Query-guided End-to-End Person Search

Bharti Munjal[1,2] Sikandar Amin[1] Federico Tombari[2] Fabio Galasso[1]
[1]OSRAM GmbH, [2]Technische Universität München

Abstract

Person search has recently gained attention as the novel task of finding a person, provided as a cropped sample, from a gallery of non-cropped images, whereby several other people are also visible. We believe that ***i.*** *person detection and re-identification should be pursued in a joint optimization framework and that* ***ii.*** *the person search should leverage the query image extensively (e.g. emphasizing unique query patterns). However, so far, no prior art realizes this.*

We introduce a novel query-guided end-to-end person search *network (QEEPS) to address both aspects. We leverage a most recent joint detector and re-identification work, OIM [37]. We extend this with* ***i.*** *a query-guided Siamese squeeze-and-excitation network (QSSE-Net) that uses global context from both the query and gallery images,* ***ii.*** *a query-guided region proposal network (QRPN) to produce query-relevant proposals, and* ***iii.*** *a query-guided similarity subnetwork (QSimNet), to learn a query-guided re-identification score. QEEPS is the first end-to-end query-guided detection and re-id network. On both the most recent CUHK-SYSU [37] and PRW [46] datasets, we outperform the previous state-of-the-art by a large margin.*

1. Introduction

Person search has recently emerged as the task of finding a person, provided as a cropped exemplar, in a gallery of non-cropped images [23, 37, 39, 46]. Person search is challenging, since the gallery contains cluttered background (including additional people) and occlusion. Furthermore, the query person may appear in the gallery under different viewpoints, poses, scale and illumination conditions. However the task is of great relevance in video surveillance, since it enables cross-camera visual tracking [3] and person verification [40].

Typical approaches to person search separate the problem into person localization (detection) and re-identification (re-id), and tackle each task sequentially via separate supervised networks. One such approach is the current best performer Mask-G [4]. But when separating the two tasks, one may remove useful contextual information for the re-id network (e.g. in Fig.1*(left)* the query is cut out from the query image). Also, the detection and localization task cannot exploit the information from the query, since the detection network runs independently, before the re-id network.

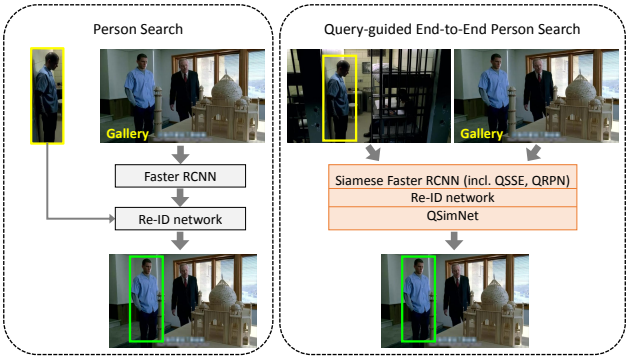


Figure 1. Person search is about finding a query person (yellow rectangle) within a gallery image (the target green rectangle). (*Left*) current approaches detect all people from the gallery image, then estimate re-identification features to match the cropped query. (*Right*) our proposed QEEPS guides the person search with an extensive use of the full query image, by means of a joint detection and re-identification network, which is end-to-end trained.

If we, as humans, were to search for a person in an image, we would not only look at each individual, but we would also search for peculiar patterns in the image, *e.g.* a distinct color or the texture of the person's t-shirt, as an additional hints. Motivated by this perspective, we introduce the first *Query-guided End-to-End Person Search* work (QEEPS). We propose the joint optimization of detection and re-identification, and to condition both aspects on the given query image, as exemplified in Fig.1*(right)*.

Our approach is the only method which is both end-to-end and query-guided. To the best of our knowledge, across the person search approaches [23, 37, 39, 42, 46], only OIM [37] and IAN [35] optimize jointly the detector and the re-identification networks (end-to-end). On the other hand, NPSM [23] is the sole to adopt a query attention mechanism, by replacing the detector RPN with an iterative query-guided search based on Conv-LSTM. In fact NPSM builds on OIM but it is not end-to-end, since it freezes the detector and re-identification network parts to pre-trained values,

i.e. its re-identification score (used for matching) does not change from the original OIM value.

We are inspired by OIM to design a model encompassing a detector with an additional re-identification branch. As in OIM, we optimize the networks jointly by adopting an OIM loss function [37]. Our model additionally features a Query-guided Siamese Squeeze-and-Excitation Network (QSSE-Net), a Query-guided RPN (QRPN) and a Query-Similarity Network (QSimNet). The QSSE-Net extends the recent squeeze-and-excitation (SE) technique to re-calibrate the channel-wise Siamese feature responses by the (global, image-level) inter-dependencies of the query and gallery channels [15]. The QRPN complements the parallel RPN with query-specific proposals. It employs a modified SE block to emphasize spatial features (in addition to feature channels), particularly bringing up the query-specific discriminant ones. QSimNet takes the query and gallery image proposal re-id features and provides a query-guided re-id score. When added to the baseline, QSimNet alone improves the mAP and CMC top-1 (hereinafter referred to as top-1) performances by as much as 7.1pp and 4.3pp on the CUHK-SYSU dataset [37] (gallery size of 100).

Altogether, QEEPS sets a novel state-of-the-art performance of 88.9% mAP and 89.1% top-1 on the CUHK-SYSU dataset [37], outperforming the prior best performer Mask-G [4] by 5.9pp mAP and 5.4pp top-1. Similarly, on the PRW dataset [46], QEEPS outperforms Mask-G [4] by 4.5pp mAP and 4.6pp top-1 setting state-of-the-art performance of 37.1% mAP and 76.7% top-1.

