Prototype-Guided Saliency Feature Learning for Person Search

Hanjae Kim[1], Sunghun Joung[1], Ig-Jae Kim[2], and Kwanghoon Sohn[1,*]
[1]Yonsei University [2]Korea Institute of Science and Technology (KIST)
{incohjk,sunghunjoung,khsohn}@yonsei.ac.kr, drjay@kist.re.kr

Abstract

Existing person search methods integrate person detection and re-identification (re-ID) module into a unified system. Though promising results have been achieved, the misalignment problem, which commonly occurs in person search, limits the discriminative feature representation for re-ID. To overcome this limitation, we introduce a novel framework to learn the discriminative representation by utilizing prototype in OIM loss. Unlike conventional methods using prototype as a representation of person identity, we utilize it as guidance to allow the attention network to consistently highlight multiple instances across different poses. Moreover, we propose a new prototype update scheme with adaptive momentum to increase the discriminative ability across different instances. Extensive ablation experiments demonstrate that our method can significantly enhance the feature discriminative power, outperforming the state-of-the-art results on two person search benchmarks including CUHK-SYSU and PRW.

1. Introduction

Person search aims to localize a target person in a gallery of pedestrian images. While inputs of person re-identification (re-ID) task are auto-detected person bounding boxes [15], those of person search task are in-the-wild image containing large amounts of backgrounds. The simplest way to solve the person search problem is to crop all the pedestrians adopting off-the-shelf detector [21], and pass them into re-ID module as [1]. This, however, divide the two tasks which affect each other into separate tasks and are inefficient in real-world applications. To address this, recent methods [30, 18, 2, 40] formulate detection and re-ID into a unified framework by sharing the backbone network, and train both tasks in an end-to-end fashion.

The main goal of person search, as well as re-ID is to generate a discriminative feature from an image for matching the same class (identity) to the target person. Existing methods focus on feature representation learning [29] to classify different instances in a gallery of pedestrian images. However, as shown in Fig. 1, misalignment issue arises from occlusion with other instances or background objects, and false positives from detector that hinder robust feature representation. To overcome this, recent methods [16, 39, 37, 24, 25] exploit attention mechanism to obtain a saliency map emphasizing discriminative parts of human (*e.g.* clothes or handbags). However, person detection and re-ID are inherently contradictory [32], since the former aims to classify everyone in an image into person class, while the latter aims to distinguish individual person for identification. It limits the applicability of attention mechanism in conventional re-ID methods to the person search framework.

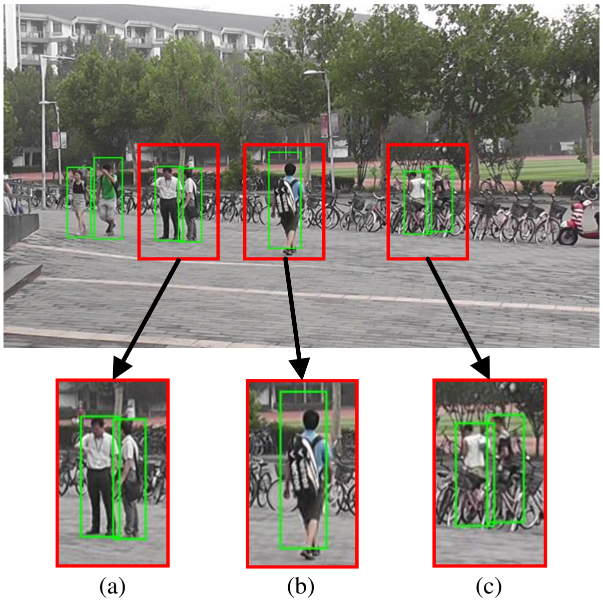


Figure 1. Examples of misalignment problem in person search task caused by (a) the presence of other instances, (b) mis-detection and (c) background objects.

On the other hand, loss functions for training the feature representation such as cross-entropy [35] or triplet loss [5] are widely used in re-ID community, in order to make the same identities be closer in embedding space, while dif-

[*] Corresponding author

ferent identities to be far apart. However, these two losses are not scalable to person search because the unified framework restrains the batch size for training to be small. To realize this, Xiao *et al*. [30] propose a novel online instance matching (OIM) loss for person search. OIM loss aggregates diverse patterns of identity into a feature vector which is representative of a certain identity, i.e., prototype [23, 34], and then optimizes all features to be closer to their corresponding prototype by minimizing intra-class variation [27]. However, the prototype is updated in an online manner regardless of the input, whereas the input identity might be affected by noise components which disturb the optimization of prototype.

In this paper, we conjecture that the prototype in OIM loss can be used as guidance to solve misalignment problems in conventional methods. Inspired by the fact that prototype optimally describes each class, we aim to learn the attention mechanism so that attention could focus on the region that is similar to its prototype. Specifically, we compute the pixel level affinity between the prototype and feature of the detected instance, and exploit it as guidance for attention learning. This map emphasizes region which highly responses to the prototype, and supervises the saliency to attention module to highlight consistent region against pose or viewpoint variation during inference. In addition, we introduce a new prototype update scheme for discriminative feature learning from OIM loss. Rather than utilizing fixed momentum for prototype update, we define the momentum as the ratio of the following two cosine similarities between 1) target and positive prototype pair and 2) target and hardest negative prototype pair. When the target feature is closer to the hardest negative prototype, low momentum is assigned to the target feature to prevent the prototype from moving closer to the hardest negative.

