Person Search by Multi-Scale Matching

Xu Lan[1], Xiatian Zhu[2], and Shaogang Gong[1]

[1] Queen Mary University of London,
x.lan@qmul.ac.uk, s.gong@qmul.ac.uk
[2] Vision Semantics Ltd
eddy@visionsemantics.com

Abstract. We consider the problem of person search in unconstrained scene images. Existing methods usually focus on improving the person detection accuracy to mitigate negative effects imposed by misalignment, mis-detections, and false alarms resulted from noisy people auto-detection. In contrast to previous studies, we show that sufficiently reliable person instance cropping is achievable by slightly improved state-of-the-art deep learning object detectors (e.g. Faster-RCNN), and the under-studied multi-scale matching problem in person search is a more severe barrier. In this work, we address this multi-scale person search challenge by proposing a *Cross-Level Semantic Alignment* (CLSA) deep learning approach capable of learning more discriminative identity feature representations in a unified end-to-end model. This is realised by exploiting the in-network feature pyramid structure of a deep neural network enhanced by a novel cross pyramid-level semantic alignment loss function. This favourably eliminates the need for constructing a computationally expensive image pyramid and a complex multi-branch network architecture. Extensive experiments show the modelling advantages and performance superiority of CLSA over the state-of-the-art person search and multi-scale matching methods on two large person search benchmarking datasets: CUHK-SYSU and PRW.

Keywords: Person Search; Person Detection and Re-Identification; Multi-Scale Maching; Feature Pyramid; Image Pyramid; Semantic Alignment.

1 Introduction

Person search aims to find a probe person in a gallery of whole unconstrained scene images [41]. It is an extended form of person re-identification (re-id) [12] by additionally considering the requirement of automatically detecting people in the scene images besides matching the identity classes. Unlike the conventional person re-id problem assuming the gallery images as either manually cropped or carefully filtered auto-detected bounding boxes [40, 24, 3, 15, 20, 44, 25, 39, 37, 2, 44], person search deals with raw unrefined detections with many false cropping and unknown degrees of misalignment. This yields a more challenging matching problem especially in the process of person re-id. Moreover, auto-detected person boxes often vary more significantly in scale (resolution) than the conventional

person re-id benchmarks (Fig. 1(b)), due to the inherent uncontrolled distances between persons and cameras (Fig. 1(a)). It is therefore intrinsically a *multi-scale matching* problem. However, this problem is currently under-studied in person search [41, 47, 28].

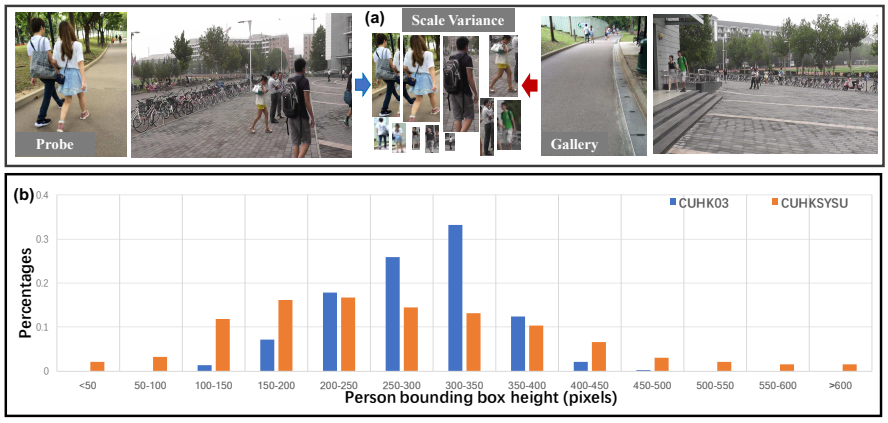

Fig. 1. Illustration of the intrinsic multi-scale matching challenge in person search. (a) Auto-detected person bounding boxes vary significantly in scale. (b) The person scale distribution of CUHK-SYSU (person search benchmark) covers a much wider range than manually refined CUHK-03 (person re-id benchmark).

In this work, we aim to address the multi-scale matching problem in person search. We show that this is a significant factor in improving the model matching performance, given the arbitrary and unknown size changes of persons in auto-detected bounding boxes. However, existing methods [41, 47, 28] focus on the person detection and localisation in scene images, which turns out not to be a severe bottleneck for the overall search performance as indicated in our experiments. For example, using the ground-truth person bounding boxes only brings a Rank-1 gain of 1.5% alongside employing ResNet-50 [14] for person search on the CUHK-SYSU benchmark [41]. In contrast, with the same ResNet-50 model, our proposed multi-scale matching learning improves the person search Rank-1 rate by 6.0% on the same benchmark (Fig. 6).

We make three **contributions** in this study: **(1)** We identify the multi-scale matching problem in person search – an element missing in the literature but found to be significant for improving the model performance. **(2)** We formulate a *Cross-Level Semantic Alignment* (CLSA) deep learning approach to addressing the multi-scale matching challenge. This is based on learning an end-to-end in-network feature pyramid representation with superior robustness in coping with variable scales of auto-detected person bounding boxes. **(3)** We improve the Faster-RCNN model for more reliable person localisation in uncontrolled scenes, facilitating the overall search performance. Extensive experiments on two bench-

marks CUHK-SYSU [41] and PRW [47] show the person search advantages of the proposed CLSA over state-of-the-art methods, improving the best competitor by 7.3% on CUHK-SYSU and 11.9% on PRW in Rank-1 accuracy.

