Norm-Aware Embedding for Efficient Person Search

Di Chen[1,3], Shanshan Zhang[1]*, Jian Yang[1,2]*, Bernt Schiele [3]

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
School of Computer Science and Engineering, Nanjing University of Science and Technology

[2] Jiangsu Key Lab of Image and Video Understanding for Social Security

[3] Max Planck Institute for Informatics, Saarland Informatics Campus

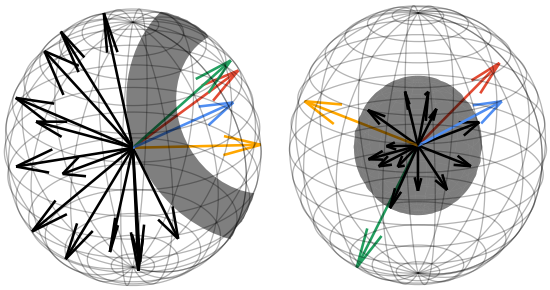
{dichen, shanshan.zhang, csjyang}@njust.edu.cn, {dichen, schiele}@mpi-inf.mpg.de

Abstract

Person Search is a practically relevant task that aims to jointly solve Person Detection and Person Re-identification (re-ID). Specifically, it requires to find and locate all instances with the same identity as the query person in a set of panoramic gallery images. One major challenge comes from the contradictory goals of the two sub-tasks, i.e.*, person detection focuses on finding the commonness of all persons while person re-ID handles the differences among multiple identities. Therefore, it is crucial to reconcile the relationship between the two sub-tasks in a joint person search model. To this end, we present a novel approach called Norm-Aware Embedding to disentangle the person embedding into norm and angle for detection and re-ID respectively, allowing for both effective and efficient multi-task training. We further extend the proposal-level person embedding to pixel-level, whose discrimination ability is less affected by misalignment. We outperform other one-step methods by a large margin and achieve comparable performance to two-step methods on both CUHK-SYSU and PRW. Also, our method is easy to train and resource-friendly, running at 12 fps on a single GPU.*

1. Introduction

In visual surveillance systems, the most fundamental problems are 1) how to locate persons within images, and 2) how to determine, if a query person is present in a particular set of images, typically across different cameras. The above two problems are usually investigated as the two independent tasks of Pedestrian Detection and Person Re-identification (re-ID). However, in practical applications, it is favorable to solve them in a joint framework, not only for convenience and high efficiency, but also for better performance. The task of Person Search, as introduced in [46], has the goal to retrieve a query person from a gallery of uncropped images captured by different cameras. Person search inherits difficulties from both re-ID and detection, *e.g.* viewpoint and illumination variance, cluttered background, occlusion, changing poses, *etc*., and is thus more challenging than either of the two tasks alone.



(a) (b)

Figure 1. Illustration on how person and background representations are scattered in the embedding space. Black arrows denote background while colorful ones denote persons with different identities. Gray surfaces are the decision boundary for person and background. (a) For L_2 normalized embeddings, the inter-class angle distances for different persons are squeezed by backgrounds. (b) Norm-aware embeddings separate persons and background by norms and discriminate person identities by angles, thus the constrain on inter-class distances is relaxed.

A standard way to address person search is to use a two-step strategy, *i.e.*, cascading a pedestrian detector and a re-ID feature extractor trained separately (*e.g.* [61, 3, 21, 16]). All candidate persons are cropped from the gallery images according to the detector, and fed into a standard person re-ID model. In contrast, others propose to share the backbone network between detection and re-ID [44, 42, 26, 2, 30, 47]. Given an uncropped image, these models output bounding box coordinates and the corresponding L_2 normalized identity embeddings for all the persons within. These works extend *e.g.* Faster R-CNN [34] by stacking an additional fully-

*Corresponding author.

connected layer to produce L_2 normalized embeddings and train the whole model jointly with standard detection losses and an identity classification loss. Nonetheless, they suffer from the conflicting objectives of detection and re-ID during training, as pointed out by [3]. An intuitive illustration for L_2 normalized embeddings is shown in Fig. 1(a). The detection classification objective tends to squeeze the embedding space for all persons regardless of the identities, so as to better separate from background. Therefore, sharing the feature space with background instances limits the angular margin between different identities. As a consequence, due to the inherent trade-off between the two tasks, their detection and re-ID performances are typically both lower than the separately-trained counterparts.