We summarize our contributions: **i.** we introduce the first query-guided end-to-end person search (QEEPS) network; **ii.** we propose a query-guided Siamese squeeze-and-excitation (QSSE) block that extends the interaction between feature channels to additionally model the global similarities between the query and gallery image pairs; **iii.** we define a novel query-guided RPN (QRPN), by extending the SE-Net squeeze-and-excitation block to the query channels and spatial features; **iv.** we define a novel query-similarity subnetwork (QSimNet) to learn a query-guided re-identification score; **v.** we achieve a new state-of-the-art performance on CUHK-SYSU [37] and PRW [46] datasets.

2. Related Work

In this section we first review prior art on the two separate tasks of person detection and person re-identification. Then we review literature on person search

Person Detection. In the past few decades, this field has witnessed steep improvement with the introduction of boosting [32], deformable parts models [10] and aggregate channel features [8]. Convolutional neural networks (CNNs) excel today at this task thanks to jointly learning the classification model and the features [29], in an end-to-end fashion. While single-stage object detectors [22, 25, 28] are preferable for runtime performance, the two-stage strategy of Faster R-CNN remains the more robust general solution [13], versatile to tailor region proposals to custom scene geometries [2] and to add multi-task branches [12, 37]. As in OIM [37], we adopt Faster R-CNN with a ResNet [14] backbone.

Person Re-Identification. Classic approaches for person re-identification have focused on manual feature design [33, 11, 9, 43] and metric learning [20, 44, 17, 19, 21, 27, 26]. As in object detection, CNNs have recently conquered the scene in re-identification, too [1, 18, 41].

While modern CNN approaches target the estimation of a re-id embedding space (whereby the same IDs lie close and further from other individuals), there are two main trends in the model learning: **i.** by Siamese networks and contrastive losses; and **ii.** by ID classification with cross-entropy losses. In the first, pairs [1, 18, 24, 31, 38, 41], triplets [6, 7] or quadruplets [5] are used to learn a corresponding number of Siamese networks, by pushing or pulling the same or the different person ids, respectively. In the second, [36, 45] define as many classes as people IDs, train classifiers with a cross-entropy loss, and take the network features as the embedding metric during inference. Best performing person search approaches [23, 37] follow this second trend, which we also adopt.

Our work also relates to the similarity-guided graph neural network of [30]. They learn the similarity among multiple query and gallery identities and use it to construct a graph, as opposed to a fixed metric, such as the cosine similarity in OIM [37]. Here we learn the similarity but do not adopt graphs, thus preserving a convenient runtime.

Person Search. The pioneering work of Xu *et al*. [39] introduces person search as re-identifying people within gallery images, where they also have to be detected and localized. The adoption of CNNs in person search is enabled by the introduction of two recent person search datasets, PRW [46] and CUHK-SYSU [37]. Initial approaches [39, 46] use separate pre-trained people detectors and only learn re-identification networks. Interestingly, the most recent work to date [4] states that detection and re-identification should be addressed separately for best performance. We contrast this statement by showing that our single end-to-end network yields better performance.

End-to-End Person Search. Xiao *et al*. [37] introduces the first end-to-end person detection and re-identification network. They propose an Online Instance Matching (OIM) loss to address the challenge of training a classifier matrix for an overwhelming number of person IDs (thousands of different people), as required for both the CUHK-SYSU [37] and PRW [46] person search datasets. In other words, they build-up a matrix look-up table by leveraging the IDs in each mini-batch at training, instead of learning ID-specific classifiers. The look-up tables are updated dur-

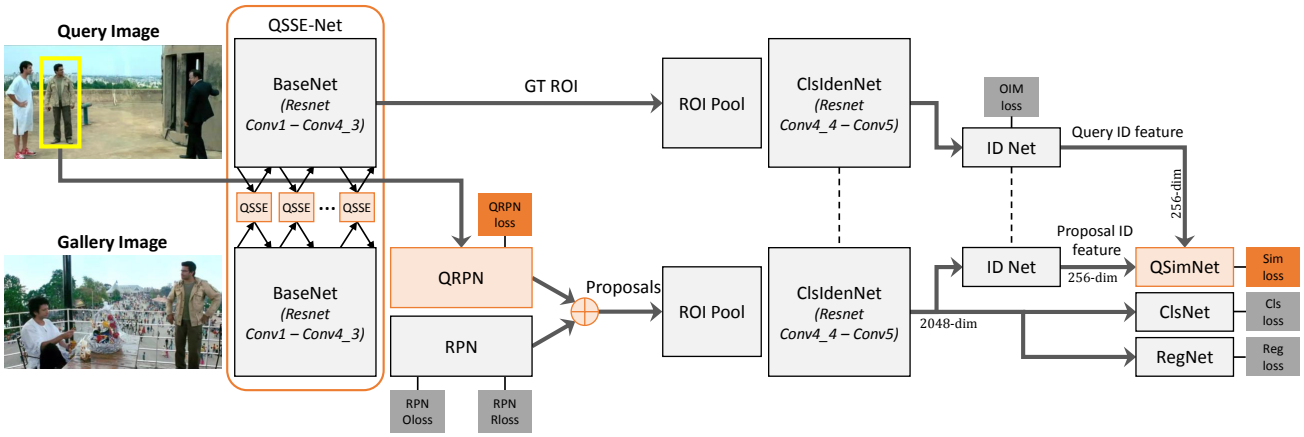


Figure 2. Our proposed QEEPS network architecture. We pair the reference OIM [37] *bottom network* with a novel Siamese *top network*, to process the query and guide the person search at different levels of supervision (cf. Sec. 4). The novel query-guidance blocks of our approach, displayed in orange, are trained end-to-end with the whole network with specific loss functions (*darker orange boxes*).

ing training by running averages and allow for employing a soft-max loss in the ID Net training with a limited number of IDs. More recently, [35] extends the OIM with an additional center loss [34], which improves the intra-class feature compactness. To our knowledge, the OIM loss is currently best for optimizing the joint network, adopted by most recent work [23, 35], including ours.