2 Related Work

Person Search Person search is a recently introduced problem of matching a probe person bounding box against a set of gallery whole scene images [41, 47]. This is challenging due to the uncontrolled false alarms, mis-detections, and misalignment emerging in the auto-detection process. In the literature, there are only a handful of person search works [41, 47, 28]. Xiao et al. [41] propose a joint detection and re-id deep learning model for seeking their complementary benefits. Zheng et al. [47] study the effect of person detection on the identity matching performance. Liu et al. [28] consider recursively search refinement to more accurately locate the target person in the scene. While existing methods focus on detection enhancement, we show that by a state-of-the-art deep learning object detector with small improvements, person localisation is not a big limitation. Instead, the multi-scale matching problem turns out a more severe challenge in person search. In other words, solving the multi-scale problem is likely to bring more performance gain than improving person detection (Fig. 6(c)).

Person Re-Identification Person search is essentially an extension of the conventional person re-id problem [12] with an additional requirement of automatic person detection in the scenes. Given the manual construction nature of re-id datasets, the scale diversity of gallery images tends to be restricted. It is simply harder for humans to verify and label the person identity of small bounding boxes, therefore leading to the selection and labelling bias towards large boxes (Fig. 1(b)). Consequently, the intrinsic multi-scale matching challenge is *artificially* suppressed in re-id benchmarks, hence losing the opportunity to test the real-world model robustness. Existing re-id methods can mostly afford to ignore the problem of multi-scale person bounding boxes in algorithm design. Whilst extensive efforts have be made to solving the re-id problem [39, 23, 7, 36, 40, 24, 3, 46, 20, 48, 37, 17, 25, 22, 38, 5, 6], there are only limited works considering multi-scale matching [5, 29]. Beyond all these existing methods, our CLSA is designed specially to explore the in-network feature pyramid in deep learning for more effectively solving the under-studied multi-scale challenge in person search.

3 Cross-Level Semantic Alignment for Person Search

We want to establish a person search system capable of automatically detecting and matching persons in unconstrained scenes with any probe person. With the arbitrary distances between people and cameras in public space, person images are inherently captured at varying scales and resolutions. This raises the multi-scale matching challenge. To overcome this problem, we formulate a Cross-Level Semantic Alignment (CLSA) deep learning approach. An overview of the CLSA

is illustrated in Fig. 2. The CLSA contains two components: (1) Person detection which locates all person instances in the gallery scene images for facilitating the subsequent identity matching. (2) Person re-identification which matches the probe image against a large number of arbitrary scale gallery person bounding boxes (the key component of CLSA). We provide the component details below.

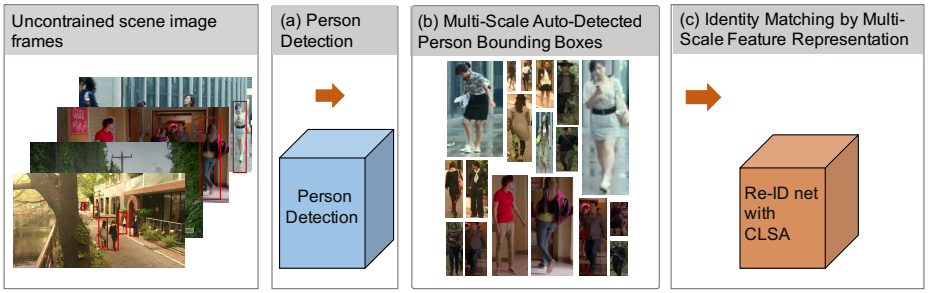

Fig. 2. Overview of the proposed multi-scale learning person search framework. (a) Person detection for cropping people from the whole scene images at (b) varying scales (resolutions). (c) Person identity matching is then conducted by a re-id model.

3.1 Person Detection

As a pre-processing step, person detection is important in order to achieve accurate search [41, 47]. We adopt the Faster-RCNN model [35] as the CLSA detection component, due to its strong capability of detecting varying sized objects in unconstrained scenes. To further enhance person detection performance and efficiency, we introduce a number of design improvement on the original model. **(1)** Instead of using the conventional RoI (Region of Interest) pooling layer, we crop and resize the region feature maps to 14×14 in pixel, and further max-pool them to 7×7 for gaining better efficiency [4]. **(2)** After pre-training the backbone ResNet-50 net on ImageNet-1K, we fix the 1^{st} building-block (the 1^{st} 4 layers) in fine-tuning on the target person search data. This allows to preserve the shared low-level features learned from larger sized source data whilst simultaneously adapting the model to target data. **(3)** We keep and exploit all sized proposals for reducing the mis-detection rate at extreme scales in uncontrolled scenes before the Non-Maximum Suppression (NMS) operation. In deployment, we consider all detection boxes scored above 0.5, rather than extracting a fixed number of boxes from each scene image [47]. This is because the gallery scene images may contain varying (unknown in priori) number of people.