The main contributions of this paper can be summarized as follows:

- We present a prototype-guided attention module, to enable learning of attention mechanism in person search by exploiting prototype as guidance.
- We introduce a new prototype update scheme to increase the discriminative ability of prototype across different instances.
- We demonstrate the advantage of our proposed method over state-of-the-art methods through extensive experimental evaluations.

2. Related Work

Person search Given a query, person search aims to match and localize a specific person in a gallery scene image. Most person search methods scale down the problem to person re-ID by cropping instances in an image level [6, 1, 9], or feature level [18, 2, 40] adopting detection module. It makes the person search task challenging due to misalignment between a matching pair caused by erroneous detection, occlusions or background clutters. To realize this, several methods have been proposed to solve the issue. Lan *et al*. [14] identify multi-scale matching and align the semantics from each scale via knowledge distillation. Han *et al*. [9] state that training detector independent from re-ID loss causes misalignment, and propose a new localization refinement framework by optimizing the detector with the supervision of re-ID. Zhong *et al*. [40] align each part of an instance by estimating visibility and introduce a partial feature matching scheme.

Attention mechanism Attention mechanism, which aims at localizing the discriminative regions in an image, has been used in re-ID [16, 37, 24, 7, 25, 31, 2] and person search [1] task to tackle misalignment problems. For example, Li *et al*. [16] learn spatial, channel and hard regional attention simultaneously to enhance the attention selection within arbitrarily-aligned bounding boxes. Zheng *et al*. [39] enforce attention consistency among images of the same person for instance invariant feature representations.

Meanwhile, there are some approaches which use additional information as guidance to learn attention. Usually, the guidance includes human semantic such as foreground mask [24, 1] or body parts [36, 7] to remove background clutters or highlight unoccluded region for partial matching. However, all of these methods require additional models trained on different datasets for guidance, which affects the performance of main task. In contrast, the proposed prototype guidance helps the prototype-based feature learning by removing noise components of an input feature, and can be easily obtained during optimization step.

Prototype Prototype, also known as proxy [17], mean [8], or center [27], is the one representative of a class among training examples. The concept of prototype is widely used in deep metric learning [27, 17, 30, 20]. Rather than focusing on individual instances in a mini-batch [4, 12], prototype-based loss optimizes all features to be close to their prototype to minimize intra-class variation [27, 20].

The true location of prototype requires the knowledge of whole dataset which is expensive. Most works [8, 27, 30] estimate prototype by updating them in an online manner, using fixed momentum during whole training process. While effective, input samples with misalignment issues in person search task can hinder the optimization with prototype. To address this, we propose a new prototype update scheme with adaptive momentum to prevent the prototype from getting close to its hard negative due to misalignment.

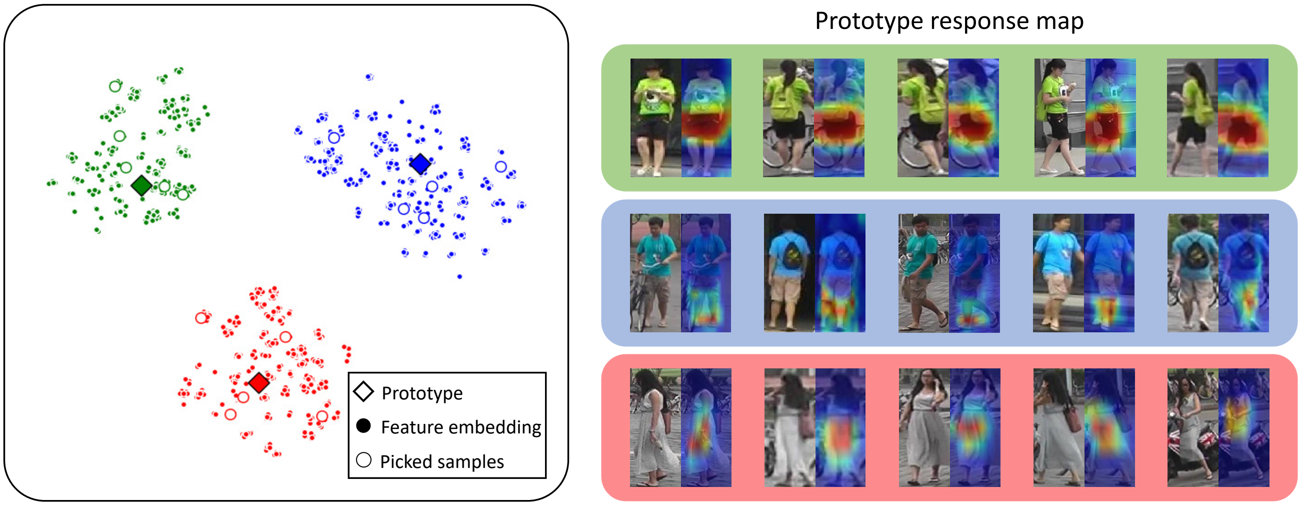


Figure 2. Illustration of feature distribution in embedding space (left), and prototype response maps of randomly picked samples per each class, marked as white circle (right). The response map highlights the consistent region of a class, invariant to pose or camera viewpoint.