Query-guided person search. To the best of our knowledge, the NPSM approach of Liu *et al*. [23] is the sole to exploit the query image. They do so by instantiating an iterative person search mechanism based on a Conv-LSTM, which re-weights attention across a number of pre-defined person detection proposals. NPSM builds upon Faster-R-CNN and OIM, but replaces the traditional RPN with the attention mechanism. We note that both the base Faster-R-CNN network and the re-identification head are pre-trained as from [37] and frozen. This implies that, upon the query-guided attention search, the final re-id score remains the same as in [37], not profiting from the proposal adjustment. We adopt the same Faster-R-CNN network and OIM loss, but optimize those end-to-end, alongside our novel query-guided proposal network.

3. Background - Online Instance Matching

We leverage the end-to-end person search architecture of [37], which we refer to as **OIM** hereinafter, since it introduces the Online Instance Matching, key to the joint detection and re-identification optimization (cf. Sec. 2).

We illustrate the base architecture of [37] in Fig. 2 (*gray blocks* from the *bottom network*, applied to the gallery image). The OIM network consists of a Faster R-CNN [29] with a ResNet backbone [14] (this accounts for the blocks BaseNet, RPN, ROI Pool and ClsIdenNet in Fig. 2). In parallel to the classification (ClsNet) and regression (RegNet) branches, [37] defines an ID Net, which provides a re-identification feature embedding, supposedly unique for the same identities but different for other people. Then they adopt cosine similarity to match cropped query identities to the estimated id embeddings from the gallery image.

4. Query-guided Person Search

Fig. 2 illustrates our proposed architecture. In more detail, we pair the OIM network [37], originally employed for the gallery image, with a second Siamese network (*top network* in Fig. 2), applied to the query image. The query network shares weights with the gallery image network. Features from the Siamese query network are used to guide the gallery image network (*bottom network* in Fig. 2) at different levels of supervision (novel query-guidance blocks are represented in *orange*).

In more details, we introduce 3 novel subnetworks: **i.** a Query-guided Siamese Squeeze-and-Excitation Network (QSSE-Net) that leverages *global* contextual information from both the query and gallery images to re-weight the feature channels; **ii.** a Query-guided Region Proposal Network (QRPN), leveraging query-ROI-Pooled features to emphasize discriminant patterns in the gallery image to produce relevant proposals; and **iii.** a Query-guided Similarity Network (QSimNet) for computing the re-identification (re-id) jointly from the query and gallery image crop features.

Note that QSSE-Net processes the full query and gallery images and considers therefore a *global* context (e.g. if one of the two images is very dark, channels expressing shape are likely to be more discriminant than those encoding color). On the other hand, QRPN and QSimNet are *local*, since they consider the person crop, and dedicated to emphasize features specific to each individual, as defined by the pair [query-gallery] image crop.

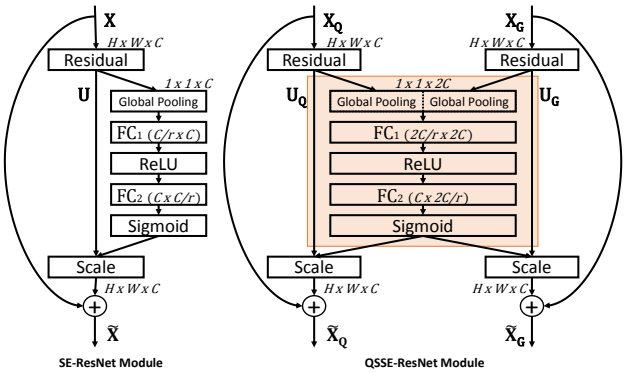


Figure 3. Our proposed Query-guided Siamese Squeeze-and-Excitation Network (QSSE-Net). QSSE-Net is integrated into the ResNet base network. It concatenates the query and gallery features, upon the residual blocks. It applies then squeeze-and-excitation [15], and re-calibrates the query and the gallery image channels according to intra- and inter-channel dependencies.

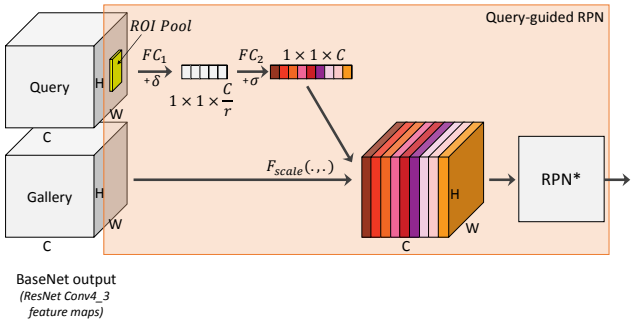


Figure 4. Our proposed Query-guided Region Proposal Network (QRPN). Based on the query guidance, QRPN adopts a modified squeeze-and-excitation net to re-calibrate the gallery image feature responses, which are then passed to a standard RPN. (*) indicates that this RPN does not compute regression offsets.