In this work, our goal is to develop a light-weight yet accurate model for person search. We adopt the one-step strategy, *i.e.*, jointly optimize detection and re-ID in an end-to-end model, and relieve the objective contradictory problem by explicit decomposition. Specifically, we share the representations for detection and re-ID completely but decompose the features in the polar coordinate system, where each embedding vector is decomposed to radial norm r and angle $\boldsymbol{\theta}$. The radial norm r is used for pedestrian detection and could be interpreted as the detection confidence of a bounding box. The angle term $\boldsymbol{\theta}$ measures the cosine similarity between persons, which is widely used in person re-ID. The principle idea is demonstrated in Fig. 1(b). During training, the embedding norms are optimized with a binary classification loss, and the angles are optimized with an OIM loss [44], which is a multi-class cross-entropy loss with normalized softmax weights. During inference, we fix the norm of the query person to 1 and calculate its similarity (dot product) to an arbitrary proposal, which is determined by both the norm and angle. Therefore, a high value of similarity indicates both high detection confidence and high identity similarity. Since the embedding norm is explicitly utilized, we call our method norm-aware embedding (NAE).

Another challenge for joint detection and re-ID is the spatial misalignment problem. Typically, when we train a detector, one proposal is sampled as positive when it has a larger Intersection over Union (IoU) than 0.5 to any ground truth box. This relatively loose matching criterion makes sure to sample enough positives in each mini-batch, but has a negative effect as it includes many mis-aligned samples. Those samples with low alignment quality are harmful for re-ID performance [61] as the included background clutter usually plays a negative role on the features' discrimination ability. In order to alleviate this problem, we propose to re-weight features of each local patch according to its confidence of belonging to a person. Specifically, we perform fine-grained person/background classification for each proposal, *i.e.*, we predict for each pixel location the confidence of belonging to a person, which is then used as the spatial attention weight for feature aggregation. After re-weighting, the features used for re-ID are expected to focus more on the person area while suppressing the background clutter, and thus become more discriminative for identity classification. Our approach using the above fine-grained classification is compatible with the norm-aware embedding, hence called NAE+.

In summary, the main contributions of this work are as follows:

- We propose the norm-aware embedding method (NAE) for person search. NAE mitigates the objective contradictory problem by decomposing the feature embedding into norm and angle for detection and re-ID respectively.
- A pixel-wise extension, denoted as NAE+, is proposed to deal with the misalignment problem for end-to-end person search.
- Our methods are fast, explainable and achieve competitive performance on standard benchmarks (CUHK-SYSU and PRW).

2. Related Work

Person Search. Recently, person search has raised a lot of interest to researchers in the computer vision community. Zheng *et al*. [61] first make a thorough evaluation on a number of combinations of different detectors and re-identifiers. They also propose a cascaded fine-tuning strategy for training and Confidence Weighted Similarity (CWS) for person matching. Lan *et al*. [21] analyze the resolution diversity problem in person search and solve the multi-scale matching problem by Cross-Level Semantic Alignment (CLSA). Chen *et al*. [3] raise attention on the contradictory objective problem in person search, and propose to avoid it by separating detection and re-identification. Han *et al*. [16] point out that the bounding boxes produced by a vanilla detector are not optimal for re-ID. Thus they develop an RoI transform layer that enables gradient flow from the re-identifier to the detector for localization refinement.

In contrast to the above two-step methods, other works aim to solve the person search problem more efficiently using one-step methods. For example, the Online Instance Matching (OIM) loss [44] and Center Loss [42, 40] are used to address the ill-conditioned training problem and enhance the feature discrimination power. Yan *et al*. [47] and Munjal *et al*. [30] propose to enrich the features with surrounding persons or the query person respectively. In [26] and [2], they discard the proposal generation operation and search the query person directly on the uncropped images by sequential decision making or reinforcement learning.

