Knowledge Distillation for End-to-End Person Search

Bharti Munjal[1,2]
m.bharti@osram.com

Fabio Galasso[2]
f.galasso@osram.com

Sikandar Amin[2]
s.amin@osram.com

[1] Technische Universität München,
Boltzmannstraße 3 Garching,
Munich 85748, Germany

[2] OSRAM GmbH,
Parkring 33 Garching,
Munich 85748, Germany

Abstract

We introduce knowledge distillation for end-to-end person search. End-to-End methods are the current state-of-the-art for person search that solve both detection and re-identification jointly. These approaches for joint optimization show their largest drop in performance due to a sub-optimal detector.

We propose two distinct approaches for extra supervision of end-to-end person search methods in a teacher-student setting. The first is adopted from state-of-the-art knowledge distillation in object detection. We employ this to supervise the detector of our person search model at various levels using a specialized detector. The second approach is new, simple and yet considerably more effective. This distills knowledge from a teacher re-identification technique via a pre-computed look-up table of ID features. It relaxes the learning of identification features and allows the student to focus on the detection task. This procedure not only helps fixing the sub-optimal detector training in the joint optimization and simultaneously improving the person search, but also closes the performance gap between the teacher and the student for model compression in this case. Overall, we demonstrate significant improvements for two recent state-of-the-art methods using our proposed knowledge distillation approach on two benchmark datasets. Moreover, on the model compression task our approach brings the performance of smaller models on par with the larger models.

1 Introduction

Person search is the complex (multi-)task of jointly localizing people and verifying their identity against a provided query person ID. Person search has recently gained attention [35, 53, 57, 58], also thanks to its numerous applications, including cross-camera visual tracking, person verification, and surveillance. The two tasks in person search, i.e. detection and re-identification are arguably in contrast with each other. In fact, detection should disregard any specific nuance of individuals and just retrieve any person, while re-identification should focus on these nuances, to distinguish individuals and retrieve the queried person. The dispute reflects in literature arguing that detection and re-identification in person search should be addressed separately [7, 57, 58] or jointly [35, 39, 53, 55].
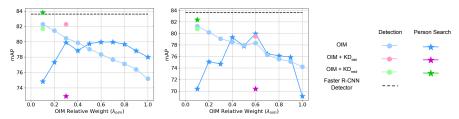
Figure 1: Joint detection and re-identification provides best performance [39, 55] but analysis shows that detection is harmed and limits performance. Cyan (dot-marker) and blue (star-marker) curves illustrate the OIM [55] detection and person search performance respectively, when varying the relative training importance via the λ_{oim} weight, cf. (3). Both in the case of ResNet50 (*left*) and ResNet18 (*right*), the more weight is given to re-id task, the more the detector is harmed (decreasing cyan curve), which remains below the original Faster R-CNN performance (*black dashed line*). A teacher-student framework for the detection part of OIM, OIM + KD$_{det}$ (*magenta*), improves detection (light magenta, dot marker) but harms re-id, which is visible by the drop in person search performance (dark magenta, star marker). The new knowledge distillation for the re-id part, OIM+KD$_{reid}$ (*green*) improves overall person search (dark green, star-marker) and also keeps the original detector performance.

Most recent state-of-the-art work in person search [39, 55] demonstrates the benefits of end-to-end optimization. The approaches add the re-identification task onto the Faster R-CNN [44] detection framework and learn both objectives jointly. As a result, the detector performance degrades, as illustrated in Fig. 1, but it still remains state-of-the-art w.r.t person search performance.

In this work, we propose knowledge distillation to address the sub-optimal detector performance for end-to-end person search, and model compression to reach the same level of performance with a smaller model. Knowledge distillation [25] stems from the belief that training a neural network from labelled data requires a large amount of over-parameterization, as it is the case for the teacher, generally a large and accurate model. The teacher supervisory training signal enables training of a student [51], typically smaller.

Here we propose two distinct ways to distill knowledge and supervise the two joint tasks of person search, i.e. detection and re-identification. We build upon the end-to-end approach of OIM [55], the de-facto building block of all most recent person search approaches [35, 39, 53]. For the detection part, we adopt the teacher-student framework of [8] which employs a multi-task loss, for intermediate feature guidance, region proposals, classification, and regression outputs. For the re-identification part, we propose a new, simple and yet effective strategy that relieves the student from learning a look-up table (LUT) for the identities in the training set. Instead, we copy and fix a pre-computed teacher LUT, which relaxes the student task of identification feature learning. We test our new distilled person search models on the two most recent CUHK-SYSU [55] and PRW-mini [39, 58] datasets, extending both the baseline OIM [55] technique and the most recent query-based method QEEPS [39]. We demonstrate performance improvement in all cases. We achieve 85.0% mAP on CUHK-SYSU and 39.7% mAP on PRW-mini. Notably, the same approach allows to train a smaller student network, realizing therefore model compression [4]. In fact, we show that a ResNet18 student still provides 84.1% mAP on CUHK-SYSU with only 46% of parameters of the larger Resnet50 teacher.