3.2 Multi-Scale Matching by Cross-Level Semantic Alignment

Given auto-detected person bounding boxes at arbitrary scales from the gallery scene images, we aim to build a person identity search model robust for multi-

scale matching. To this end, we explore the seminal image/feature pyramid concept [1, 21, 31, 8]. Our motivation is that a single-scale feature representation blurs salient and discriminative information at different scales useful in person identity matching; And a pyramid representation allows to be "scale-invariant" (more "scale insensitive") in the sense that a scale change in matching images is counteracted by a scale shift within the feature pyramid.

Build-In Feature Pyramid We investigate the multi-scale feature representation learning in deep Convolutional Neural Network (CNN) to exploit the built-in feature pyramid structure formed on a single input image scale. Although CNN features have shown to be more robust to variance in image scale, pyramids are still effective in seeking more accurate detection and recognition results [27].

For the CNN architecture, we adopt the state-of-the-art ResNet-50 [14] as the backbone network (Fig. 3) of the identity matching component. In this study, we particularly leverage the feature pyramid hierarchy with low-to-high levels of semantics from bottom to top layers, automatically established in model learning optimisation [43]. Given the block-wise net structure in ResNet-50, we build a computationally efficient K-levels feature pyramid using the last conv layer of top-K ($K = 3$ in our experiments) blocks. The deepest layer of each block is supposed to have the most semantic features.

Nonetheless, it is not straightforward to exploit the ResNet-50 feature hierarchy. This is because the build-in pyramid has large semantic gaps across levels due to the distinct depths of layers. The features from lower layers are less discriminative for person matching therefore likely hurt the overall representational capacity if applied jointly with those from higher layers.

Cross-Level Semantic Alignment To address the aforementioned problems, we improve the in-network feature pyramid by introducing a Cross-Level Semantic Alignment (CLSA) learning mechanism. The aim is to achieve a feature pyramid with all levels encoding the desired high-level person identity semantics. Formally, to train our person identity matching model, we adopt the softmax Cross-Entropy (CE) loss function to optimise an identity classification task. The CE loss on a training person bounding box $(\boldsymbol{I}, y)$ is computed as:

$$\mathcal{L}_{\text{ce}} = -\log\Big(\frac{\exp(\boldsymbol{W}_y^\top \boldsymbol{x})}{\sum_{i=1}^{|\mathcal{Y}|}\exp(\boldsymbol{W}_i^\top \boldsymbol{x})}\Big) \tag{1}$$

where $\boldsymbol{x}$ specifies the feature vector of $\boldsymbol{I}$ by the last layer, $\mathcal{Y}$ the training identity class space, and $\boldsymbol{W}_y$ the y-th ($y \in \mathcal{Y}$) class prediction function parameters.

In our case, $\boldsymbol{x}$ is the top pyramid level, also denoted as $\boldsymbol{x}^K$. For anyone of the top-K ResNet blocks, we obtain $\boldsymbol{x}$ by applying an average pooling layer and a FC layer on the output feature maps (Fig. 3 (b)). Consider the different feature scale distributions across layers [30], we further normalise $\boldsymbol{x}$ by batch normalisation and ReLU non-linearity. In this way, we compute the feature representations for all K pyramid layers $\{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^K\}$.

Recall that we aim to render all levels of feature representations identity semantic. To this end, we first project each of these features $\{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^K\}$ by a FC layer into the identity semantic space with the same dimension as $\mathcal{Y}$. The

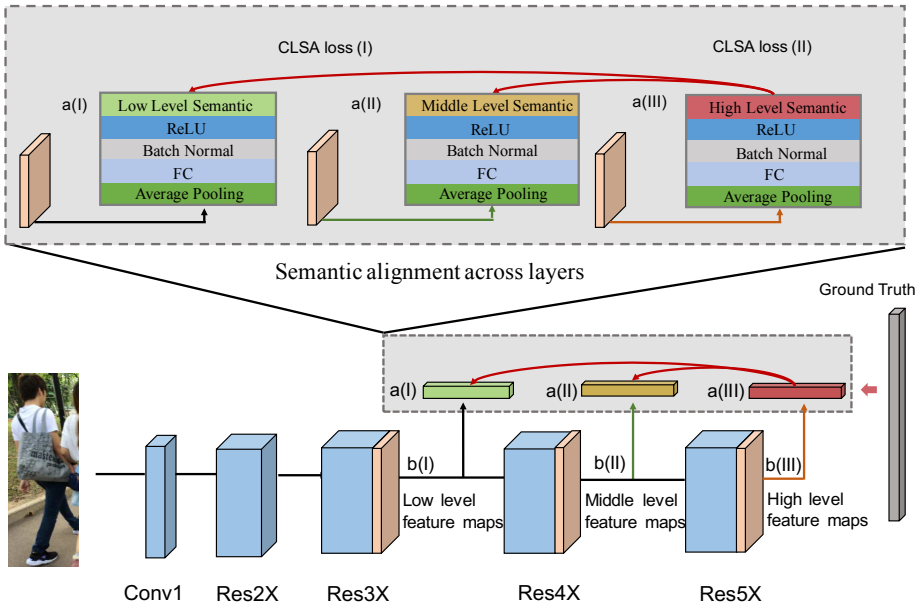

Fig. 3. Overview of the proposed Cross-Level Semantic Alignment (CLSA) approach in a ResNet-50 based implementation.