In this paper, we also adopt the one-step strategy. Based on the OIM model [44], we improve the feature learning with our norm-aware embedding. Additionally, the final

similarity calculation of our method is similar to CWS. Different from the original form which is used in a post-processing step, CWS in our method is naturally induced from the explicit decomposition in the polar coordinate system. Therefore, it is also useful to guide the training process for better feature learning.

Person re-ID. Early person re-ID models focus on designing features manually [37, 11, 58, 24] and learning effective distance metrics [20, 23, 50]. Recently, CNNs have become the de facto standard for building a re-ID model. Such models are usually trained as a feature extractor with siamese loss [49, 22, 1, 36, 27, 45], triplet loss [7, 4] or cross-entropy loss [43, 59, 61, 10, 41]. Instead of averaging the convolutional features from all locations, latest methods extract part-level features and join them together as the final person embedding [35, 39, 57, 48]. These methods usually partition the feature maps into horizontal stripes for fine-grained feature learning. Our pixel-wise extension of the norm-aware embedding is also inspired by this approach. Instead of dividing the feature maps into blocks, we use a pixel-wise probability map to re-weight the features at every location, which is further supervised by a segmentation loss with bounding box annotations.

Pedestrian Detection. Similar to person re-ID, early pedestrian detection methods are also based on hand-crafted features [12, 8, 9, 51, 54]. Deep neural networks, as versatile feature extractors, have dominated this task in recent years [52, 53, 32, 31]. Successful general object detection models are adapted for pedestrians, such as R-CNN [13, 52, 53] and Faster R-CNN [34, 55, 56]. In this work, we also build our model based on the adapted Faster R-CNN, which is extensible for fine-grained feature learning and reaches a sweet spot between speed and accuracy.

Embedding Norms. It is common practice to normalize the deep embeddings with unit length in face recognition [28, 29, 6], person re-ID [10, 41] and person search [44]. To the best of our knowledge, only two papers discuss the efficacy of embedding norms [15, 38]. Guo *et al*. [15] find that the norm of the softmax weight vector is related to the sample number of this class. They further propose to promote the norms of underrepresented classes in order to improve the performance of one-shot face recognition. Wang *et al*. [38] also use normalized embeddings to represent face identities. Additionally, they regress the norm of the embedding to the age of the given person by reducing the mean squared error between these two during training. However, the norm information is then ignored for age-invariant face recognition when matching identities. Different from the above two works, our method *makes explicit use* of the embedding norms rather than employing them as a regularization term during training. By using the norm for the classification task (person vs. background), we endow the norm with a clear semantic meaning, *i.e*., the detection confidence, which is essential for person search.

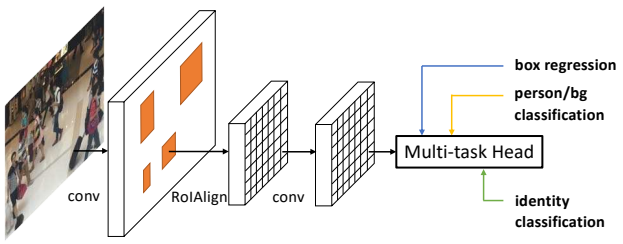


Figure 2. Overall architecture for one-step methods based on Faster R-CNN [34]. Black arrows denote the forward pass and colorful ones denote different supervision signals. Region Proposal Net is omitted for simplicity.

3. Methodology

A typical one-step person search method based on Faster R-CNN [34] is illustrated in Fig. 2. A multi-task head for localization, detection and re-ID is added on top of the top convolutional features of Faster R-CNN.

The first and most representative one-step method is OIM [44], where an L_2 normalized fully connected layer is concatenated to the global average pooled convolutional features. As is shown in Fig. 3(a), the box regression and region classification losses remain the same as in Faster R-CNN, with an identity classification loss supervising the person embeddings produced by the fully connected layer. In contrast, our norm-aware embedding method, illustrated in Fig. 3(b), removes the original region classification branch and uses the embedding norm as the binary person/background classification confidence.

In this section, we will describe the norm-aware embedding head in detail and present the pixel-wise extension for fine-grained feature learning.