3. Method

3.1. Preliminary

Typical end-to-end person search methods [30, 18, 40, 2] is based on Faster R-CNN [21] with re-ID head as a unified framework. Given a single image, it first localizes pedestrian candidates with a region proposal network (RPN), then crops the feature according to the region of interests (ROI) to extract proposal features at a ROI pooling layer. Let us denote the extracted feature as $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, where C is the number of feature channels with height H, and width W. Re-ID head then generates the feature embedding $\mathbf{f} \in \mathbb{R}^{d\times H\times W}$ from $\mathbf{F}$, then aggregates spatial information of $\mathbf{f}$ to $\mathbf{x} \in \mathbb{R}^{d}$, where d is the channel dimension.

For optimization of the embedding, OIM loss [30] is used for dealing with few numbers of identities in uncropped full scene input images. They estimate prototypes representing each class via online update scheme with all training samples, and memorize them into lookup table V. To use unlabeled identities for learning, the features from them is also memorized in a circular cue U. Given $\mathbf{x}_i$ the embedding of i-th class, the probability p_i of $\mathbf{x}_i$ being recognized as i could be estimated from the softmax function with every elements in V and U as:

$$p_i = \frac{e^{(\mathbf{v}_i^{\mathsf{T}}\mathbf{x}_i/\tau)}}{\sum_{j=1}^{N} e^{(\mathbf{v}_j^{\mathsf{T}}\mathbf{x}_i/\tau)} + \sum_{k=1}^{Q} e^{(\mathbf{u}_k^{\mathsf{T}}\mathbf{x}_i/\tau)}}, \tag{1}$$

where $[\mathbf{v}_1, \cdots, \mathbf{v}_N] \in V$, $[\mathbf{u}_1, \cdots, \mathbf{u}_Q] \in U$ and τ is temperature parameter for smoothness of probability distribution. The final objective of OIM loss is to maximize the log-likelihood of p_i: $\mathcal{L}_{oim} = E_x[\log p_i]$. After the forward path, prototype v_i is updated with a momentum of η following L_2 normalization:

$$\mathbf{v}_i \leftarrow \eta\mathbf{v}_i + (1-\eta)\mathbf{x}_i, \quad \eta \in [0,1]. \tag{2}$$

3.2. Motivation and Overview

The baseline in Sec. 3.1 simply uses global average pooling (GAP) for embedding and thus suffers from the misalignment caused by detection module. Moreover, such misaligned features disturb the prototype updated from Equ. (2) to be discriminative in embedding space, which degrades feature learning of OIM loss.

In this paper, we aim to solve the misalignment issue caused by detection module with prototype guided attention (PGA) module, which produces a saliency map for localizing discriminative regions inside bounding boxes. We guide the learning of saliency map with a prototype to highlight the consistent region across the same class as shown in Fig. 2, and make the feature from detection module embedded be closer to its prototype, reducing intra-class variation. Furthermore, to increase the separability among prototypes, we propose a new prototype update scheme with adaptive momentum which constraints feature involved on prototype update.

3.3. Prototype Guided Attention Module

Attention module for saliency feature learning As illustrated in Fig. 3, we present an attention module to provide the saliency map enhancing the discriminative regions while suppressing the noise component. In this work, we adopt SE block [13] as a base for designing our attention model for its lightweight mechanism and scalability [1, 18, 36]. We consider both channel and spatial attention for generating a saliency map following [3, 28].

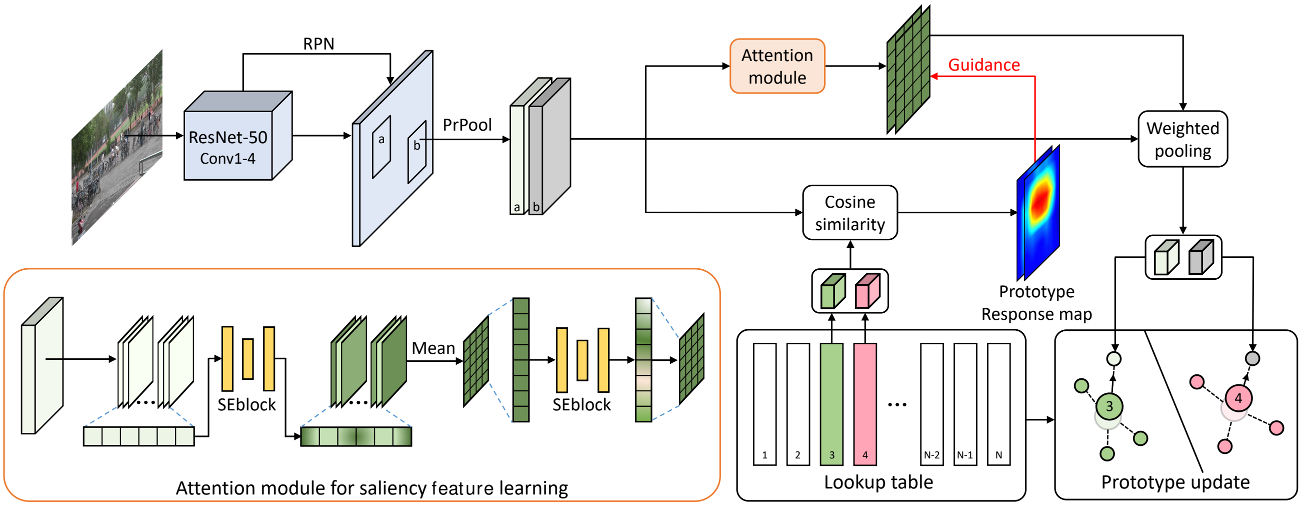


Figure 3. The overall architecture of the proposed PGA module. It consists of attention module for saliency feature learning, prototype guidance generation, and prototype update with the embeded training samples.