4.1. Query-guided Siamese Squeeze-and-Excitation Network (QSSE-Net)

The QSSE-Net block is integrated into the ResNet base network. QSSE-Net is inspired by the recent squeeze-and-excitation network (SE-Net) [15], the main difference being the extension to a Siamese-like model which includes both the query and the gallery, as illustrated in Fig. 3. The very recent Mask-G [4] also utilizes squeeze-and-excitation block in their pipeline to re-weight the feature channels.

As proposed in [15], a QSSE block performs two operations, namely *squeeze* and *excitation*, i.e. compute a weight vector and re-weight the feature maps per channel. The squeeze operation condenses the spatial information of each of the C channels of both query and gallery by global average pooling, resulting in channel-descriptors $\mathbf{z}_q$ and $\mathbf{z}_g \in \mathbb{R}^C$, respectively.

The excitation operation applies a non-linear bottleneck function of two fully-connected layers using concatenated query and gallery channel-descriptors $[\mathbf{z}_q, \mathbf{z}_g] \in \mathbb{R}^{2C}$. The first FC_1 layer reduces the dimensionality $2C$ by a factor of r, to obtain $\frac{2C}{r}$ channels. The second layer FC_2 re-expands those to C followed by a sigmoid activation σ. This results in the weight vector $\mathrm{s} \in \mathbb{R}^C$ being as follows

$$\mathbf{s} = \mathrm{F}_{ex}(\mathbf{z}_q, \mathbf{z}_g; \mathbf{W}) = \sigma(\ \mathbf{W}_2\ \delta(\ \mathbf{W}_1[\mathbf{z}_q, \mathbf{z}_g]\ )\ ) \qquad (1)$$

whereby, $\mathbf{W}_1 \in \mathbb{R}^{\frac{2C}{r} \mathrm{x} 2C}$ are the parameters of the first fully-connected layer placed for dimensionality-reduction, while the second fully-connected layer with parameters $\mathbf{W}_2 \in \mathbb{R}^{C \mathrm{x} \frac{2C}{r}}$ is for dimensionality-expansion. The reduction ratio r is set to 16 in all our experiments as proposed in [15]. We refer to δ as the ReLU non-linearity that models nonlinear interactions between channels.

As shown in Fig. 3 (blocks *"Scale"* and *skip connections*), the outputs of a QSSE-ResNet block $\widetilde{\mathbf{X}}_{\mathrm{Q}}$ and $\widetilde{\mathbf{X}}_{\mathrm{G}}$ for the respective query and gallery images are:

$$\begin{aligned} \widetilde{\mathbf{X}}_{\mathrm{Q}} &= \mathbf{X_Q} + \mathbf{s} \odot \mathbf{U_Q} \\ \widetilde{\mathbf{X}}_{\mathrm{G}} &= \mathbf{X_G} + \mathbf{s} \odot \mathbf{U_G} \end{aligned} \qquad (2)$$

where $\odot$ means channel-wise multiplication, re-weighting the residual outputs $\mathbf{U_Q}$ and $\mathbf{U_G}$. We connect a QSSE block to each ResNet block within the BaseNet (cf. Fig. 2).

Note that, differently from [15], our QSSE-Net concatenates globally-average-pooled features from the query and the gallery networks, and then uses the channel excitation to re-weight both of them. In this way, QSSE-Net re-calibrates channel weights to take into account *intra-network* channel dependencies and *inter-network* channel similarities.

4.2. Query-guided RPN (QRPN)

Our proposed Query-guided Region Proposal Network (QRPN) re-weights the BaseNet feature responses of the gallery image by means of the cropped query features. As illustrated in Fig. 4, QRPN includes a channel-wise attention mechanism (query guidance) and a standard RPN [29], extracting the proposal boxes from the gallery re-weighted image features. The novel query guidance is inspired by the SE block of [15] but it features some key differences. As in [15], we adopt a bottleneck design with two fully-connected layers, FC_1 and FC_2, which squeeze and expand the features, so as to highlight important signal correlations. The reduction ratio r is set to 16 as in [15]. The resulting weights (excitations), upon the sigmoid activation σ, are applied to the gallery BaseNet feature maps per channel (channel-wise multiplication).

Here for the first time, we apply the SE idea to the pooled feature maps of the query crop. In more details, first we pool the query crop feature map by ROI Pool. Then we apply FC_1 to all channels and all pixels of the pooled map

(*i.e.*, not just to the channels). Finally, the excitations are applied to the gallery image features, not to the own query features. Our query guidance may therefore emphasize specific gallery channels, based on *local* (spatially-localized) channel-wise query patterns.

Our proposed QRPN complements the standard (query-agnostic) RPN (cf. the parallel QRPN and RPN in Fig. 2). QRPN extracts proposal boxes featuring a *query-similarity* score, while RPN pursues the standard *objectness* score. Notably, QRPN includes an RPN with the same design as the parallel standard RPN, *e.g.* the same anchor boxes. As illustrated in Sec. 5, we obtain the best performance by simply summing up the scores from the QRPN and the RPN for each anchor. The usual non-maximum-suppression (NMS) is finally applied on the resulting score, while we adopt the regression offsets of RPN, thus query- and class-agnostic.

4.3. Query-guided Similarity Net (QSimNet)

The baseline OIM network [37] compares the re-identification features from the query and the gallery image crops by means of cosine similarity. In other words, re-id features are computed for the query and gallery image crops independently, and then used to retrieve the query individual by matching. We maintain that the similarity score should depend on the specific query re-id features and be end-to-end trainable, so the network could emphasize and tailor the similarity metric for each query (e.g. balancing color, shape and other attributes for each specific person).