We summarize our contributions as follows: **i.** we introduce knowledge distillation for person search and propose two distinct teacher-student frameworks, i.e. for the detector and for re-identification parts; **ii.** we integrate our approach into the OIM [55] and QEEPS [39] person search models; **iii.** we show significant improvement over baseline methods on both the CUHK-SYSU [55] and PRW-mini [39, 58] datasets; **iv.** we also show that our knowledge

distillation approach enables model compression without drop in performance.

2 Related Work

In this section, we first discuss the literature on the multiple tasks encompassed within person search. Then we review prior art on knowledge distillation and model compression.

Person Detection. There is a large body of work on person detection, from methods based on hand-crafted features [16, 18] to deep convolutional neural network (CNN) feature-learning methods [19, 20, 60]. Best CNN detectors are either one-stage [37, 43] or two-stage [19, 20, 44]. The latter select object proposals via a region proposal network and then classify those into persons vs background. In line with OIM [55], we adopt Faster R-CNN with a ResNet [24] backbone, due to its robustness and flexibility.

Person Re-Identification. Person re-identification is the task of classifying the same individuals as provided by a query sample, within a gallery of cropped, centered and aligned persons. Earlier approaches have focused on manual feature design [17, 22, 51, 65] and metric learning [26, 32, 33, 34, 38, 40, 66]. More recent re-identification approaches are based on CNNs [1, 31, 52, 64] and mainly concerned with the estimation of a ID-feature embedding spaces, either via Siamese networks and contrastive losses [1, 9, 11, 15, 31, 36, 50, 56, 62], or with ID-classification and cross-entropy losses [54, 57].

Person Search. Xu *et al*. [57] introduced this as finding a person in a set of non-cropped gallery images, given a crop of the queried person. Person search involves detecting people in gallery images as well as verifying their ID against the provided query-ID. It encompasses therefore the two tasks of detection and re-identification. Thanks to large-scale datasets (CUHK-SYSU [55] and PRW [68]), person search has witnessed progress but it remains divided into approaches addressing the two tasks separately [7, 27, 57, 68] *vs* jointly [35, 39, 53, 55, 59]. We consider the two tasks jointly, since it was proven most recently beneficial [39].

End-to-End Person Search. Xiao *et al*. [55] introduced a model for the end-to-end training of joint person search. They extended Faster R-CNN to estimate an ID embedding for re-identification, and introduced an Online Instance Matching (OIM) loss, to effectively train it. IAN [53] also proposed an end-to-end approach using a center loss [52] with the goal to improve the intra-class feature compactness. More recently, Munjal *et al*. [39] proposed query-guidance for OIM, dubbed QEEPS, i.e. conditioning person search on the non-cropped query image. We propose knowledge distillation for the end-to-end OIM [55] and demonstrate the generality of our approach by also applying it to the current state-of-the-art QEEPS [39].

Knowledge Distillation and Model Compression. Knowledge distillation, proposed by [2, 5] and popularized by [25], aims to train a small neural network from a large and more accurate one. It has gained attention for its promise of more effective training [45], better accuracy [3] and efficiency [2, 41, 49, 58], but it remains strongly limited to networks solving the single classification task. When moving from classification to detection, as only few works [8, 30, 46] attempted, complex modelling questions arise as to where and how to supervise, which are not entirely answered yet. Such complex questions include the class importance imbalance, as for the background *vs* the other classes, and the implicit multi-task objective, since detection implies the joint bounding box regression and classification. Here, we apply knowledge distillation to the even more complex multi-task person search.

Distilling knowledge from a larger to a smaller network realizes model compression, i.e. train a smaller but accurate network. This has also been addressed for classification
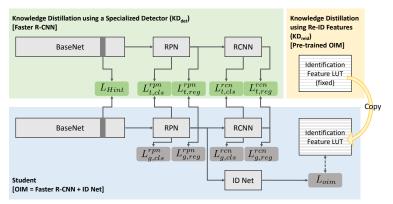
Figure 2: We propose two knowledge distillation approaches for person search. The first approach is motivated from [8] and uses the output of a specialized teacher detector (shown in green) to guide the detector of our person search student network (shown in blue). The second approach uses a copy of the LUT from a person search teacher (shown in yellow) and fix it during the student's training, thereby relaxing the task of ID feature learning and allowing the student to focus on the detection task.

via quantization [21, 59, 70] and binarization [12, 42] of floating point operations, network pruning [6, 10, 23, 29, 53] and factorization [28, 48]. The proposed knowledge distillation is directly applicable and realizes model compression for person search for the first time.