resulted semantic class probability vectors are denoted as $\{\boldsymbol{p}^1, \cdots, \boldsymbol{p}^K\}$ with $\boldsymbol{p}^k = [p_1^k, \cdots, p_{|\mathcal{Y}|}^k]$, $k \in \{1, \cdots, K\}$. To transfer the strongest semantics from the top (K-th) pyramid level to a lower (s-th) level, we introduce a Kullback-Leibler divergence based Cross-Level Semantic Alignment (CLSA) loss formulation inspired by knowledge distillation [16]:

$$\mathcal{L}_{\text{clsa}}(s) = \sum_{j=1}^{|\mathcal{Y}|} \tilde{p}_j^K \log \frac{\tilde{p}_j^K}{\tilde{p}_j^s}. \tag{2}$$

where $\tilde{p}_j^k$ is a *softened* per-class prediction semantic score obtained by

$$\tilde{p}_j^k = \frac{\exp(p_j^k/T)}{\sum_{j=1}^{|\mathcal{Y}|} \exp(p_j^k/T)}, \tag{3}$$

where the temperature parameter T controls the softening degree (higher values meaning more softened predictions). We set T=3 following the suggestion in [16]. To enable end-to-end deep learning, we add this CLSA loss on top of the conventional CE loss (Eq (1)):

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + T^2 \sum_{s=1}^{K-1} \mathcal{L}_{\text{clsa}}(s) \tag{4}$$

where T^2 serves as a weighting parameter between the two loss terms.

Identity Matching by CLSA Feature Pyramid In deployment, we first compute a CLSA feature pyramid by forward propagating any given person bounding box image. We then concatenate the feature vectors of all pyramid levels as the final representation for person re-id matching.

Remarks The CLSA is similar in spirit to a few person re-id matching methods [5, 29]. However, these methods adopt the image pyramid scheme, in contrast to the CLSA leveraging the in-network feature pyramid on a single image scale therefore more efficient. The FPN model [27] also exploits the build-in pyramid. The CLSA differs from FPN in a number of fundamental ways: (1) FPN focuses on object detection and segmentation, whilst CLSA aims to address fine-grained identity recognition and matching. (2) FPN additionally performs feature map unsampling hence less efficient than CLSA. (3) CLSA performs semantic alignment and transfer in the low-dimensional class space, in comparison to more expensive FPN's feature alignment. We will evaluate and compare these multi-scale learning methods against CLSA in our experiments (Table 4).

4 Experiments

Datasets To evaluate the CLSA, we selected two person search benchmarks: CUHK-SYSU [41] and PRW [47]. We adopted the standard evaluation setting as summarised in Table 1. In particular, the CUHK-SYSU dataset contains 18,184 scene images, 8,432 labelled person IDs, and 96,143 annotated person bounding boxes. Each probe person appears in two or more scene gallery images captured from different locations. The training set has 11,206 images and 5,532 probe persons. Within the testing set, the probe set includes 2,900 person bounding boxes and the gallery contains a total of 6,978 whole scene images. The PRW dataset provides a total of 11,816 video frames and 43,110 person bounding boxes. The training set has 482 different IDs from 5,704 frames. The testing set contains 2,057 probe people along with a gallery of 6,112 scene images. In terms of bounding box scale, CUHK-SYSU and PRW range from 37×13 to 793×297, and 58×21 to 777×574, respectively. This shows the two person search datasets present the intrinsic multi-scale challenge. Example images are shown in Fig. 4.

Performance Metrics For person detection, a person box is considered as correct if overlapping with the ground truth over 50% [41, 47]. For person identity matching or re-id, we adopted the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP). The CMC is computed on each individual rank k as the probe cumulative percentage of truth matches appearing at ranks $\leq k$. The mAP measures the recall of multiple truth matches, computed by first computing the area under the Precision-Recall curve for each probe, then calculating the mean of Average Precision over all probes [46].

Implementation Details We adopted the Pytorch framework [33] to conduct all the following experiments. For training the person detector component, we adopted the SGD algorithm with the momentum set to 0.9, the weight decay to 0.0001 the iteration to 110,000, and the batch size to 256. We initialised

Fig. 4. Example probe person and unconstrained scene images on (a) CUHK-SYSU [41] and (b) PRW [47]. Green bounding box: the ground truth probe person in the scene. ✓: Contain the probe person. ✗: Not contain the probe person.

Table 1. Evaluation setting, data statistics, and person bounding box scale of the CUHK-SYSU and PRW benchmarks. Bbox: Bounding box.