Given the feature map $\mathbf{f}$ in Sec. 3.1, we aggregate the feature in spatial domain to produce channel descriptors $\mathbf{f}_c \in \mathbb{R}^d$. The descriptors are fed into two consecutive FC layers to capture channel-wise dependencies, and the channel-wise attention $\mathbf{z}$ is obtained by applying a sigmoid function on the descriptors for normalization. The process is formulated as:

$$\mathbf{z} = \sigma(\mathbf{W}_2^c \, \delta(\mathbf{W}_1^c \, \mathbf{f}_c)), \tag{3}$$

where $\mathbf{W}_1^c \in \mathbb{R}^{\frac{d}{r} \times d}$ and $\mathbf{W}_2^c \in \mathbb{R}^{d \times \frac{d}{r}}$ are the parameters of FC layers for channel reduction and expansion by a factor or r and σ and δ refer to sigmoid and ReLU function.

To extract the spatial attention from the informative channels, $\mathbf{f}$ is channel-wise pooled with the weight $\mathbf{z}$ then flattened, resulting a spatial descriptor $\mathbf{f}_s \in \mathbb{R}^{HW}$. Similar to channel attention, we use two FC layers and sigmoid function to generate a saliency map $\mathbf{A}$:

$$\mathbf{A} = \sigma(\mathbf{W}_2^s \, \delta(\mathbf{W}_1^s \, \mathbf{f}_s)), \tag{4}$$

where $\mathbf{W}_1^s \in \mathbb{R}^{\frac{HW}{r'} \times (HW)}$ and $\mathbf{W}_2^s \in \mathbb{R}^{(HW) \times \frac{HW}{r'}}$ are the parameters of SE block with a factor or r'. The spatial attention $\mathbf{A}$ is reshaped back to the same spatial size of $\mathbf{f}$.

Prototype guidance generation A straightforward method to learn saliency from attention module is self-guided learning from the main task's loss function. However in person search, attention module is affected from both detection and re-ID loss which are conflicting each other.

Instead, we exploit prototype, which is the optimal representation of class in OIM loss, as the guidance to help the attention learning. Specifically, we measure the response of prototype in OIM loss $\mathbf{v}$ with the feature $\mathbf{f}$ from Sec. 3.1. We define the response map $\mathbf{M}$ as cosine similarity between $\mathbf{f}$ and $\mathbf{v}$:

$$m^{h,w} = \frac{\mathbf{v} * \mathbf{f}^{h,w}}{\|\mathbf{v}\|_2 * \|\mathbf{f}^{h,w}\|_2}, \tag{5}$$

where $m_{h,w} \in \mathbf{M}$. We further attach ReLU function to $\mathbf{M}$ to suppress the weak prototype responses.

After generating the guidance, the attention map $\mathbf{A}$ is optimized from it by a Mean Squared Error loss defined as

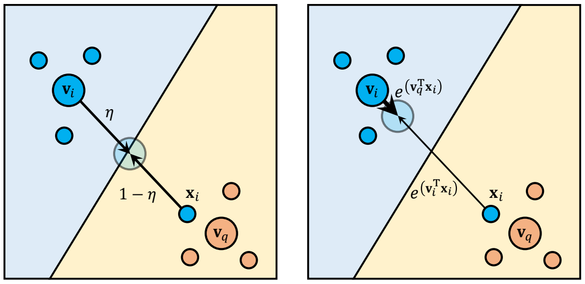
$$\mathcal{L}_{att} = \sum_{h}^{H} \sum_{w}^{W} \|\mathbf{M}^{h,w} - \mathbf{A}^{h,w}\|_2. \tag{6}$$

Note that back-propagated gradient flew to $\mathbf{M}$ is set to zero, to prevent the optimization of prototype on this loss function.

3.4. Prototype Update with Adaptive Momentum

In OIM loss, the location of prototype in embedding space is estimated by updating the prototypes online as Equ. (2). The equation can be seen as an exponentially weighted moving average [22], where features involving the update process recently have high weights while older features are progressively downweighted. Therefore, if the recent feature has noise components such as background clutters or occlusion, it could transfer the prototype close to prototypes of other classes, which reduces inter-class differences between feature embeddings.

