiority of our model over the existing person search methods in a more challenging application scenario.

Evaluation on DukeMTMC-PS. We further evaluated the performance of our HDL model on the newly introduced DukeMTMC-PS benchmark. Compared to CUHK-SYSU and PRW, test scene images from this benchmark are more than two times larger, therefore presenting a more challenging person search task. We compared with the only scalable joint learning competitor OIM and an independent learning baseline using Faster R-CNN+ResNet-50. The results in Table **??** show the consistent performance and efficiency superiority of HDL and the knowledge distillation efficacy from the stronger teacher model to the lightweight student model. Encouragingly, HDL(S) even surpasses the independent learning model, Faster R-CNN+ResNet50, by 2.9% (71.8-68.9) in Rank-1 and 2.8% (45.5-42.7) in mAP, in addition to more than one order of magnitude inference efficiency advantage.

Table 3: Performance evaluation on **PRW**. IL: Independent Learning; JL: Joint Learning; T: Teacher; S: Student; R101: ResNet-101; G: GFLOPs (1×10^9); M: MFLOPs (1×10^6).

Type	Metric (%)	Rank-1	mAP	Cost (scene/person)
IL	MGTS [**?**]	**72.1**	32.6	>1725.6G/52.8G
	CLSA [**?**]	65.0	**38.7**	>1330.7G/26.4G
JL	OIM [**?**]	49.9	21.3	1330.7G/2.0G
	IAN(R101) [**?**]	61.9	23.0	3713.7G/2.0G
	NPSM [**?**]	53.1	24.2	-
	HDL(T)	69.2	33.6	1381.6G/2.1G
	HDL(S)	64.4	28.2	**121.4G/76.4M**

Table 4: Performance evaluation on **DukeMTMC-PS**. IL: Independent Learning; JL: Joint Learning; T: Teacher; S: Student; FRCNN+R50: Faster R-CNN + ResNet-50; G: GFLOPs (1×10^9); M: MFLOPs (1×10^6).

Type	Metric (%)	Rank-1	mAP	Cost (scene/person)
IL	FRCNN+R50	68.9	42.7	>1330.7/26.4G
JL	OIM [**?**]	50.5	34.5	1330.7G/2.0G
	HDL(T)	**74.3**	**50.0**	1381.6G/2.1G
	HDL(S)	71.8	45.5	**121.4G/76.4M**

4.2. Further Analysis and Discussions

Attention Learning. We evaluated the benefits of our attention learning design. It is evident from Table **??** that, both the teacher and student models benefit significantly. In particular, our attention learning not only improves the quality of teacher's knowledge, but also facilitates the knowledge transfer process given that the student acquires more gains in most cases. This verifies our design consideration of integrating attention with feature and prediction in HDL.

Knowledge Distillation. We examined the effect of different distillation and their combinations on CUHK-SYSU. Table **??** reveals a couple of observations: (1) Each distillation *alone* brings about model improvements, with prediction

Table 5: Evaluating attention (A) learning. T: Teacher; S: Student. Setting: The gallery size for CUHK-SYSU is 6,978.

Dataset	CUHK		PRW		Duke	
Metric (%)	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
T(w/o A)	68.5	63.8	62.1	26.3	69.9	44.3
T(w/ A)	**73.2**	**69.7**	**69.2**	**33.6**	**74.3**	**50.0**
Gain	+4.7	+5.9	+7.1	+7.3	+4.4	+5.7
S(w/o A)	42.6	38.6	50.6	16.8	56.5	26.2
S(w/ A)	**49.5**	**45.1**	**59.1**	**22.8**	**61.4**	**33.4**
Gain	+6.9	+6.5	+8.5	+6.0	+4.9	+7.2

distillation contributing the most. This is because as the model output the prediction encodes the most discriminative abstraction information. (2) As the low-level knowledge, transferring attention and feature further enhances model learning on top of high-level prediction distillation. This verifies the complementary benefits of exploiting different model knowledge in HDL design.

5. Conclusion

In this work, we present a novel *Hierarchical Distillation Learning* (HDL) method for person search in unconstrained surveillance scene images. This method is designed particularly for addressing the largely ignored *scalability* problem in person search. It is in contrast to existing alternative methods that typically focus on model performance improvement alone. Specifically, we formulate a comprehensive knowledge distillation method for transferring feature representation, attention map, and class prediction from a strong and heavy teacher model to a weak and lightweight student model. This addresses the hard-to-optimise challenge for small models. We also contribute a simple and powerful joint learning teacher model, potentially motivating the further development of new models of its kind. Extensive comparative evaluations have been conducted on three large person search benchmarks. The results validate

Table 6: Evaluating different distillation. T: Teacher; S: Student. FD: Feature Distillation; AD: Attention Distillation; PD: Prediction Distillation. Setting: The gallery size for CUHK-SYSU is 6,978.

	Distillation			CUHK-SYSU	
	FD	AD	PD	Rank-1	mAP
1	-	-	-	49.5	45.1
2	✓	-	-	58.1	54.1
3	-	✓	-	52.0	47.8
4	-	-	✓	65.8	62.6
5	✓	✓	-	59.4	55.8
6	✓	-	✓	66.4	63.0
7	-	✓	✓	68.2	65.4
8	✓	✓	✓	**70.0**	**66.4**

the scalability advantages of our HDL model over a variety of state-of-the-art person search methods. We provide in-depth component analysis to give the insights on model performance gain and design rationale.

Acknowledgement

This work is supported by the China Scholarship Council, the Alan Turing Institute, and Innovate UK Industrial Challenge Project (98111-571149). We are especially grateful to the QMUL ITS Research group for their support.