As illustrated in Fig. 5, we propose a simple query-guided similarity subnetwork (QSimNet) to compare the re-id features of the query against the gallery image crops. Upon the L2 distance (element-wise subtraction and square) of the re-id features, we apply batch normalization [16] and a fully connected layer, followed by softmax. QSimNet is learned end-to-end with the rest of the network. At inference time, we use its output scores to perform non-maximum suppression (NMS) for the final matches of the query probe in the gallery image. We do not therefore use the classification scores from the original detection network, ClsNet in Fig. 2, but ClsNet is used for training the detector branch and to remove the least-confident person detections during inference, with score $< 10^{-2}$.

4.4. End-to-end Joint Optimization

We jointly optimize, in an end-to-end fashion: **i.** the person detection network searching people in the gallery image; **ii.** the identification network for learning a discriminative feature embedding per ID in the training data; and **iii.** the novel query-guided subnetworks QSSE-Net, QRPN and QSimNet. We pursue the joint optimization by means of loss functions for each task, represented in Fig. 2 as the darker (gray or orange) loss boxes. In more details, the Faster R-CNN detector is supervised with loss functions for classification (L_{cls}), regression (L_{reg}), RPN objectness (L_{rpn_o}), and RPN box regression (L_{rpn_r}); while the identification subnetwork is supervised by the OIM loss (L_{oim}) [37]. We introduce two new loss functions, the QRPN loss (L_{qrpn}) and the Sim loss (L_{sim}), to directly supervise QRPN and QSimNet, while the QSSE-Nets gets the same implicit supervision as the BaseNets. The overall objective is given by:

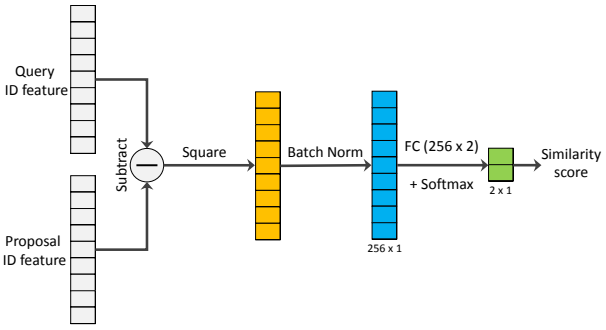


Figure 5. Our proposed Query-guided Similarity Network (QSimNet). QSimNet introduces a simple query-guided net to estimate the similarity between the query and the gallery image. This is learned end-to-end with the rest of the network.

$$L = \lambda_1 L_{cls} + \lambda_2 L_{reg} + \lambda_3 L_{rpn_o} + \lambda_4 L_{rpn_r} + \lambda_5 L_{oim} + \lambda_6 L_{qrpn} + \lambda_7 L_{sim} \quad (3)$$

whereby λ_{1-7}, responsible for the relative loss importance, are here all set to 1.

Siamese Design. Note that a query-guided person search network implies passing both the query and the gallery images through the Siamese network. Finding a person within a gallery image is still a fast operation (our Pytorch implementation runs in 300msec on an Nvidia P6000), as it only requires a forward pass through the network. But comparing a query against a gallery image set needs re-running the query at all times and, during training, it requires storing the intermediate features and the gradients for each image pair. Our current batch only contains a single query-gallery pair, but this does not affect performance (cf. Sec.5).

QRPN loss. Similarly to the RPN loss [29], we define L_{qrpn} as a cross-entropy loss:

$$L_{qrpn} = -\frac{1}{N} \sum_{N} \log(p_n^u) \quad (4)$$

whereby N is the size of the mini-batch, p_n^u is the probability of the assigned true class u for the n^{th} anchor box in the mini-batch. Anchor boxes that overlap with the query individual are marked as positives. We set on purpose to not sample negatives from the other people present in the gallery image, to avoid setting diverging objectives for the parallel QRPN and RPN (since other people in the gallery image are positives for the standard RPN).

Method	Gallery Size 50		Gallery Size 100		
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)	Max-recall (%)
OIM [37]	80.0	-	75.5	78.7	-
+ *QRPN*	82.1	82.7	79.6	80.4	96.6
+ *QSimNet*	85.1	85.6	82.6	83.0	98.1
+ *QRPN* + *QSimNet*	86.2	86.7	83.1	83.3	98.2
+ *QSSE* + *QRPN* + *QSimNet* (= QEEPS)	**87.0**	**87.1**	**84.4**	**84.4**	**98.8**

Table 1. Importance of each proposed model component, as evaluated on the CUHK-SYSU dataset [37], for gallery sizes of 50 and 100. OIM [37] results are reported from the original paper. The best performer OIM + QSSE + QRPN + QSimNet makes our proposed complete model, which we dub QEEPS. We also report the maximum detector recall, which depends on the subset of galleries containing the query.

Sim loss. We define L_{sim} as the binary cross-entropy loss function which maximizes the similarity score between the query crop and the corresponding individual in the gallery image proposals. L_{sim} takes similar form as Eq. (4).

Positive/negative ratio. To alleviate for the few positives in the mini-batch (since a gallery image contains at most one query id), we augment the data via jittering and relax the IoU overlap for the anchor box positive assignment to 0.6. On the other hand, we sample fewer negatives resulting in a mini-batch of size 128 instead of 256. We keep 0.3 as the maximum IoU overlap of anchor boxes for negative assignment as in standard RPN.