3.1. Norm-Aware Embedding

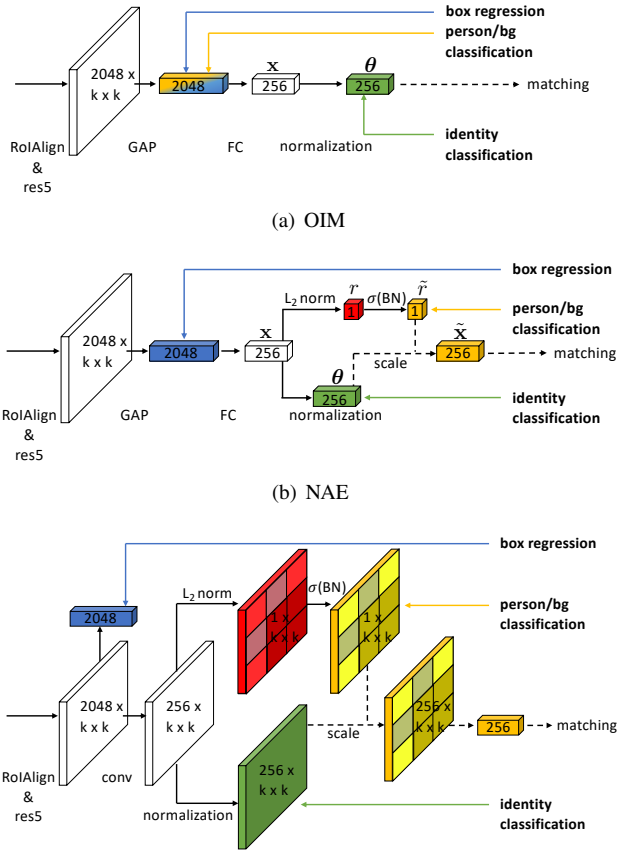
On top of the final convolutional features, we first apply global average pooling (GAP) and a fully connected (FC) layer to get the d dimensional feature vector $\mathbf{x}$, where d is fixed to 256 following [44]. Then $\mathbf{x}$ is decomposed explicitly in the polar coordinate system as:

$$\mathbf{x} = r \cdot \boldsymbol{\theta}, \tag{1}$$

where norm $r \in [0, +\infty)$ and angle $\boldsymbol{\theta}$ is a 256-dimensional vector with unit length.

To interpret the norm r as the detection confidence, we use a monotonic mapping to squeeze its magnitude to the range of $[0, 1]$:

$$\tilde{r} = \sigma\left(\frac{r - \mathbb{E}[r]}{\sqrt{\mathrm{Var}[r] + \epsilon}} \cdot \gamma + \beta\right), \tag{2}$$



(a) OIM

(b) NAE

(c) NAE+

Figure 3. Multi-task head architecture for Online Instance Matching (OIM), norm-aware embedding (NAE) and its pixel-wise extension (NAE+). Dashed arrows indicate that the procedure is only enabled during inference.

where σ is the sigmoid activation, within which is a batch normalization [19] layer.

The original embedding $\mathbf{x}$ is then scaled into our norm-aware embedding $\tilde{\mathbf{x}}$:

$$\tilde{\mathbf{x}} = \tilde{r} \cdot \boldsymbol{\theta} \tag{3}$$

This procedure is represented by the dashed arrows in Fig. 3(b).

Inference & Matching. For a query person, we first extract its embedding $\tilde{\mathbf{x}}_q$ by removing the RPN module and setting the proposal coordinate with the given bounding box. Since the query bounding box definitely contains a person, we manually set the norm of $\tilde{\mathbf{x}}_q$ to 1. Then, the similarity of the query person and an arbitrary detected person $\mathbf{x}_g$ in the gallery is calculated as follows:

$$\text{sim}(\tilde{\mathbf{x}}_q, \tilde{\mathbf{x}}_g) = \tilde{\mathbf{x}}_q^T \tilde{\mathbf{x}}_g = \tilde{r}_g \cdot \boldsymbol{\theta}_q^T \boldsymbol{\theta}_g \tag{4}$$

In the above equation, $\boldsymbol{\theta}_q^T \boldsymbol{\theta}_g$ is the cosine similarity between the query and the gallery person. Thus, the final similarity equals the cosine similarity weighted by the detection confidence, which is especially useful to suppress false detections. Meanwhile, it also shares the same formation as Class Weighted Similarity (CWS) [61]. However, instead of just using CWS as a post-processing step, we leverage it to explicitly decompose the embedding for the detection and re-ID objectives during training. We further demonstrate the efficacy of CWS in Sec. 4.3.