From this observation, we propose a new prototype update process considering the distribution of prototypes. To maintain the efficiency on non-parametric prototype estimation in OIM loss, we modulate momentum η in Equ. (2) at every prototype update to transfer the prototype far apart



(a) (b)

Figure 4. Illustration of prototype update scheme with (a) fixed momentum Equ. (2) and (b) adaptive momentum based on the similarity between a prototype and its hard negative. Thickness of each arrow stands for the magnitude of the corresponding component's momentum for update.

from its hard negative [4]. Suppose that there is a target annotated to identity i as $\mathbf{x}_i$, and prototype set V. We first compare similarities of two pairs: $\mathbf{v}_i^\mathsf{T}\mathbf{x}_i$ and $\mathbf{v}_q^\mathsf{T}\mathbf{v}_i$, where q is the index satisfying the condition

$$q = \underset{p \in \{1 \dots N\} \setminus i}{\operatorname{argmax}} \ \mathbf{v}_p^\mathsf{T}\mathbf{v}_i. \tag{7}$$

As shown in Fig. 4, when prototype $\mathbf{v}_i$ is more similar to its hard negative prototype $\mathbf{v}_q$ than to target $\mathbf{x}_i$, the importance of $\mathbf{v}_i$ on update process should be low compared to that of $\mathbf{x}_i$. So we define the adaptive momentum from the similarities:

$$\mathbf{v}_i \leftarrow \frac{e^{(\mathbf{v}_q^\mathsf{T}\mathbf{x}_i/\tau)}}{e^{(\mathbf{v}_q^\mathsf{T}\mathbf{x}_i/\tau)} + e^{(\mathbf{v}_i^\mathsf{T}\mathbf{x}_i/\tau)}}\mathbf{v}_i + \frac{e^{(\mathbf{v}_i^\mathsf{T}\mathbf{x}_i/\tau)}}{e^{(\mathbf{v}_q^\mathsf{T}\mathbf{x}_i/\tau)} + e^{(\mathbf{v}_i^\mathsf{T}\mathbf{x}_i/\tau)}}\mathbf{x}_i, \tag{8}$$

where τ is the softmax temperature value. Note that when the value $\mathbf{v}_i$ is empty in the initial state , we just put the target feature to V.

3.5. Implementation Details

Network details We implement our proposed method on Pytorch [19]. We build our model upon OIM [30] which consists of backbone ResNet-50 [19] and Faster R-CNN [19] as a detection module. For training a region proposal network (RPN), we adjust the anchor scales $\{8, 16, 32\}$ and aspect ratio $\{1, 2, 3\}$. We adopt other settings such as foreground overlap threshold or number of proposals per batch the same as [21]. Next, RoIAlign [10] is utilized for feature extraction on 'conv4' block of ResNet-50. After transforming the dimension of extracted features to 2048 on 'conv5' block, we add detection and re-ID head to get a bounding box and feature embedding. We use the same anchor setting with RPN for training the detection head. All features are then transformed to 256 dimension and pass through attention module which consists of two SEblock having ratio $r = 8$ and $r' = 7$ respectively to generate a saliency map per each feature. Finally, the feature is weighted average pooled by the saliency map followed by L2 normalization. For base OIM loss setting, the temperature parameter τ is set to 0.033 and momentum η to 0.5.

Model training We train our network on a single NVIDIA TITAN X GPU with 4 images per one batch. All images are resized to 900 on a shorter side and 1500 on a larger side. The learning rate is initialized to 0.0005 for CUHK-SYSU and 0.0001 for PRW dataset, and decreased by 10 at 30k-th iteration. In both datasets, the training continues until 40k-th iteration. We use SGD for optimization with a weight decay of 0.0001 and momentum of 0.9.

Loss function We optimize our proposed method with linear combination of detection loss $\mathcal{L}_{det}$, OIM loss $\mathcal{L}_{oim}$, and attention guidance loss $\mathcal{L}_{att}$ as follows:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{oim} + \lambda\mathcal{L}_{att}, \tag{9}$$

where λ is the loss weight of $\mathcal{L}_{att}$. The initial value of λ equals to zero, and 0.1 after all prototypes are saved at V.

4. Experiments

4.1. Datasets and metrics

CUHK-SYSU CUHK-SYSU dataset [30] is a large-scale person search dataset collected from urban scenes and movies. The dataset contains 18,184 images, 96,143 annotated bounding boxes and 8,432 labeled identities. All unlabeled identities are served as negative samples and excluded from the training of re-ID. We follow the train/test split provided from the dataset: 11,206 images with 5,532 identities for training, and 6,978 gallery images with 2,900 query images for testing. The dataset provides gallery subsets with different sizes, and we use the default gallery size as 100 in the experiment.

PRW PRW dataset [38] is a collection of 11,816 video frames captured from the university. The dataset contains 43,110 bounding boxes and 932 identities. Similar to CUHK-SYSU, the annotation has labeled and unlabeled instances and 34,304 bounding boxes are annotated with identities. We also follow the train/test split in the dataset: 5,704 images with 482 identities for training, and 6,112 images with 2057 query images for testing. In PRW, the search space for a query is the whole gallery set.

Evaluation metrics We report performance using mean Average Precision (mAP) and Common Matching Characteristic (CMC top-K), following the common practice in person search. The first metric calculates an averaged precision (AP) score of searching a query from gallery images and averages the AP scores from all query identities to get mAP score. The second metric counts the case when

Methods	res.	mAP(%)	top1(%)
Baseline	7×7	35.5	77.2
Self-attention	7×7	36.2	77.7
Self-attention*	14×14	36.8	78.3
PGA	7×7	40.3	79.9
PGA*	14×14	**42.7**	**82.8**

Table 1. Component analysis of prototype guided attention module (PGA) with the baseline and self-attention module which shares the same network to PGA trained without prototype guidance. 'res.' denotes the spatial size of saliency map. All experiments are done on PRW.