Dataset	Images	Bboxes	IDs	**Bbox Scale**	ID Split		Bbox Split	
					Train	Test	Train	Test
CUHK-SYSU	18,184	96,143	8,432	$37\times13\sim793\times297$	5,532	2,900	55,272	40,871
PRW	11,816	43,110	932	$58\times21\sim777\times574$	482	450	18,048	25,062

the learning rate at 0.001, with a decay factor of 10 at every 30,000 iterations. For training the identity matching component, we used both annotated and detected (over 50% Intersection over Union (IoU) with the annotated and sharing the identity labels) boxes as [47]. We set the momentum to 0.9, the weight decay to 0.00001, the batch size to 64, and the epoch to 100. The initial learning rate was set at 0.01, and decayed by 10 at every 40 epochs. All person bounding boxes were resized to 256×128 pixels. To construct the in-network feature pyramid, we utilised the top 3 (Res3x, Res4x, Res5x) blocks in our final model implementation, i.e. $K=3$ in Eq. (4). We also evaluated other pyramid constructing ways in the component analysis (Sec. 4.3).

4.1 Comparisons to State-Of-The-Art Person Search Methods

We compared the proposed CLSA method with two groups of existing person search approaches: (1) Three most recent state-of-the-art methods (NPSM [28], OIM [41], CWS [47]); and (2) Five popular person detectors (DPM [10], ACF [9], CCF [42], LDCF [32], and R-CNN [11]) with hand-crafted (BoW [46], LOMO [26], DenseSIFT-ColorHist (DSIFT) [45]) or deep learning (IDNet [41]) features based re-id metric learning methods (KISSME [18], XQDA [26]).

Evaluation on CUHK-SYSU Table 2 reports the person search performance on CUHK-SYSU with the standard gallery size of 100 scene images. It is clear

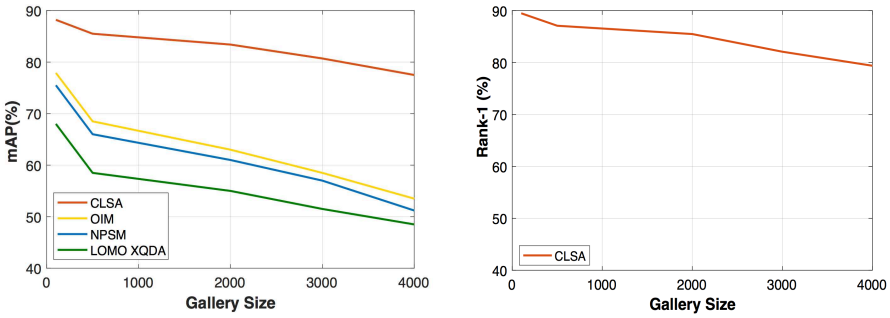

Fig. 5. Model scalability evaluation over different gallery search sizes on CUHK-SYSU.

that the CLSA significantly outperforms all other competitors. For instance, the CLSA surpasses the top-2 alternative models NPSM and OIM (both are end-to-end deep learning models) by 7.3% (88.5-81.2) and 9.8% (88.5-78.7) in Rank-1, 9.3% (87.2-77.9) and 11.7% (87.2-75.5) in mAP, respectively. The performance margin of CLSA against other non-deep-learning methods is even larger, due to that these models rely on less discriminative hand-crafted features without the modelling advantage of jointly learning stronger representation and matching metric model. This shows the overall performance superiority of the CLSA over current state-of-the-art methods, thanks to the joint contributions of improved person detection model (see more details below) and the proposed multi-scale deep feature representation learning mechanism.

To evaluate the model efficiency, we conducted a person search test among 100 gallery images on CUHK-SYSU. We deployed a desktop with a Nvidia Titan X GPU. Applying CLSA, OIM, and NPSM takes 1.2, 0.8, and 120 seconds, respectively. This indicates that the performance advantages of our CLSA do not sacrifice the model efficiency.

To test the model performance scalability, we further evaluated top-3 methods under varying gallery sizes in the range from 100 to 4,000 (the whole test gallery set). We observed in Fig. 5 that all methods degrade the performance given larger gallery search pools. When increasing the gallery size from 100 to 4,000, the mAP performance of NPSM drops from 77.9% to 53.0%, i.e. -24.9% degradation (no reported Rank-1 results). In comparison, the CLSA is more robust against the gallery size, with mAP/Rank-1 drop at -9.7% (77.5-87.2) and -9.1% (79.4-88.5). This is primarily because more distracting people are involved in the identity matching process, presenting more challenging tasks. Importantly, the performance gain of CLSA over other competitors becomes even higher at larger search scales, desirable in real-world applications. This indicates the superior deployment scalability and robustness of CLSA over existing methods in tackling a large scale person search problem, further showing the importance of solving the previously ignored multi-scale matching challenge given auto-detected noisy bounding boxes in person search.

Table 2. Evaluation on CUHK-SYSU. Gallery size: 100 scene images. The best and second-best results are in red and blue.