Training. As can be seen from Eq. 4, our norm-aware embedding is able to discriminate person identity as well as suppress false detections. Therefore, it can be supervised by re-ID and detection signals simultaneously during training. Specifically, the detection signal is cast on the scaled norm $\tilde{r}$ and formulated as a binary classification:

$$\mathcal{L}_{\text{det}} = -y \log(\tilde{r}) - (1-y) \log(1-\tilde{r}) \tag{5}$$

where y is a $\{0, 1\}$ label indicating if this proposal is considered as background or person. Meanwhile, we use an OIM loss [44] $\mathcal{L}_{\text{reid}}$ on the normalized angular vector $\boldsymbol{\theta}$, which is a multi-class cross-entropy loss that minimizes the angular margin for the same identity and maximizes that of different identities. The bounding box regression loss $\mathcal{L}_{\text{box}}$ remains identical to the form defined by Faster R-CNN. The three loss functions are illustrated by the yellow, green and blue arrows respectively in Fig. 3(b). Together with RPN classification and regression losses, they are jointly optimized by Stochastic Gradient Descent (SGD).

3.2. Pixel-Wise Extension

In Sec. 3.1, the convolutional features of each proposal are collapsed into a vector by global average pooling, losing spatial information. In this way, the person embeddings would suffer from distracting noise of the misaligned regions (the black region in Fig. 4). To address this problem, we propose NAE+, which is a pixel-wise extension of NAE. We carefully leverage the spatial information via highlighting the body part and suppressing the misaligned regions. Specifically, we first predict a $256 \times k \times k$ tensor from the top feature map with a 1×1 convolutional layer. Then the 256-dimensional vectors at all locations can be normalized and scaled into norm-aware embeddings, while still preserving the spatial structure. An illustration is shown in Fig. 3(c). In this way, the mapped norm $\tilde{r}_i$ at each location acts as a spatial attention, calibrating the per-pixel importance before the tensor is collapsed into the final matching vector.

The training of NAE+ can be formulated in a semantic segmentation manner, *i.e.*, supervising all the mapped norms with a per-pixel cross-entropy loss. Different from the standard semantic segmentation approach, the ground truth class map is not available in person search datasets, hence we need to generate the coarse ground truth from bounding box annotations. The generation process is shown in Fig. 4. For each RoI, we set its intersection to the ground truth bounding box as 1 and leave the rest as 0. Bilinear

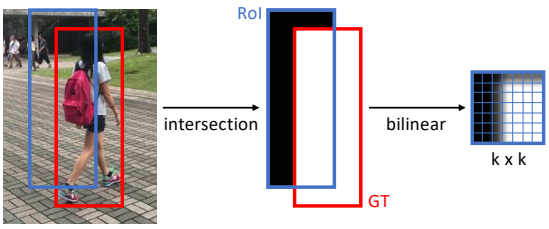


Figure 4. Pixel-wise label generation from bounding box annotation. Red box is the ground truth while blue box is the proposal. Black region is marked as 0, white being one and gray being values in between. Bilinear interpolation is used to match the label size to the feature map size.

interpolation is used to resize the RoI box into $k \times k$. The loss is then formulated as follows:

$$\mathcal{L}_{\text{det+}} = -\frac{1}{k^2}\sum_{i=1}^{k^2} y_i \log(\tilde{r}_i) + (1-y_i)\log(1-\tilde{r}_i) \quad (6)$$

where y_i is the generated per-pixel label with its value lying between $[0, 1]$. The label generation procedure is illustrated in Fig. 4.

During inference, even though the per-pixel probability map is trained, a single probability is needed to measure the detection confidence. A direct approach is to use the averaged probability $\mathbb{E}_i[\tilde{r}_i]$ across all spatial locations. However, we find through our experiments that this approach works poorly, *i.e.*, the confidence of a valid bounding box is too low since the mean value would be diluted by the inevitable background regions with low confidence. We address this problem by simply stretching the magnitude of $\mathbb{E}_i[\tilde{r}_i]$. Specifically, for each image, all the detection confidences are divided by the maximum value among them, such that all of them are expanded and still range between 0 and 1.