Methods	τ	mAP(%)	top1(%)
Baseline	-	35.5	77.2
Baseline w/ AMU	0.03	38.7	79.8
	0.05	38.9	80.0
	0.1	38.6	79.4
PGA w/o AMU	-	40.3	79.9
PGA w/ AMU	0.03	42.4	83.1
	0.05	**42.5**	**83.5**
	0.1	42.2	82.9

Table 2. Component analysis of adaptive momentum update (AMU) with the baseline and PGA module on PRW.

there exists at least one of the top-K predicted bounding boxes overlapping with the ground truth with the overlap rate larger than 0.5. In this experiment, we adopt CMC top-1 only as it is the hardest condition among all possible Ks.

4.2. Ablation Study

Effectiveness of PGA We analyze our PGA module with the ablation evaluations with respect to the effectiveness of prototype guidance in attention module. In order to demonstrate the spatial size of saliency map on performance, we decrease the stride to 1 in conv5 block [11] using dilated convolution, denoted as 'PGA*'. As in Tab. 1, we observe that the usage of saliency map for feature extraction boosts the re-ID performance. While self-attention module improves the mAP and top1 performance of the OIM baseline by 0.7% and 0.5% only in 7×7 saliency map, PGA improves the performance by 4.8% and 2.7%. This indicates that prototype guidance helps the attention learning on person search. Moreover, increasing the spatial size of saliency map contributes to the performance gain by providing fine-grained information about the discriminative region inside the bounding box.

Effectiveness of AMU In Tab. 2, we evaluate the impact of adaptive momentum update (AMU) on the baseline and PGA with the different temperature setting in Equ. (8). Compared to the baseline using fixed momentum update of prototype in Equ. (2), AMU improves the performance of mAP by 3.2% and top1 by 2.6%, for $\tau = 0.03$. Adding AMU on PGA, we achieve higher improvement of mAP by 6.9% and top1 by 5.9%, for $\tau = 0.03$. This demonstrates that the AMU supports the robust prototype guidance generation from PGA. For all experiments, we observe that temperature 0.05 provides the best performance so we adopt the value on our module.

Moreover, we conduct an experiment to validate the effectiveness of AMU for increasing the separability among prototypes in Fig. 5. Specifically, we track a prototype and collect cosine similarity of the hardest negative and second hardest negative at every update in the training process to compare the distance between them. It is clear that, the prototype updated from AMU gets further away from its hard negatives at every epoch, which greatly benefits the feature learning of OIM loss.

Saliency map visualization For the qualitative analysis, we visualize the saliency map from PGA module. In Fig. 6, we observe that saliency map supervised from the prototype guidance localizes consistent region to the same person which leads to invariant representation from the viewpoint or pose variation. In addition to validate the robustness of the saliency map, we visualize the map with respect to the misaligned cases in Fig. 7. In all cases, the saliency map improves the matching performance by highlighting the discriminative region and suppressing the background clutters or other instances.

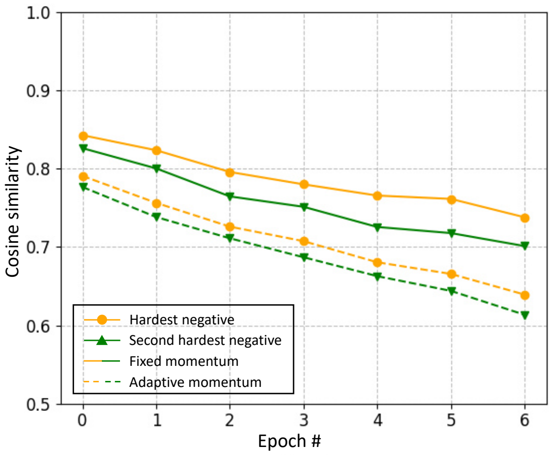


Figure 5. Comparison of average cosine similarity between a prototype and its hard negatives per epoch on training processes.

Methods	# of Parameters (K)	speed (sec.)
Baseline	3596	0.355
PGA	3804	0.356

Table 3. Comparison of the number of parameters and runtime between PGA and baseline. All values are measured on TITAN X GPU.

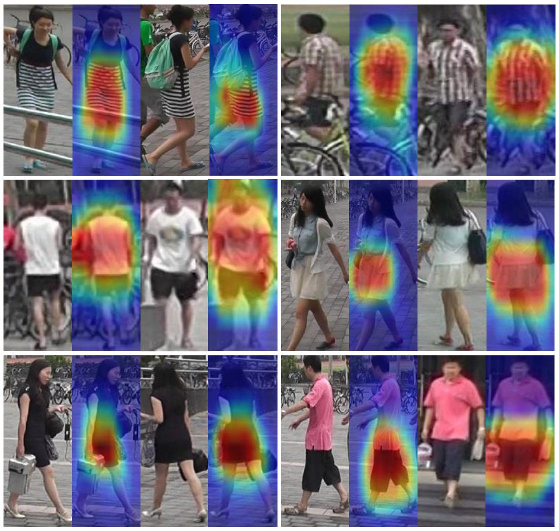


Figure 6. Visualization of saliency map generated by our PGA module. Note that red color indicates the salient regions and blue color indicates suppressed regions.