Method	Rank-1 (%)	mAP (%)
ACF[9]+DSIFT[45]+Euclidean	25.9	21.7
ACF[9]+DSIFT[45]+KISSME[18]	38.1	32.3
ACF[9]+LOMO[26]+XQDA[26]	63.1	55.5
CCF[42] +DSIFT[45]+Euclidean	11.7	11.3
CCF[42]+DSIFT[45]+KISSME[18]	13.9	13.4
CCF[42]+LOMO[26]+XQDA[26]	46.4	41.2
CCF[42]+IDNet[41]	57.1	50.9
CNN[35]+DSIFT[45]+Euclidean	39.4	34.5
CNN[35]+DSIFT[45]+KISSME[18]	53.6	47.8
CNN[35]+LOMO[26]+XQDA[26]	74.1	68.9
CNN[35]+IDNet[41]	74.8	68.6
OIM[41]	78.7	75.5
NPSM[28]	**81.2**	**77.9**
CLSA	**88.5**	**87.2**

Evaluation on PRW We further evaluated the CLSA against 11 existing competitors on the PRW dataset under the benchmarking setting with 11,816 gallery scene images. Overall, we observed similar performance comparisons with the state-of-the-art methods as on CUHK-SYSU. In particular, the CLSA is still the best person search performer with significant accuracy margins over other alternative methods, surpassing the second-best model NPSM by 11.9% (65.0-53.1) and 14.5% (38.7-24.2) in Rank-1 and mAP, respectively. This consistently suggests the model design advantages of CLSA over existing person search methods in a different video surveillance scenario.

Table 3. Evaluation on PRW. The best and second-best results are in red and blue.

Method	Rank-1 (%)	mAP (%)
ACF-Alex[9]+LOMO[26]+XQDA[26]	30.6	10.3
ACF-Alex[9]+IDE$_{det}$[47]	43.6	17.5
ACF-Alex[9]+IDE$_{det}$[47] +CWS [47]	45.2	17.8
DPM-Alex[10]+LOMO[26]+XQDA[26]	34.1	13.0
DPM-Alex[10]+IDE$_{det}$[47]	47.4	20.3
DPM-Alex[10]+IDE$_{det}$[47]+CWS[47]	48.3	20.5
LDCF[32]+LOMO[26]+XQDA[26]	31.1	11.0
LDCF[32]+IDE$_{det}$[47]	44.6	18.3
LDCF[32]+IDE$_{det}$[47] +CWS[47]	45.5	18.3
OIM[41]	49.9	21.3
NPSM[28]	**53.1**	**24.2**
CLSA	**65.0**	**38.7**

Table 4. Evaluating different multi-scale deep learning methods on CUHK-SYSU in the standard 100 sized gallery setting. FLOPs: FLoating point OPerations.

Method	Rank-1 (%)	mAP (%)	FLOPs ($\times 10^9$)
ResNet-50	82.5	81.6	**2.678**
In-Network Pyramid	81.1	80.2	**2.678**
DeepMu [34]	78.3	75.8	-
MST [13]	82.7	81.9	8.034
DPFL [5]	84.7	83.8	5.400
FPN [27]	85.5	85.0	4.519
CLSA	**88.5**	**87.2**	2.680

4.2 Comparisons to Alternative Multi-Scale Learning Methods

Apart from existing person search methods, we further evaluated the effectiveness of CLSA by comparing with the in-network feature pyramid (baseline) and four state-of-the-art multi-scale deep learning approaches including DeepMu [34], MST [13], DPFL [5], and FPN [27] on the CUHK-SYSU benchmark. We used the standard 100 sized gallery setting in this test. For all compared methods, we utilised the same person detection model and the same backbone identity matching network (except DeepMu that exploits a specially proposed CNN architecture) as the CLSA for fair comparison.

Table 4 shows that the proposed CLSA is more effective than other multi-scale learning algorithms in person search. In particular, we have these observations: **(1)** The in-network feature pyramid decreases the overall performance as compared to using the standard ResNet-50 features (no pyramid) by a margin of 1.4% (82.5-81.1) in Rank-1 and 1.4% (81.6-80.2) in mAP. This verifies our hypothesis that directly applying the CNN feature hierarchy may harm the model performance due to the intrinsic semantic discrepancy across different pyramid levels. **(2)** CLSA improves the baseline in-network feature pyramid by a gain of 7.4% (88.5-81.1) in Rank-1 and 7.0% (87.2-80.2) in mAP. This indicates the exact effectiveness of the proposed cross-level semantic alignment mechanism in enhancing the person identity matching capability of the CNN feature representation in an end-to-end learning manner. **(3)** Three ResNet-50 based competitors all bring about person search performance improvement although less significant than the CLSA. This collectively suggests the importance of addressing the multi-scale matching problem in person search. **(4)** For model computational efficiency in FLOPs (FLoating point OPerations) per bounding box, CLSA has the least (a marginal) cost increase compared to other state-of-the-art multi-scale learning methods. This shows the superior cost-effectiveness of CLSA over alternative methods in addition to its accuracy advantages.

4.3 Further Analysis and Discussions

Effect of Person Detection We analysed the effect of person detection on the person search performance using the CUHK-SYSU benchmark. We started

with the three customised components of Faster-RCNN (Sec 3.1). Table 5 shows that: **(1)** The region proposal resizing and max-pool operation does not hurt the model performance. In effect, this is a replacement of ROI pooling. In the context of an average pooling to 1×1 feature map followed, such a design remains the capability of detecting small objects therefore imposing no negative effect. **(2)** Freezing the first block's parameters in fine-tuning detector helps due to the commonality of source and target domain data in low-level feature patterns. **(3)** Using all sized proposals improves the result. It is worthy noting this does not reduce the model efficiency, because only top 256 boxes per image are remained after the Non-Maximum Suppression operation, similar to the conventional case of selecting larger proposals. There are an average of 6.04 bounding boxes per image on CUHK-SYSU.