Compared to NAE, NAE+ does not increase the number of parameters. It only adds a small overhead on computation, while improving the person search accuracy as demonstrated by experiments.

4. Experiments

In this section, we perform a thorough evaluation of our NAE and NAE+. We begin by introducing the datasets and evaluation protocols, after which we describe the implementation in detail. Comprehensive analysis and visual inspections are conducted to explore the efficacy of our method. We further compare our method with the state-of-the-arts w.r.t. both search performance and running speed.

4.1. Datasets and Settings

CUHK-SYSU [44] is a hybrid dataset consisting of city scenes shot by a moving camera and screenshots of movies. A total of 18,184 uncropped images and 96,143 bounding boxes are collected, among which 11,206 images and 55,272 pedestrians are used for training. The testing set includes 2,900 query persons and 6,978 gallery images. For each query, different gallery sizes are defined by the benchmark to assess the scaling ability of different models. If not specified, we use the gallery size of 100 by default.

PRW [61] is extracted from video frames recorded by 6 stationary cameras that are installed at different locations in a university campus. There are 11,816 frames with 43,110 bounding boxes, where 34,304 of them are annotated with 932 identities and the rest marked as unknown identities. In the training set, there are 5,704 images with 482 identities. The testing set contains 2,057 query persons and each of them are to be searched in a gallery with 6,112 images. Therefore, the gallery size is significantly larger than the default setting of CUHK-SYSU.

Evaluation Protocol. Similar to person re-ID [60], Mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC top-K) are standard metrics used to measure person search performance. However, a candidate in the ranking list would only be considered correct if its IoU to the ground truth bounding box is larger than 0.5, which is the main difference from the re-ID approach.

4.2. Implementation Details[1]

Our model consists of three major parts: a stem network for spatial feature extraction, a region proposal network (RPN) for candidate bounding box sampling and a head network for proposal classification and regression.

We adopt an ImageNet-pretrained [5] ResNet-50 [18] as our backbone network, with the foremost four residual blocks, *i.e.*, 'conv1' to 'conv4', used as the stem network.

A standard RPN is built on top of the stem network to generate pedestrian candidate bounding boxes. We follow the anchor settings in [25] and sample the positive proposals with a lower bound IoU of 0.5 to the ground truth, and the IoU interval for negative proposals is $[0.1, 0.5)$.

Next, the proposals are cropped and reshaped to 14×14 by an RoIAlign layer [17]. The head network, which is the 'conv5' residual block of ResNet-50, is used to transform the proposals into 2048-dimensional 7×7 feature maps. Task-specific heads for bounding box regression and norm-aware embedding generation are added on top of the feature maps. We set the spatial size k to 7 for NAE+, which is depicted in Fig. 3(c).

During training, we sample 5 images for each batch, which are resized to $900 \times 1{,}500$. Our model is trained on a single NVIDIA Tesla P40 GPU for 22 epochs, with an initial learning rate of 0.003 which is progressively warmed up during the first epoch and decreased by 10 at the 16-th epoch. The momentum and weight decay of SGD are set to

[1] https://github.com/DeanChan/NAE4PS

Detector	Recall	AP	Re-idetntifier	mAP	top-1
OIM-base	89.3	79.7	OIM-base	84.4	86.1
			NAE	90.0	91.8
NAE	92.6	86.8	OIM-base	85.9	87.6
			NAE	91.5	92.4
GT	100	100	OIM-base	90.7	91.2
			NAE	93.5	94.0

Table 1. Analytical experiment results on CUHK-SYSU. The upper block uses the detected boxes of OIM-base, while the lower block uses the NAE detection results.

Method	mAP	top-1	ΔmAP	Δtop-1
OIM-base	84.4	86.1		
OIM-base w/ CWS	87.1	88.5	+2.7	+2.4
NAE	91.5	92.4		
NAE w/o CWS	89.9	91.3	-1.6	-1.1

Table 2. Ablation experiments on Class Weighted Similarity.