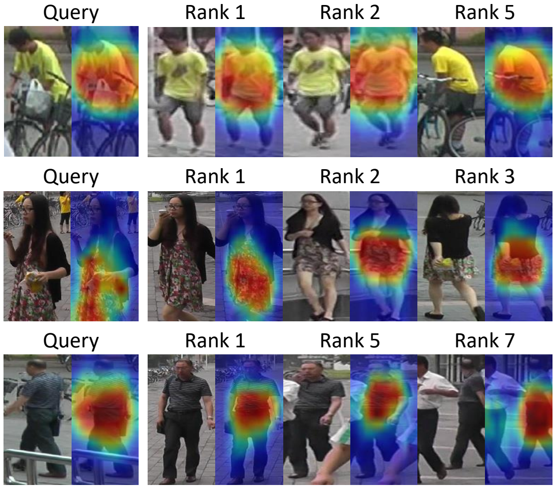


Figure 7. Demonstration of the robustness of proposed saliency map against occlusions with background (tow row), bounding box covering half of a person (middle row), and the presence of other instances (bottom row).

Runtime comparison In Tab. 3, we compare the number of parameters and runtime speed to process a gallery image with the baseline OIM. Likewise in baseline, our module shows the high computational speed of joint framework about 0.3 seconds, which is applicable in real world. It is noticeable that with a small increase about $200K$ in attention module, we achieve the state-of-the-art performance.

4.3. Comparison with State-of-the-Art Methods

In this section, we compare our model with several state-of-the-art methods on person search. For better comparison, we categorize the person search methods into two-step with two separate detection and re-ID models and one-step with a unified model.

	Method	mAP (%)	top1 (%)
two-step	MGTS [1]	83.0	83.7
	RDLR [9]	93.0	94.2
	IGPN [6]	89.1	90.5
	TCTS [26]	**93.9**	**95.1**
one-step	OIM [30]	75.5	78.7
	Context Graph [33]	84.1	86.5
	QEEPS [18]	88.9	89.1
	APNet [40]	88.9	89.3
	NAE+ [2]	92.1	92.9
	Ours	90.2	91.8
	Ours*	**92.3**	**94.7**

Table 4. Comparison of performance on CUHK-SYSU.

	Method	mAP (%)	top1 (%)
two-step	MGTS [1]	32.6	72.1
	RDLR [9]	42.9	70.2
	IGPN [6]	46.2	86.1
	TCTS [26]	**46.8**	**87.5**
one-step	OIM [30]	21.3	49.9
	Context Graph [33]	33.4	73.6
	QEEPS [18]	37.1	76.7
	APNet [40]	41.9	81.4
	NAE+ [2]	44.0	81.1
	Ours	42.5	83.5
	Ours*	**44.2**	**85.2**

Table 5. Comparison of performance on PRW.

Results on CUHK-SYSU In Tab. 4, we show the person search results on CUHK-SYSU with a gallery size of 100. Our method achieves the mAP of 92.3% and top1 accuracy of 94.7%, surpassing other methods in one-step method. Compared to QEEPS [18] that uses siamese network to guide the query information to the main network or Context Graph [33] adopting additional graph networks, our method reuse the prototype in optimization process requiring less memories. It also outperforms APNet [40] which adopts the part-based model to solve the misalignment problem. Note that NAE+ [2] also exploits pixel-wise saliency map, but the saliency is generated from coarse bounding box annotation. Our methods are also comparable to the two-step methods including MGTS [1] which uses mask information for person search, and RDLR [9] that adopts bounding box refinement. TCTS [26] produces the highest performance of all methods, but they require a cascade process which is complicated. Compared to the two-step methods, our method is a unified model which requires less parameters and memories.

Results on PRW In Tab. 5, we compare our results to state-of-the-art approaches on PRW dataset. We achieve

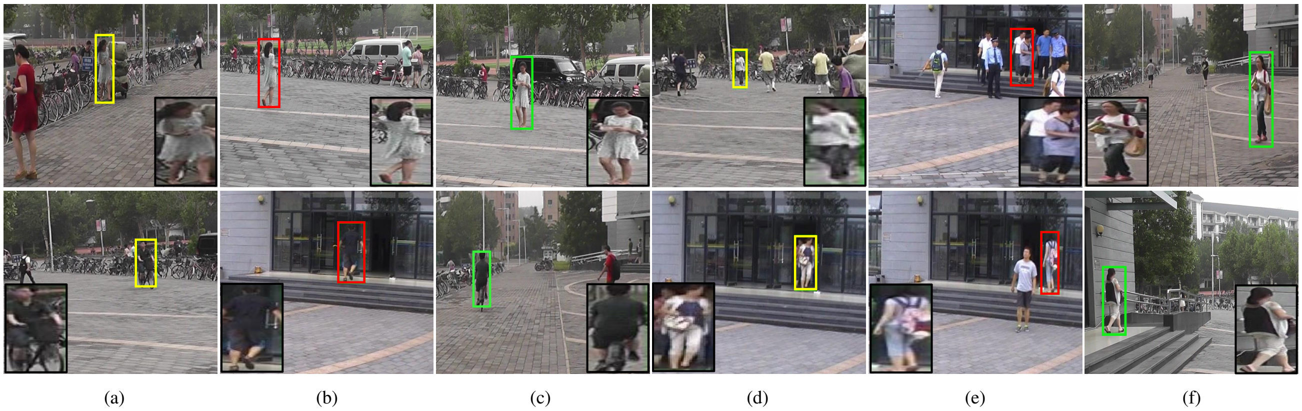


Figure 8. Qualitative result of top1 person search results in PRW dataset: (a),(d) are query images, (b),(e) are top1 retrieved results from baseline OIM, and (c),(f) are the results from our model. We visualize the red box as a failure, and the green box as a correct case. The black box represents the detected instance in a large scale for better comparison.