3 Background - Online Instance Matching

Online Instance Matching (OIM) loss and the end-to-end architecture proposed by Xiao *et al*. [55] is currently the de-facto standard for identification feature learning in end-to-end person search [35, 39, 53]. The architecture of [55] is based on Faster R-CNN [44] with a ResNet backbone [24], a Region Proposal Network (RPN) and a Region Classification Network (RCN), as illustrated in Fig. 2 (blue region). In parallel to the RCN, [55] defines an ID Net, which provides an identification feature embedding. They introduce an OIM loss as an additional objective in the Faster R-CNN framework focusing on the task of learning unique identity (ID) features for the image instances of the same person. This is accomplished by learning a lookup table (LUT) for all the identities in the training set. We refer to this approach as **OIM** in the text.

In more details, given ID features $x \in R^D$ where $D = 256$, OIM maintains a LUT $V \in R^{D \times P}$ for D-dimensional ID features corresponding to the P ID labels, and also a circular queue $U \in R^{D \times Q}$ containing most recent Q ID features of the unlabeled identities that appear in the recent mini-batches. During the forward pass of the network, the computed ID feature of each mini-batch sample is compared with all the entries in V and U, to compute the OIM loss. The OIM objective is to minimize the negative expected log-likelihood. Given the softmax probabilities p_t of the positive IDs in the mini-batch, the OIM loss is given by $L_{oim} = -E_x[\log p_t]$.

During the backward pass, entries of V corresponding to the identities in the current mini-batch are updated using moving-average. The OIM detection and re-identification objectives are in conflict during the optimization, which results in a significantly sub-optimal detector performance, as illustrated in Fig. 1. Intuitively, one would expect adjusting the relative weights of the two tasks might solve this problem. In fact we notice that the relative weights

do play a role, and by decreasing the weight of the ReID task the detector approaches the performance of the specialized detector (standalone Faster R-CNN). However, at the same time the ReID task becomes harder to train and therefore its performance drops significantly.

4 Knowledge Distillation in OIM

We propose two independent approaches for knowledge distillation in OIM for person search. First, we draw from most recent literature in knowledge distillation for object detection to supervise the detector of the OIM model with the help of a specialized (hence stronger) person detector. The second approach is to relax the task of re-identification with the help of a pre-computed identification feature table. As a result the detector becomes the focus during the optimization without compromising on the quality of identification features. Both proposed approaches contribute in recovering the sub-optimal detector performance, as we would illustrate experimentally in Sec. 6. However, we would demonstrate that the two approaches are not complementary. In the following we discuss both approaches in detail.

4.1 KD using a Specialized Detector

As a first approach, we propose that the detector of the student model (OIM) to mimic the superior output of a stronger detector (a teacher). The teacher in this case is a specialized Faster R-CNN with the same backbone architecture as the student. This approach stems from the belief that a better dedicated detector be a good teacher for the detector in OIM model, without modifying the OIM training.

To this end, we adopt the approach of Chen *et al.* [8] for the supervision of the student at various levels, from mimicking of the base features to the supervision of the detector. This corresponds to the green region in Fig. 2. As illustrated in the figure, we employ supervision for three different components of the Faster R-CNN [44] object detection framework. Following [8] these components are, **i.** intermediate base feature representations using a hint based learning [45], **ii.** RPN classification and regression modules to produce better region proposals. **iii.** likewise RCN classification and regression modules to generate stronger object detections. We refer to this approach as **KD**$_{\textbf{det}}$ in the text.

Following [8, 45], hint loss (L_{Hint}) is given as an L_2 loss between an intermediate layer of teacher BaseNet (F_t) and an adapted intermediate layer of student BaseNet ($f_a(F_s)$).

$$L_{Hint} = \|f_a(F_s) - F_t\|_2^2 \tag{1}$$

The classification loss (L_{cls}) and bounding box regression loss (L_{reg}) for both RPN and RCN sub-networks are given as:

$$\begin{aligned} L_{cls} &= \mu L_{g,cls} + (1 - \mu)L_{t,cls} \\ L_{reg} &= L_{g,reg} + \gamma L_{t,reg} \end{aligned} \tag{2}$$

$L_{g,\cdot}$ corresponds to the losses w.r.t the ground-truth labels [44] and $L_{t,\cdot}$ corresponds to the losses w.r.t the output of the teacher network. Motivated from [8], we use soft cross entropy loss i.e $-\sum P_t \log P_s$ as $L_{t,cls}$ where P_t and P_s are the softened teacher and student classification probabilities with temperature 10. For $L_{t,reg}$ we use teacher bounded regression loss as in [8]. The overall optimization objective for this approach is given as follows:

$$L = L_{cls}^{rpn} + L_{reg}^{rpn} + L_{cls}^{rcn} + L_{reg}^{rcn} + \lambda_{Hint}L_{Hint} + \lambda_{oim}L_{oim} \tag{3}$$

whereby, L_{oim} represents the OIM loss given in Sec. 3. For details on the contribution of each loss term (except L_{oim}), we refer the reader to [8]. We set μ, γ and λ_{Hint} to 0.5 as in [8], while we investigate different values of λ_{oim} in Sec. 6 and also Fig. 1.