Table 5. Detection model component analysis on CUHK-SYSU.

	Full		No resize&max-pool		Not fix 1st block		Not all sized proposals	
Metric (%)	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CLSA	**88.5**	87.2	88.3	**87.3**	87.7	86.8	87.9	86.9

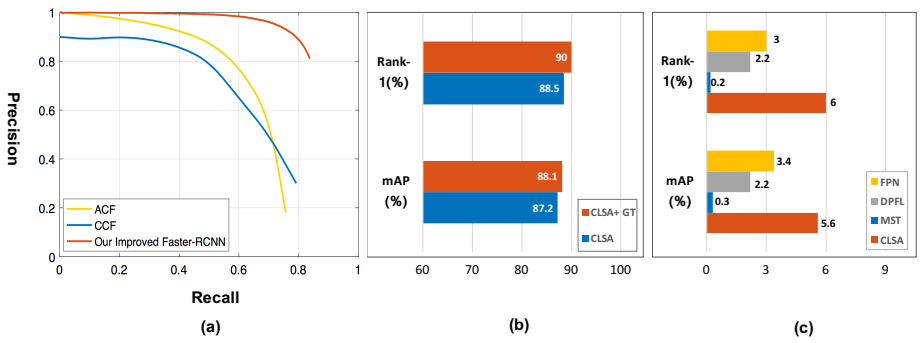

Fig. 6. Evaluation of person detection on CUHK-SYSU in the standard 100 sized gallery setting. **(a)** Person detection precision-recall performance. **(b)** The person search performance of the CLSA based on auto-detected *or* ground-truth person bounding box images. **(c)** Person detection *versus* multi-scale learning on the effect of person search performance.

We then evaluated the holistic person detection performance with comparison to other two detection models (ACF [9] and CCF [42]). For person detection, it is shown in Fig. 6 (a) that the precision performance of both ACF and CCF drops quickly when increasing the recall rate, whilst our improved Faster-RCNN remains more stable. This shows the effectiveness of deep learning detectors along additional model improvement from our CLSA. This is consistent with

the results in Table 2 and Table 3 that the CLSA outperforms ACF or CCF based methods by 20+% in both rank-1 and mAP.

We further tested the person search effect of our detection model by comparing with the results based on ground-truth bounding boxes. It is found in Fig. 6 (b) with perfect person detection, the CLSA gives only a gain of 0.9% (88.1-87.2) in mAP and 1.5% (90.0-88.5) in Rank-1. This indicates that the person detection component is not necessarily a major performance bottleneck in person search, thanks to modern object detection models. On the other hand, Table 4 also shows that addressing the multi-scale challenge is more critical for the overall model performance on person search, e.g. CLSA brings a performance boost of 6.0% (88.5-82.5%) in Rank-1 and 5.6% (87.2-81.6) in mAP over the baseline network ResNet-50.

Effect of Feature Pyramid We evaluated the performance effect of feature pyramid of CLSA on CUHK-SYSU. Recall that the in-network feature pyramid construction is based on the selection of ResNet blocks (see Sec. 3.2 and Fig. 3). We tested three block selection schemes: 5-4, 5-4-3 (used in the final CLSA solution), and 5-4-3-2. Table 6 shows that a three-level pyramid is the optimal. It also suggests that performing semantic alignment directly with elementary features such as those extracted from the Res2X block may degrade the overall representation benefit in the pyramid, due to the hard-to-bridge semantic gap.

Table 6. Effect of in-network feature pyramid construction on CUHK-SYSU.

Blocks Selection	5-4	5-4-3	5-4-3-2
Rank-1 (%)	87.3	**88.5**	85.3
mAP (%)	86.2	**87.2**	84.3

Effect of Temperature Softness We evaluated the impact of the temperature parameter setting in Eq. (3) in the range from 1 to 7. Table 7 shows that this parameter is not sensitive with the best value as 3.

Table 7. Effect of temperature softness (Eq. (3)) on CUHK-SYSU.

Temperature T	1	3	5	7
Rank-1 (%)	88.3	**88.5**	88.3	88.1
mAP (%)	87.0	87.2	**87.3**	86.9

Evaluating Person Re-ID and Object Classification We evaluated the effect of CLSA on person re-id (Market1501 [46], CUHK03 [23]) and object image classification (CIFAR100 [19]), in comparison to ResNet-50. Table 8 shows the positive performance gains of our CLSA method on both tasks. For example, the CLSA improves person re-id by 3.5%(88.9-85.4) in Rank-1 and 4.5% (73.1-68.6) in mAP on Market-1501. This gain is smaller than that on the same source video

based PRW (see Table 3), due to the potential reason that person bounding boxes of Market-1501 have been manually processed with limited and artificial scale variations. Moreover, our method also benefits the CIFAR object classification with a 1.5% (76.2-74.7) top-1 rate gain. These observations suggest the consistent and problem-general advantages of our model in addition to person search in unconstrained scene images.