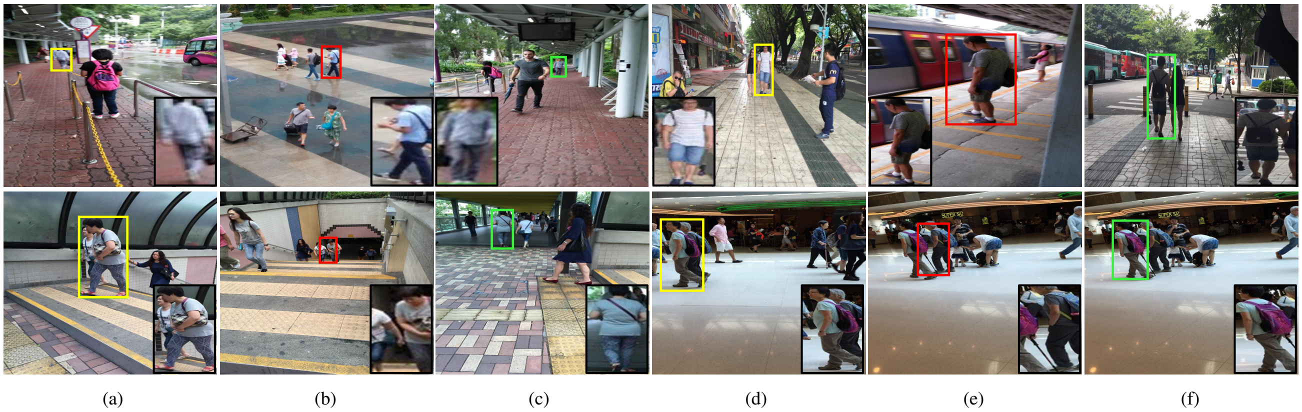


Figure 9. Qualitative result of top1 person search results in CUHK-SYSU dataset: (a),(d) are query images, (b),(e) are the results from baseline OIM, and (c),(f) are the results from our model. We visualize the red box as a failure, and the green box as a correct case. The black box represents the detected instance in a large scale for better comparison.

the mAP of 44.2% and top1 accuracy of 85.2%, outperforming the most person search methods. Compared to CUHK-SYSU, PRW has a larger gallery size containing many identities and thus more challenging. The saliency map from our model could emphasize the discriminative region which is effective on such a hard dataset. It is noticeable that our method shows high improvement of top1 accuracy, surpassing 3.9% compared to NAE+. The saliency map learned from the prototype guidance provides consistent region across the same person, thus it helps retrieval performance.

5. Conclusion

We have introduced a novel framework to learn the discriminative representation of each person instance under severe geometric variation. We account for the fact that prototype in OIM loss can optimally describe each class. To this end, we propose a prototype guided attention module by exploiting prototype in OIM loss as guidance for saliency feature learning. It allows solving misalignment problem in person search task with the additional supervisory signal from prototype guidance. Furthermore, we introduce a prototype update scheme with adaptive momentum to increase the discriminative ability of prototype across different instances. Our experiments have shown that our method effectively learned the saliency feature of each person instance, outperforming state-of-the-art methods. In the future, we will extend our model to part-based representation of one identity, to leverage partial matching without using extra part annotation.

Acknowledgements This research was supported by the Yonsei University Research Fund of 2021 (2021-22-0001), and R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA (NRF) funded by Ministry of Science and ICT (NRF-2018M3E3A1057289).

References

[1] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 734–750, 2018. 1, 2, 3, 7

[2] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12615–12624, 2020. 1, 2, 3, 7

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017. 3

[4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017. 2, 5

[5] 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. In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2016. 1

[6] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2585–2594, 2020. 2, 7

[7] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11744–11752, 2020. 2

[8] Samantha Guerriero, Barbara Caputo, and Thomas Mensink. Deepncm: Deep nearest class mean classifiers. In *International Conference on Learning Representations Workshop*, 2018. 2

[9] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9814–9823, 2019. 2, 7

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 5

[11] 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. 6

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3

[14] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *Proceedings of the European Conference on Computer Vision*, pages 536–552, 2018. 2

[15] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 1

[16] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018. 1, 2

[17] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017. 2

[18] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2019. 1, 2, 3, 7

[19] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[20] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6450–6458, 2019. 2

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 3, 5

[22] SW Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959. 4

[23] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 2

[24] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. 1, 2

[25] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7134–7143, 2019. 1, 2

[26] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11952–11961, 2020. 7

[27] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 499–515. Springer, 2016. 2

[28] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module.

In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 3

[29] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016. 1

[30] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 1, 2, 3, 5, 7

[31] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018. 2

[32] 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*, pages 937–940, 2014. 1

[33] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2158–2167, 2019. 7

[34] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3482, 2018. 2

[35] Yao Zhai, Xun Guo, Yan Lu, and Houqiang Li. In defense of the classification loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1

[36] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6995–7003, 2018. 2, 3

[37] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020. 1, 2

[38] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 5

[39] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5735–5744, 2019. 1, 2

[40] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6827–6835, 2020. 1, 2, 3, 7