5. Experiments

5.1. Datasets and metrics

CUHK-SYSU. The CUHK-SYSU dataset [37] consists of 18,184 images, labeled with 8,432 identities and 96,143 pedestrian bounding boxes (23,430 boxes are ID labeled). The images, captured in urban areas by hand-held cameras or from movie snapshots, vary largely in viewpoint, lightning, occlusion and background conditions. We adopt the standard train/test split, where 11,206 images and 5,532 identities are used for training, 2,900 queries and overall 6,978 gallery images for testing. We experiment with gallery sizes of 50 and 100 for each query.

PRW. The PRW dataset [46], acquired in a university campus from six cameras, consists of 11,816 images with 43,110 bounding boxes (34,304 boxes are ID labeled) and 932 identities. Compared to CUHK-SYSU, PRW features less images and IDs but many more bounding boxes per id (36.8, against 2.8 in CUHK-SYSU), which makes it more challenging. The training set consists of 5,134 images with 482 identities, while the test set consists of 6,112 images (gallery size) with 450 identities and provides 2057 queries.

PRW-mini. The PRW test evaluation may become impractical for person search techniques which are query-based. In fact, conditioning on the query requires jointly processing each [query-gallery] pair and the exhaustive evaluation of the product space, *i.e.* $2,057 \times 6,112$ [1] .

We introduce the PRW-mini, which we publicly release[2], to reduce the evaluation time while maintaining the difficulty. PRW-mini tests 30 query images against the whole gallery. To maintain difficulty, we have sampled multiple sets of 30 query images and selected the one where the baseline OIM [37] performs at the same accuracy as in PRW (OIM [37] is a de facto baseline for most recent person search techniques [35, 23]).

Evaluation Metrics. We report two commonly adopted performance metrics [37, 35, 23]: mean Average Precision (mAP) and Common Matching Characteristic (CMC top-K) for evaluation. mAP is derived from the detection literature and reflects the accuracy in localizing the query in all gallery images (AP is computed for each ID and averaged to compute the mAP). 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). More specifically, we evaluate on the CUHK-SYSU dataset [37] by using the provided scripts, and we evaluate on PRW [46] with the same scripts as adopted by Mask-G [4], which we publicly provide[2].

5.2. Implementation Details

We build upon OIM [37] for the design, setup and pre-training of the base feature network and the network head (BaseNet and ClsIdenNet in Fig. 2), as well as for the ID-Net. As in [2], we adjust the anchor sizes to the objects in the dataset: we adopt scales $\{2, 4, 8, 16, 32\}$ and aspect ratios $\{1, 2, 3\}$. We adopt the same anchors for the RPN and QRPN. The input images are re-scaled such that their shorter side is 600 pixels. We pad or crop the query images to the same size of the gallery one. We train the whole network using SGD with momentum for 2 epochs, with a base learning rate of 0.001 which is reduced by a factor of 10 after the first epoch. For training, we consider all query-gallery image pairs for the CUHK-SYSU dataset, but we only use three gallery images per query for the PRW dataset (since this is already large). We augment the data by flipping both the query and the gallery image.

[1] By contrast, the baseline OIM [37] computes query and gallery re-id features separately and requires $2,057 + 6,112$ network forward passes.

[2] PRW-mini and the evaluation script (for PRW and PRW-mini) are at: `https://github.com/munjalbharti/Query-guided-End-to-End-Person-Search`

Method	mAP(%)	top-1 (%)
OIM [37] (Baseline)	75.5	78.7
IAN [35]	76.3	80.1
NPSM [23]	77.9	81.2
QEEPS (ours)	**84.4**	**84.4**
Mask-G [4]	83.0	83.7
OIM‡ (Baseline)	83.3	84.2
QEEPS (ours)	**88.9**	**89.1**

Table 2. Comparison with the state-of-the-art on the CUHK-SYSU dataset for the gallery size 100. Methods above the dashed line employ the standard Faster R-CNN image re-sizing to 600 pixels (shorter image side), while those below use larger images with shorter sides of 900 pixels. Note the strength of our baseline OIM‡, which is already above the state-of-the-art performance.

5.3. Ablation Study

As ablation study, we consider the OIM [37] baseline and evaluate the separate benefits of the proposed QRPN, QSimNet and QSSE-Net on the CUHK-SYSU dataset. In Table 1, we observe that adding QRPN to OIM provides 79.6% mAP and 80.4% top-1, improving mAP by 4.1pp and top-1 by 1.7pp, for a gallery size of 100. Adding QSimNet, we achieve an even higher improvement of 7.1pp mAP and 4.3pp top-1 for a gallery size of 100. Combining QRPN and QSimNet improves on both results (83.1% mAP and 83.3% top-1 for a gallery size of 100), demonstrating the complementary benefit of considering query guidance for the proposal generation and for the similarity score.

We achieve the best performance (84.4% mAP, 84.4% top-1 for a gallery size of 100) of our proposed QEEPS network (OIM+QSSE+QRPN+QSimNet) by integrating additionally the QSSE blocks into the Siamese BaseNets. Differently from QRPN and QSimNet, QSSE-Net acts on the channels and the feature maps of the entire images, not just the local crops, which provides complementary benefits. We consider the complete QEEPS in the next experiments.

In the last column of Table 1, we show a similar improvement trend for the detector recall (for a fixed number of region proposals of 300). Our full model QEEPS achieves a nearly-perfect recall of 98.8%, indicating the importance of query guidance also for learning higher-quality proposals.