4.2 KD using Pre-Trained Re-ID Features

The OIM optimization includes an iterative estimation of **i.** the ID feature embeddings, and **ii.** the lookup table (LUT) for the evaluation of these ID feature embeddings and computation of the OIM loss. The LUT in the original model [55] is randomly initialized. It eventually converges over time, however this iterative complexity impacts the learning of the parallel detection task.

We propose a new approach to person search with OIM, whereby we leverage knowledge distillation to relax one of the two tasks, i.e. estimating the LUT. In other words, distilling knowledge for the re-identification means that the student is not tasked any more with training for the OIM LUT. Most importantly, originally an optimization goal becomes instead a supervisory signal which eases the training and improves the performance.

To accomplish this, we fix the LUT V of our student OIM model using a copy from a pre-trained OIM model. This approach for knowledge distillation is illustrated by yellow components in Fig. 2. We refer to this approach as **KD$_{reid}$** in the text. The pre-computed V is fixed, hence not updated during the back-propagation step. The network is able to obtain the optimal supervision for the ID features directly from the very first iteration.

The proposed KD$_{reid}$ approach is in contrast with the KD$_{det}$ approach, since it aims to simplify the learning of the ID features while the latter focuses on directly supervising and improving the detector. Moreover, the proposed method do not add any overhead to the network training as compared to the original OIM. In fact it reduces the computations (FLOPS), since LUT V is fixed and we skip the step of updating it during back-propagation. Whereas, for KD$_{det}$ a forward pass over teacher network is also required.

Finally, we would like to emphasize that the proposed knowledge distillation KD$_{reid}$ is applicable to any OIM based person search method, since we just require to copy and fix the LUT V from a pre-trained model. We show this by applying KD$_{reid}$ to the most recent work Query-guided End-to-End Person Search (QEEPS) [39] in Sec. 6.

5 Model Compression

In this work, we also demonstrate achieving model compression using our proposed knowledge distillation approaches. We do not require any modification to the proposed approaches which we discussed in section 4.1 and 4.2. The only difference is that for model compression the backbone architecture of our student OIM network (blue region in Fig. 2) is much smaller than the specialized detector (green region), and the pre-trained OIM model (in yellow).

Prior works in knowledge distillation [8, 14] show that neural networks are often over-parametrized, and a proper teacher-student knowledge distillation has the potential to scale down the redundancy while keeping the performance intact. In other words, supervision from a stronger model as a teacher allows a weaker model to reach the level of performance which the smaller model itself, with current training procedures, is unable to arrive at.

In our evaluation we focus on different sizes of the Resnet [24] architecture to study model compression. In particular, our teacher has a larger backbone architecture (Resnet50), while our student is based on Resnet18, which is the smallest standard variant of this architecture. In Sec. 6, we discuss this further.

6 Experiments

CUHK-SYSU. The CUHK-SYSU is a large scale person search dataset [55] consisting of 18,184 images, labeled with 8,432 identities and 96,143 pedestrian bounding boxes (23,430 boxes are ID labeled). The dataset displays a large variation in person appearance, backgrounds, illumination conditions etc. For our evaluations, we adopt the standard train/test split as detailed in [39, 55].

PRW-mini. PRW [68] is another important dataset focusing on the task of person search. Different from CUHK-SYSU, PRW is acquired in a single setup, i.e. a university campus using six cameras. Overall it consists of 11,816 images with 34,304 bounding boxes and 932 identities. The diversity in the background and the appearance of the persons is limited compared to CUHK-SYSU. Munjal *et al.* [39] proposed PRW-mini[1], a subset of PRW, for a faster and representative benchmarking of the full dataset. The motivation came from the huge computational complexity of their query-guided person search method (QEEPS), which we also evaluate in this work.

Evaluation Metrics. Following [55, 39, 53, 55], we adopt mean Average Precision (mAP) and Common Matching Characteristic (CMC top-K) for results on person search. On the other hand, we report mAP for person detection results. mAP metric is common in the detection literature, reflecting both precision and recall of results. CMC is specific to re-identification and reports the probability of retrieving at least one correct ID within the top-K predictions (CMC top-1 is adopted here, which we refer to as top-1).