Table 8. Evaluating the CLSA on re-id and object classification benchmarks.

Dataset	Market-1501 [46]		CUHK03 [23]		Dataset	CIFAR100 [19]
Metric (%)	Rank-1	mAP	Rank-1	mAP	Metric (%)	Top-1 rate
ResNet-50	85.4	68.6	48.8	47.5	ResNet-110	74.7
CLSA	**88.9**	**73.1**	**52.3**	**50.9**	CLSA	**76.2**

5 Conclusion

In this work, we present a novel *Cross-Level Semantic Alignment* (CLSA) deep learning framework for person search in unconstrained scene images. In contrast to existing person search methods that focus on improving the people detection performance, our experiments show that solving the multi-scale matching challenge is instead more significant for improving the person search results. To solve this under-studied cross-scale person search challenge, we propose an end-to-end CLSA deep learning method by constructing an in-network feature pyramid structural representation and enhancing its representational power with a semantic alignment learning loss function. This is designed specially to make all feature pyramidal levels identity discriminative therefore leading to a more effective hierarchical representation for matching person images with large and unconstrained scale variations. Extensive comparative evaluations have been conducted on two large person search benchmarking datasets CUHK-SYSU and PRW. The results validate the performance superiority and advantages of the proposed CLSA model over a variety of state-of-the-art person search, person re-id and multi-scale learning methods. We also provide comprehensive in-depth CLSA component evaluation and analysis to give the insights on model performance gain and design considerations. In addition, we further validate the more general performance advantages of the CLSA method on the person re-identification and object categorisation tasks.

Acknowledgements

This work was partly supported by the China Scholarship Council, Vision Semantics Limited, the Royal Society Newton Advanced Fellowship Programme (NA150459), and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

1. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., Ogden, J.M.: Pyramid methods in image processing. RCA Engineer **29**(6), 33–41 (1984)
2. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, p. 2 (2018)
3. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2 (2017)
4. Chen, X., Gupta, A.: An implementation of faster rcnn with study for region sampling. arXiv (2017)
5. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: Workshop of IEEE International Conference on Computer Vision. pp. 2590–2600 (2017)
6. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 392–408 (2018)
7. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1335–1344 (2016)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
9. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(8), 1532–1545 (2014)
10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1627–1645 (2010)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587 (2014)
12. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person re-identification. Springer (2014)
13. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European Conference on Computer Vision. pp. 346–361 (2014)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
15. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv (2017)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv (2015)
17. Jiao, J., Zheng, W.S., Wu, A., Zhu, X., Gong, S.: Deep low-resolution person re-identification. In: AAAI Conference on Artificial Intelligence (2018)
18. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2288–2295 (2012)

19. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
20. Lan, X., Wang, H., Gong, S., Zhu, X.: Deep reinforcement learning attention selection for person re-identification. arXiv (2017)
21. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
22. Li, M., Zhu, X., Gong, S.: Unsupervised person re-identification by deep learning tracklet association. In: European Conference on Computer Vision (2018)
23. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014)
24. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. arXiv (2017)
25. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
26. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE International Conference on Computer Vision. pp. 2197–2206 (2015)
27. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
28. Liu, H., Feng, J., Jie, Z., Jayashree, K., Zhao, B., Qi, M., Jiang, J., Yan, S.: Neural person search machines. In: IEEE International Conference on Computer Vision (2017)
29. Liu, J., Zha, Z.J., Tian, Q., Liu, D., Yao, T., Ling, Q., Mei, T.: Multi-scale triplet cnn for person re-identification. In: Proceedings of the 2016 ACM on Multimedia Conference. pp. 192–196. ACM (2016)
30. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv (2015)
31. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
32. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: Advances in Neural Information Processing Systems. pp. 424–432 (2014)
33. Paszke, A., Gross, S., Chintala, S., Chanan, G.: Pytorch (2017)
34. Qian, X., Fu, Y., Jiang, Y.G., Xiang, T., Xue, X.: Multi-scale deep learning architectures for person re-identification. In: IEEE International Conference on Computer Vision (2017)
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
36. Wang, H., Gong, S., Zhu, X., Xiang, T.: Human-in-the-loop person re-identification. In: European Conference on Computer Vision. pp. 405–422 (2016)
37. Wang, H., Zhu, X., Gong, S., Xiang, T.: Person re-identification in identity regression space. International Journal of Computer Vision (2018)
38. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
39. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: European Conference on Computer Vision. pp. 688–703 (2014)

40. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1249–1258 (2016)
41. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3376–3385 (2017)
42. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: IEEE International Conference on Computer Vision. pp. 82–90 (2015)
43. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833 (2014)
44. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
45. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3586–3593 (2013)
46. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision. pp. 1116–1124 (2015)
47. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Tian, Q.: Person re-identification in the wild. arXiv (2017)
48. Zhu, X., Wu, B., Huang, D., Zheng, W.S.: Fast openworld person re-identification. IEEE Transactions on Image Processing (2017)