5.4. Comparison to the State-of-the-art

CUHK-SYSU. In Table 2, we compare QEEPS to state-of-the-art approaches in person search [4, 23, 35] and to OIM [37]. It should be noted that OIM [37], IAN [35] and NPSM [23] build on Faster R-CNN and therefore presumably re-scale the images such that shorter side is 600 pixels. All three methods (shown above the *dashed* lines) argue for a joint detection and re-identification network. For this image resolution, our approach QEEPS achieves 84.4% mAP and 84.4% top-1, surpassing the state-of-the-art NPSM [23] by 6.5pp mAP and 3.2pp top-1 respectively.

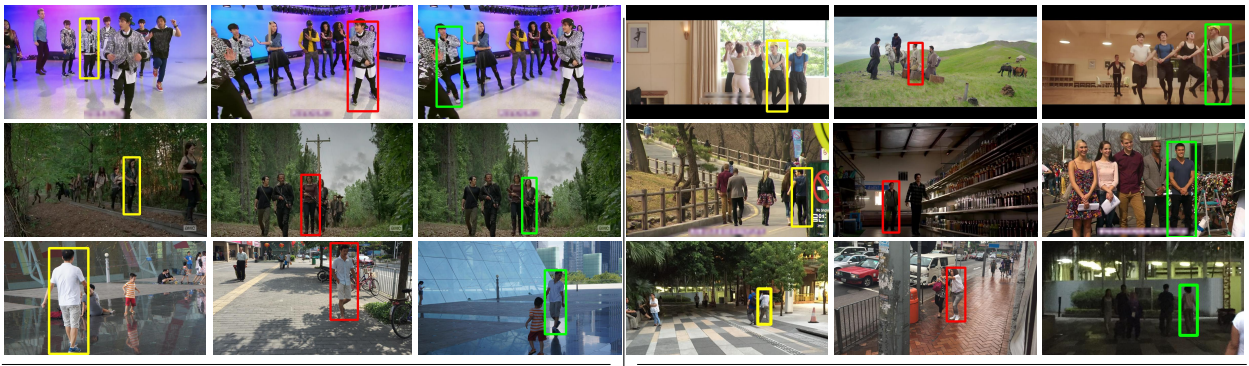
Method	mAP(%)	top-1 (%)
OIM [37]	21.3	49.9
IAN [35]	23.0	61.9
NPSM [23]	24.2	53.1
Mask-G [4]	32.6	72.1
OIM‡ (Baseline)	36.9	75.7
QEEPS (ours)	**37.1**	**76.7**
Mask-G [4]	33.1	70.0
OIM‡ (Baseline)	38.3	70.0
QEEPS (ours)	**39.1**	**80.0**

Table 3. Comparison with the state-of-the-art on the PRW dataset [46], above the dashed line, and on the proposed subset PRW-mini (cf. Sec. 5), below it.

Below the *dashed* line, Mask-G [4] considers a similar Faster R-CNN but with larger images (shoter side 900 pixels). In order to have a fair comparison with Mask-G [4], we consider the baseline OIM‡, same as OIM but with the input images re-scaled to a shorter side of 900 pixels, as long as the larger side be less than 1500 pixels. Mask-G argues that its better performance (83.0% mAP, 83.7% top-1 for a gallery of 100) is due to considering detection and re-identification independently. However, when run on the larger images, the strong baseline OIM‡ surpasses all prior art with a performance of 83.3% mAP and 84.2% top-1. We argue that this reasserts the validity of considering detection and re-identification jointly. On the same setup, our QEEPS achieves 88.9% mAP and 89.1% top-1, improving on best published results (Mask-G) by 5.9pp mAP and 5.4pp top-1. We attribute the further leap in performance to the proposed query guidance, both as a global and local cue.

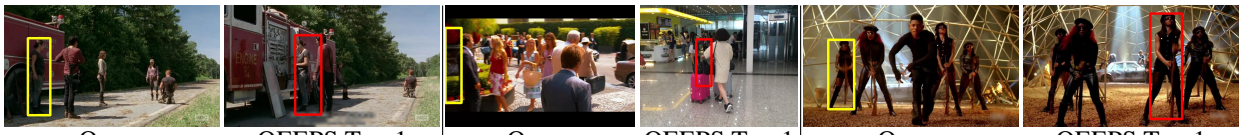
PRW/PRW-mini. In Table 3, we compare state-of-the-art techniques [4, 23, 35] to OIM [37], to the baseline OIM‡ and to the proposed QEEPS. Results above the dashed line refer to the full PRW dataset [46]. Note how Mask-G [4] (32.6% mAP, 72.1% top-1) and OIM‡ (36.9% mAP, 75.7% top-1) neatly surpass all other approaches. While one cannot draw a clear conclusion on the employed technique, it seems clear that processing larger input images yields a strong benefit, since these two methods are the only to re-scale them to short sides of 900 pixels, instead of 600. Finally, QEEPS outperforms the Mask-G by 4.5pp in mAP and 4.6pp in top-1, setting the novel state-of-the-art performance of 37.1% mAP and 76.7% top-1.

Below the dashed line, we report results on the PRW-mini subset, introduced in Sec. 5.1. Note how the ranking of OIM‡, Mask-G [4] and QEEPS is preserved in the PRW-mini as compared to PRW, and that all algorithms report similar mAP and top-1 performances (*e.g.* for Mask-G the gaps are only 0.5pp and 2.1pp respectively). PRW-mini maintains therefore the same difficulty as PRW, while reducing the evaluation time for query-based techniques by 2 orders of magnitude.