Implementation Details. Our implementation of OIM is based on Resnet [24] and uses first four blocks (conv1 to conv4) as BaseNet. The input images are re-scaled such that their shorter side is 600 pixels, unless specified otherwise. We employ the same training strategy as in [39]. Note that, we augment the data by flipping and initialize the backbone architecture using pre-trained ImageNet [13] weights.

6.1 Knowledge Distillation in OIM

In Table 1, we summarize the ablation study of our proposed knowledge distillation approaches, i.e. KD_{det} and KD_{reid}, on CUHK-SYSU [55] dataset. All experiments in this section consider Renset50 as the backbone architecture. We begin with the evaluation of a pure detector (Faster R-CNN [44]). We obtain a mAP of 83.6% and recall of 88.1% as the baseline for detection accuracy. We refer to this as DET_{50} in the text. Then, we evaluate a basic OIM [55] model with a relative weight of ReID task $\lambda_{oim} = 1$. We set this as our baseline for person search for rest of the experiments in this section and refer it as OIM_{50}, which achieves 78% mAP and 77.9% CMC top-1 for person search, meanwhile significantly underperforms on detector accuracy compared to DET_{50} by 8.4% mAP. In Fig. 1, we investigate the effect of relative weighting of the detection and re-identification tasks by varying λ_{oim} demonstrating contrasting nature of both tasks.

Next, we detail the results for our distillation approaches. We show that applying KD_{det} to OIM with DET_{50} as the teacher and $\lambda_{oim} = 1$ does not change the person search performance, however reasonably improves the detector accuracy by 3.7pp mAP and 2.3pp recall. On the other hand, selecting $\lambda_{oim} = 0.3$ degrades the person search by more than 5% in terms of mAP as well as top-1, but the detector recovers almost entire performance of teacher DET_{50}. This result is intuitive since the detector of the student model is getting an

---

[1] Publicly available at https://github.com/munjalbharti/Query-guided-End-to-End-Person-Search

Student	λ_{oim}	Type of KD	Teacher	Person Search		Detection	
Resnet50 Models				mAP(%)	top-1(%)	mAP(%)	Recall(%)
Faster R-CNN (DET$_{50}$)*	-	-	-	-	-	**83.6**	**88.1**
OIM (OIM$_{50}$, Baseline)*	1.0	-	-	78.0	77.9	75.2	82.7
OIM	1.0	KD$_{det}$	DET$_{50}$	78.3	77.5	78.9	85.0
OIM	0.3	KD$_{det}$	DET$_{50}$	72.9	72.0	82.3	87.2
OIM	0.1	KD$_{reid}$	OIM$_{50}$	81.2	81.0	81.9	86.9
OIM	0.1	KD$_{det}$, KD$_{reid}$	DET$_{50}$, OIM$_{50}$	80.3	79.9	83.2	87.8
OIM	0.1	KD$_{reid}$	OIM$_{18}$	75.0	73.9	81.7	86.8
OIM	0.1	KD$_{reid}$	QEEPS$_{50}$	83.8	84.2	81.7	86.8
QEEPS[39] (QEEPS$_{50}$)*	1.0	-	-	84.4	84.4	-	-
QEEPS	0.1	KD$_{reid}$	OIM$_{50}$	83.2	82.9	-	-
QEEPS	0.1	KD$_{reid}$	QEEPS$_{50}$	**85.0**	**85.5**	-	-

Table 1: Knowledge distillation for Resnet50 student models. Above the dashed line, are the detection and person search results of the student model using the two proposed distillation methods (KD$_{det}$ and KD$_{reid}$). For KD$_{det}$, Resnet50 Faster R-CNN detector (DET$_{50}$) is used as teacher. For KD$_{reid}$ Resnet50 OIM model (OIM$_{50}$) is used as teacher. Below the dashed line, are the results of using different teachers with OIM and QEEPS student. OIM$_{18}$ represents Resnet18 OIM model, QEEPS$_{50}$ represents Resnet50 QEEPS model. (*) indicates models trained without KD.

additional signal from the teacher network and at the same time we decreased the relative weight of the contrasting re-id task.

Keeping the results of KD$_{det}$ in view, for our second distillation approach KD$_{reid}$, we select $\lambda_{oim} = 0.1$ to ensure the detector of the student gets the required focus during the joint optimization. Whereas, supervision through KD$_{reid}$ is supposed to simplify the re-id task, hence justifying the lower value of its relative weight λ_{oim}. Interestingly, we observe significant improvements for both detection (6.7pp mAP and 4.2pp recall) and person search (3.2pp mAP and 3.1pp top-1) over the baseline OIM$_{50}$. Note that, KD$_{reid}$ is significantly simplified knowledge distillation approach in comparison to KD$_{det}$, and still it improves significantly on both detection and re-id tasks unlike KD$_{det}$. This result further motivates the importance of research into appropriate supervision and optimal training procedures.