0.9 and 5×10^{-4} respectively. As for NAE+, we initialize the weights with a trained NAE model by converting the FC layer weights into 1×1 convolution weights. It is then fine-tuned for 11 epochs. The learning rate is set to 0.003 for the first 8 epochs, and then decayed to 0.0003 for the remaining 3 epochs.

At test time, the number of proposals is set to 300. Non-maximum Suppression [14] with a threshold of 0.4 is used to filter out redundant boxes.

4.3. Analytical Experiments

As mentioned in the introduction section, person search accuracy is affected by both the detection quality and the identity recognition accuracy. In order to better understand how well our NAE method handles the above two sub-tasks, we disentangle the person search into detection and re-ID and evaluate their performances individually.

We implement the analysis on our norm-aware embedding and the OIM baseline model. Four variants are evaluated, namely

- OIM-base: Our re-implementation of the OIM model [44] which shares the same architecture settings as our NAE model described in Sec. 4.2. Benefiting from large input image size [56], dense anchor setting [25] and RoIAlign [17], our OIM-base is significantly better than the original implementation.
- OIM-base w/ CWS: Using the trained model of OIM-base and apply Class Weighted Similarity [61] when matching gallery persons to the query.
- NAE: Our norm-aware embedding model as described in Sec. 3.1.
- NAE w/o CWS: Identical to NAE but only using the normalized embedding $\boldsymbol{\theta}$ without the scale operation (the vector marked in green in Fig. 3(b)).

All models are trained on CUHK-SYSU and tested under a gallery size of 100.

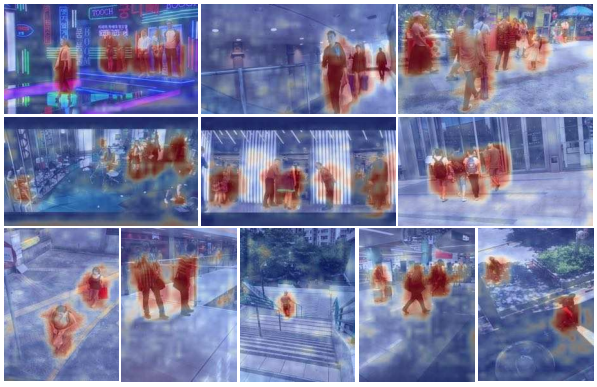


Figure 5. Pixel-wise norm predictions of NAE+ on CUHK-SYSU. Warmer color represents larger norm, which indicates a higher probability of this position being a person. The detection performance of NAE+ is 93.0% and 82.1% w.r.t. Recall and AP, remaining similar to to that of NAE.

For pedestrian detection, we use Recall and Average Precision (AP) as the performance metrics. For person re-ID, mAP and top-1 accuracy are adopted. They are the same as in person search, but the embeddings for matching are extracted differently. We remove the RPN of the trained model and set the proposals manually with the boxes to be inspected. Therefore, an end-to-end person search model serves solely as a re-ID feature extractor.

The evaluation results are collected in Tab. 1 and Tab. 2, from which we make the following conclusions.

The detection quality of NAE is better. The detection results are recorded in the second column of Tab. 1, from which we can see that our NAE model achieves 92.6% and 86.8% w.r.t. Recall and AP, surpassing OIM-base by 3.3 and 7.1 pp. respectively. The better detection quality indicates that the detection objective in NAE is optimized more smoothly and effectively than in our OIM-base. The final person search performance of NAE is also better than OIM-base, thanks to the high-quality bounding boxes.

NAE is more discriminative for re-identification. In the lower block of Tab. 1, we can see that NAE achieves 91.5% and 92.4% w.r.t. mAP and top-1, outperforming OIM-base with NAE detected boxes by 5.6 and 4.8 pp. The performance improvement also holds when switching the bounding boxes to the ground truth boxes or OIM-base detections, as is shown in the upper and lower block of Tab. 1. These results suggest that NAE has better re-ID accuracy, which indicates that the discrimination power of NAE is superior to OIM.