(a) Query (b) OIM‡ Top-1 (c) QEEPS Top-1 (a) Query (b) OIM‡ Top-1 (c) QEEPS Top-1

Figure 6. Qualitative *Top-1* person search results for a number of challenging query examples. For each example, we show (a) the query images with the bounding box of the query-person, in yellow, (b) their corresponding output matches given by the baseline OIM‡ , and (c) results of our proposed approach QEEPS. Notice that, our approach is able to fetch difficult gallery images as its first estimate. Red bounding boxes are failures, while green represent correct matches.



Query QEEPS Top-1 Query QEEPS Top-1 Query QEEPS Top-1

(a) (b) (c)

Figure 7. Observed trends in failure cases, such as, (a) localization errors of the person detector, (b) incorrect ground-truth annotations of bounding box and the person ID, and (c) extremely challenging examples due to similar appearance and/or low visibility.

Runtime Comparison. In Table 4, we report the time taken by Mask-G to process a gallery image (given a pre-processed query) compared to ours (processing both query and gallery images at all times). Since both methods use different GPUs, we report the TFLOPs too. Upon normalization with TFLOPs, ours is 3.14 times faster and requires 4.27 times less memory.

Method	# params (*M*)	Time (*sec*)	GPU (TFLOPs)
Mask-G [4]	209	1.3	K80 (8.7)
QEEPS	49	0.3	P6000 (12.0)

Table 4. Runtime comparison of QEEPS with Mask-G [4] for image size 900×1500.

5.5. Qualitative Results

As illustrated in Fig. 6, our approach performs person search successfully in a number of challenging scenarios, where the baseline method fails. For instance, in the top-left example, QEEPS retrieves the correct guy dancing in the line, while OIM‡ selects a different individual, but similarly dressed. Also quite convincingly, in the middle-right example, the query person is provided from the back (at low resolution) and found in a frontal-view gallery image. In the last row, we show two interesting examples depicting the importance of the global context. Notice how QEEPS compensates for global illumination changes (*e.g.* the blueish image in the bottom-left example) and retrieves the correct person.

We show in Fig. 7 the three most common failure cases. In column (a), the same person is retrieved successfully but wrongly localized (IoU < 0.5). In column (b), we illustrate an annotation mistake. Finally in column (c) the failure is most likely due the extreme difficulty of some examples on the CUHK-SYSU and PRW datasets, since several people look alike (illustrated) and some others have low-visibility issues. These may be challenging to a human, too.

6. Conclusions

We have proposed a novel QEEPS network, which jointly addresses detection and re-identification in an end-to-end fashion. Our results demonstrate that the joint consideration of detection and re-identification is a valid approach to person search, as it intuitively allows each separate module to account and to change with each other, during the joint training. Furthermore, the large and consistent improvement in performance provided by our proposed query guidance highlights the importance of this aspect. When searching for a person in a gallery image, we should consider the query for its global context (e.g. the overall illumination may shift the importance of color as a cue) and for its local cues (e.g. specific patterns which ease the creation of tailored proposals and better similarity scores).

Acknowledgements This research was partially funded by the BMWi – German Federal Ministry for Economic Affairs and Energy (MEC-View Project). We are grateful to the Mask-G team for their great support on the PRW evaluation.

References

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

[2] S. Amin and F. Galasso. Geometric proposals for faster r-cnn. *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017.

[3] J. O. N. Castañeda, A. Frias-Velazquez, N. B. Bo, M. Slembrouck, J. Guan, G. Debard, B. Vanrumste, T. Tuytelaars, and W. Philips. Scalable semi-automatic annotation for multi-camera person tracking. *IEEE Transactions on Image Processing*, 25:2259–2274, 2016.

[4] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai. Person search via a mask-guided two-stream cnn model. In *The European Conference on Computer Vision (ECCV)*, 2018.

[5] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48:2993–3003, 2015.

[8] P. Dollár, R. Appel, S. J. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36:1532–1545, 2014.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. *The IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[10] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32:1627–1645, 2009.

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *The European Conference on Computer Vision (ECCV)*, 2008.

[12] I. Hasan, T. Tsesmelis, F. Galasso, A. Del Bue, and M. Cristani. Tiny head pose classification by bodily cues. In *The IEEE International Conference on Image Processing (ICIP)*, 2017.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *The International Conference on Machine Learning (ICML)*, 2015.

[17] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[19] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[21] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[22] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[23] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan. Neural person search machines. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[24] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26:3492–3506, 2017.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *The European Conference on Computer Vision (ECCV)*, 2016.

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

[27] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[28] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[30] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang. Person re-identification with deep similarity-guided graph neural network. In *The European Conference on Computer Vision (ECCV)*, 2018.

[31] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *The European Conference on Computer Vision (ECCV)*, 2016.

[32] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[33] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *The IEEE International Conference on Computer Vision (ICCV)*, 2007.

[34] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *The European Conference on Computer Vision (ECCV)*, 2016.

[35] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng. IAN: the individual aggregation network for person search. *CoRR*, abs/1705.05552, 2017.

[36] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[38] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Y. Xu, B. Ma, R. Huang, and L. Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Association for Computing Machinery (ACM) International Conference on Multimedia (MM)*, 2014.

[40] Y. Yang, S. Liao, Z. Lei, and S. Z. Li. Large scale similarity learning using similar pairs for person verification. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.

[41] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. *22nd International Conference on Pattern Recognition (ICPR)*, pages 34–39, 2014.

[42] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[43] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[44] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by saliency learning. *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39 2:356–370, 2017.

[45] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *The European Conference on Computer Vision (ECCV)*, 2016.

[46] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.