We further combine KD$_{det}$ with KD$_{reid}$ in our evaluations in Table 1. We notice that the detector of the student reached the Faster R-CNN detector performance (83.2% mAP and 87.8% recall), however there is a drop in person search results as compared to employing only KD$_{reid}$ (80.3 vs 81.2 in mAP). Clearly, the addition of KD$_{det}$ increases the focus on the detection task; weakening the training of the re-id branch, hence declining the person search results. This result indicates the challenges in combining the two knowledge distillation approaches (KD$_{det}$, KD$_{reid}$). One could adjust the relative weighting of these approaches, we however adopt KD$_{reid}$ for all of our next experiments, due to its superior performance and simplicity.

Significance of Teachers' Quality. In Table 1 below the dashed line, we study the effect of using different teacher models in KD$_{reid}$, namely Resnet18 (referred as OIM$_{18}$) and recently proposed QEEPS [39] with Resnet50 (referred as QEEPS$_{50}$), in addition to the baseline OIM$_{50}$. Notice, how using OIM$_{18}$ as teacher drops the person search performance to 75% mAP and 73.9% top-1 which is 3pp mAP and 4pp top-1 lower than the baseline OIM$_{50}$. While, using QEEPS$_{50}$ as teacher gives significant improvement in person search over the baseline, 5.8pp mAP and 6.3pp top-1. This is 2.6pp mAP and 3.2pp top-1 better than using OIM$_{50}$ as a teacher. Overall, we can conclude that stronger teachers provide stronger supervision hence better results, while inferior teachers may harm the performance. However, we also demonstrate that students using same model as teacher, e.g. OIM Resnet50 student, OIM$_{50}$ teacher (cf. Table 1, 5th row) also improves the results due to improved train-

Method	mAP(%)	top-1 (%)
OIM [55]	75.5	78.7
Distilled OIM	**83.8**	**84.2**
QEEPS [39]	84.4	84.4
Distilled QEEPS	**85.0**	**85.5**
IAN [53]	76.3	80.1
NPSM [35]	77.9	81.2
Context Graph [59]	84.1	86.5
CLSA [27]	87.2	88.5

(a) CUHK-SYSU

Method	mAP(%)	top-1 (%)
OIM‡[55]	38.3	70.0
Distilled OIM‡	**39.5**	**73.3**
QEEPS [39]	39.1	80.0
Distilled QEEPS	**39.7**	**80.0**
Mask-G [8]	33.1	70.0

(b) PRW-mini

Table 2: Comparison to the state-of-the-art on, (a) CUHK-SYSU [55] (image size = 600), and (b) PRW-mini [39] (image size = 900), where OIM‡ is same as in [39].

ing conditions. It is worth noting that the detector performance, in this case, remains almost the same when using different teachers for KD_{reid} and $\lambda_{oim} = 0.1$ (mAP 81.9%, 81.7% and 81.7% for OIM_{50}, OIM_{18} and $QEEPS_{50}$, respectively), since a higher relative importance of the detector encourages its convergence to the performance of a pure detector (mAP 83.6%).

Next, we also evaluate the state-of-the-art QEEPS model [39] with Resnet50 as a student with supervision of OIM_{50} and $QEEPS_{50}$ as teachers. QEEPS does not consider intermediate detection stage, therefore we only report person search results. As shown in the table, using OIM_{50} to train QEEPS dropped the person search performance by 1.2pp mAP and 1.5pp top-1, while using $QEEPS_{50}$, results in an improvement of 0.6pp mAP and 1.1pp top-1.

Comparison to the State-of-the-Art. It is important to note that our knowledge distillation approach KD_{reid} can be directly applied to all methods that use OIM loss [55] for learning the identification features [7, 39, 55, 59]. In Table 2 (a), we demonstrate our approach on CUHK-SYSU [55] dataset for two such state-of-the-art methods, i.e. OIM [55] and QEEPS [39]. Our distilled OIM (above the dashed line) using KD_{reid} from $QEEPS_{50}$ outperforms OIM [55] by 8.3pp in mAP and 5.5pp in top-1. Similarly, our distilled QEEPS outperforms QEEPS [39] by 0.6pp mAP and 1.1pp top-1, achieving overall 85.0% mAP and 85.5% top-1. Also notice that our distilled QEEPS outperforms IAN [53], NPSM [35] and a recent context graph based approach [59] by 8.7pp, 7.1pp and 0.9pp mAP, respectively.

Our knowledge distillation approach is also applicable to other methods like IAN [53], NPSM [35] and CLSA [27] that learn the identification features using softmax loss. In this case, first we would need to compute an OIM-like ID feature table for the teacher model and then use it to additionally supervise the identification feature learning of the student model.

In Table 2 (b), we report results on PRW-mini. In this case we adopt image size of 900 pixels, following [39, 47]. Our distilled OIM‡[2] (above the dashed line) trained with KD_{reid} using $QEEPS_{50}$ surpasses OIM‡ by 1.2pp mAP and 3.3pp top-1. Similarly our distilled QEEPS surpasses QEEPS [39] by .6pp mAP.

6.2 Model Compression

In Table 3, we report the model compression results for our two knowledge distillation methods, KD_{det} and KD_{reid}. We employ Resnet18 (with ~46% of Resnet50 parameters), as the network architecture for all the student entries in this table. We keep DET_{50} as the teacher for distillation using KD_{det}.

---

[2]OIM‡ uses image size 900 whereas OIM uses image size 600.

Student Resnet18 Models	λ_{oim}	Type of KD	Teacher	Person Search		Detection	
				mAP(%)	top-1(%)	mAP(%)	Recall(%)
Faster R-CNN (DET_{18})*	-	-	-	-	-	**82.4**	**87.2**
OIM (OIM_{18}, Baseline)*	1	-	-	69.1	68.0	74.2	81.6
OIM	1	KD_{det}	DET_{50}	69.1	68.0	77.5	83.7
OIM*	0.6	-	-	79.9	80.4	78.4	84.4
OIM	0.6	KD_{det}	DET_{50}	70.4	69.0	79.7	85.4
OIM	0.1	KD_{reid}	OIM_{50}	80.5	80.9	80.6	85.9
OIM	0.1	KD_{det}, KD_{reid}	DET_{50}, OIM_{50}	78.4	78.0	82.0	86.9
OIM	0.1	KD_{reid}	OIM_{18}	72.9	71.1	80.6	85.9
OIM	0.1	KD_{reid}	$QEEPS_{50}$	82.4	83.0	80.8	86.0
QEEPS*	1	-	-	76.6	76.0	-	-
QEEPS	0.1	KD_{reid}	OIM_{50}	82.1	81.4	-	-
QEEPS	0.1	KD_{reid}	$QEEPS_{50}$	**84.1**	**84.3**	-	-

Table 3: Knowledge distillation for Resnet18 student models. Above the dashed line, are the detection and person search results of the student model using the two proposed distillation methods (KD_{det} and KD_{reid}). For KD_{det}, Resnet50 Faster R-CNN detector (DET_{50}) is used as teacher. For KD_{reid} Resnet50 OIM model (OIM_{50}) is used as teacher. Below the dashed line, are the results of using different teachers with OIM and QEEPS as students. OIM_{18} represents Resnet18 OIM model, $QEEPS_{50}$ represents Resnet50 QEEPS model. (*) indicates models trained without KD.

The person search results for our baseline model OIM_{18} i.e. 69.1% mAP and 68% top-1 are significantly lower than OIM_{50} by around 9%. Whereas, the detector of OIM_{18} is only slightly worse than OIM_{50} but significantly below the pure detector accuracy of DET_{18} (mAP 82.4%). As shown in Table 3, the application of KD_{det} to OIM at $\lambda_{oim} = 1$, improves the detection performance by 3.3pp mAP and 2.1pp recall with no effect on person search results. We also evaluate the KD_{det} at $\lambda_{oim} = 0.6$, as it provides the best trade-off between detection and re-id tasks (cf. Fig. 1 (right)). We notice that the detection improves by 1.3pp mAP and 1pp recall, while person search performance drops by a large value of 9.5pp mAP and 11.4pp top-1. This result indicates that while detector is being directly supervised through KD_{det}, a lower λ_{oim} for re-id task during optimization is counter-productive for person search.

Next we notice that, our second approach for distillation KD_{reid} is quite effective also for the model compression scenario. It brings a significant improvement over the baseline on both person search and detection benchmarks. Specifically, it improves by 11.4pp mAP and 12.9pp top-1 for person search and 6.4pp mAP and 4.3pp recall for detection. Quite interestingly, our trained Resnet18 OIM model outperforms Resnet50 baseline OIM model both for person search (78% mAP *vs* 80.5%) and detection (75.2% mAP *vs* 80.6%).

Finally, below the dashed line in Table 3, we report results for Resnet18 based OIM and QEEPS student models through supervision from different teacher models such as OIM_{18}, OIM_{50}, and $QEEPS_{50}$. In particular, supervision of Resnet18 QEEPS with $QEEPS_{50}$ supervision achieves on par performance to the state-of-the-art results of $QEEPS_{50}$ [39].

7 Conclusions

We have introduced knowledge distillation for person search and proposed two approaches to supervise either the detector or the re-identification part of two most recent models, OIM [55] and QEEPS [39]. In both cases we improve performance on the CUHK-SYSU [55] and PRW-mini [39] datasets, which extends to model compression.

Our approach is the first-of-a-kind, relaxing the multi-task person search optimization by transferring one task to a teacher. We plan to further investigate whether this multi-task relaxation approach may apply to other multi-task goals.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[2] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems 27*. 2014.

[3] Vasileios Belagiannis, Azade Farshad, and Fabio Galasso. Adversarial network compression. In *ECCV Workshops*, 2018.

[4] Cristian Bucilă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, 2006.

[5] R. Caruana C. Bucilua and A. Niculescu-Mizil. Model compression. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[6] Miguel ÃĄ. Carreira-PerpiÃśÃ₫n and Yerlan Idelbayev. âĂIJlearning-compressionâĂİ algorithms for neural net pruning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[8] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems 30*. 2017.

[9] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1320–1329, 2017.

[10] Wenlin Chen, James T. Wilson, Stephen Tyree, Kilian Q. Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. In *ICML*, 2015.

[11] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, 2016.

[12] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*. 2015.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[14] Misha Denil, Babak Shakibi, Laurent Dinh, Marc Aurelio Ranzato, and Nando de Freitas. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems 26*. 2013.

[15] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48:2993–3003, 2015.

[16] Piotr Dollár, Ron Appel, Serge J. Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1532–1545, 2014.

[17] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, 2010.

[18] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2009.

[19] Ross B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[20] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[21] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir D. Bourdev. Compressing deep convolutional networks using vector quantization. *CoRR*, 2014.

[22] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[23] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *NIPS*, 2015.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

[26] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[27] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person Search by Multi-Scale Matching. In *ECCV2018*, 2018.

[28] Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan V. Oseledets, and Victor S. Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *CoRR*, 2015.

[29] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *CoRR*, 2017.

[30] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, June 2014. doi: 10.1109/CVPR.2014.27.

[32] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3765–3773, 2015.

[33] Shengcai Liao and Stan Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3685–3693, 2015.

[34] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, 2015.

[35] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan. Neural person search machines. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 493–501, Oct 2017. doi: 10.1109/ICCV.2017.61.

[36] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26:3492–3506, 2017.

[37] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[38] T M Feroz Ali and Subhasis Chaudhuri. Maximum margin metric learning over discriminative nullspace for person re-identification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[39] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1846–1855, 2015.

[41] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *ICLR*, 2018.

[42] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, 2016.

[43] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, pages 91–99, 2015.

[45] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *In Proceedings of ICLR*, 2015.

[46] Jonathan Shen, Noranart Vesdapunt, Vishnu Naresh Boddeti, and Kris M. Kitani. In teacher we trust: Learning compressed models for pedestrian detection. *CoRR*, abs/1612.00478, 2016. URL http://arxiv.org/abs/1612.00478.

[47] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[48] Amos Sironi, Bugra Tekin, Roberto Rigamonti, Vincent Lepetit, and Pascal Fua. Learning separable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[49] Gregor Urban, Krzysztof J Geras, Samira Ebrahimi Kahou, Ozlem Aslan, Shengjie Wang, Rich Caruana, Abdelrahman Mohamed, Matthai Philipose, and Matt Richardson. Do deep convolutional nets really need to be deep and convolutional. In *ArXiv*, 2016.

[50] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[51] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter H. Tu. Shape and appearance context modeling. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[52] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[53] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. Ian: The individual aggregation network for person search. *CoRR*, abs/1705.05552, 2017.

[54] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1249–1258, 2016.

[55] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3376–3385. IEEE, 2017.

[56] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[57] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 937–940, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10.1145/2647868. 2654965. URL http://doi.acm.org/10.1145/2647868.2654965.

[58] Zheng Xu, Yen-Chang Hsu, and Jiawei Huang. Training student networks for acceleration with conditional adversarial networks. In *BMVC*, 2018.

[59] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[60] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li. Convolutional channel features. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 82–90, 2015.

[61] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan Yuille. Knowledge distillation in generations: More tolerant teachers educate better students. In *AAAI*, 2018.

[62] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. *2014 22nd International Conference on Pattern Recognition*, pages 34–39, 2014.

[63] Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I. Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S. Davis. Nisp: Pruning networks using neuron importance score propagation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[64] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[65] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[66] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by saliency learning. *IEEE transactions on pattern analysis and machine intelligence*, 39 2:356–370, 2017.

[67] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.

[68] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[69] Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *CoRR*, 2017.

[70] Chenzhuo Zhu, Song Han, Huizi Mao, and William J. Dally. Trained ternary quantization. *CoRR*, 2017.