Class Weighted Similarity is helpful. In Tab. 2, we can see that adding CWS to OIM-base yields a gain of $+2.7$ and

	Method	CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
one-step	OIM [44]	75.5	78.7	21.3	49.9
	IAN [42]	76.3	80.1	23.0	61.9
	NPSM [26]	77.9	81.2	24.2	53.1
	RCAA [2]	79.3	81.3	-	-
	CTXGraph [47]	84.1	86.5	33.4	73.6
	QEEPS [30]	88.9	89.1	37.1	76.7
	OIM-base (ours)	84.4	86.1	34.0	75.9
	NAE (ours)	91.5	92.4	43.3	80.9
	NAE+ (ours)	**92.1**	**92.9**	**44.0**	**81.1**
two-step	DPM+IDE [61]	-	-	20.5	48.3
	CNN+MGTS [3]	83.0	83.7	32.6	**72.1**
	CNN+CLSA [21]	87.2	88.5	38.7	65.0
	FPN+RDLR [16]	**93.0**	**94.2**	**42.9**	70.2

Table 3. Comparison with state-of-the-arts. One-step methods are gathered in the upper block while two-step methods are in the lower block. Best results in each block are marked in **bold**.

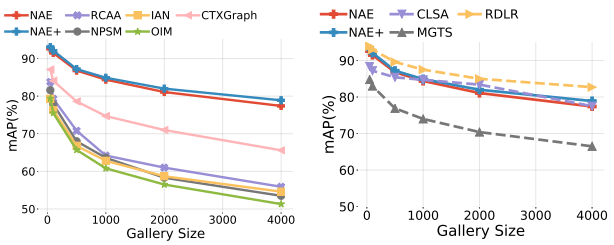
+2.4 pp. for mAP and top-1 respectively. Meanwhile, removing CWS from NAE makes mAP and top-1 drop from 91.5 to 89.9 and 92.4 to 91.3. These results confirm the positive efficacy of CWS. As a naturally induced form of NAE, CWS also contributes to the person search performance of our method.

In conclusion, our norm-aware embedding successfully alleviates the contradictory objectives of detection and re-ID by decomposing embedding explicitly into norm and angle. The detection and re-ID sub-tasks both achieve better results than the baseline. As a result, the final person search performance of our method is remarkable, which can be attributed to the improvements on the two sub-tasks.

4.4. Visualized Inspections

To inspect the efficacy of the NAE+ method, we visualize the output probability maps in Fig. 5. Specifically, we remove the RPN and RoIAlign modules from a trained NAE+ model and forward the input image directly through the whole network. The output probability map, composed of the mapped norms $\tilde{r}_i$ at each location, is upsampled with bilinear interpolation to match the input image size. We then represent the probability maps with different colors and overlay them with the corresponding input images. We observe from Fig. 5 that NAE+ successfully highlights the human body region and suppresses background clutters, which makes the embedding more robust to noises. As is shown in Tab. 3, NAE+ outperforms NAE consistently on CUHK-SYSU and PRW.

We also show some qualitative search results in Fig. 7. The selected cases are representative hard ones, including crowd overlapping (case a, f), confusing appearance (c, d, f, g), viewpoint change (c, d, f, g) and obstacle occlusion



(a) (b)

Figure 6. Performance comparison on CUHK-SYSU with varying gallery sizes. Dashed lines represent two-step methods while solid lines denote one-step methods.

GPU (TFLOPs)	MGTS	QEEPS	NAE	NAE+
K80 (4.1)	1269	-	663	606
P6000 (12.6)	-	300	-	-
P40 (11.8)	-	-	158	161
V100 (14.1)	-	-	83	98

Table 4. Speed comparison on different GPUs. Running time is measured in milliseconds.

(e, g). Our NAE method successfully localize and match the query person in most of the hard cases, although there is still room to improve the performance on extreme instances like (f) and (g). Moreover, our NAE+ method is better than NAE as it returns the correct result for all scenarios.

4.5. Comparison to the State-of-the-arts

In this section, we compare our NAE and NAE+ to state-of-the-art methods on person search in Tab. 3. All the results are gathered according to their search strategies, *i.e.*, one-step methods in the upper block and two-step method in the lower block. 'DPM', 'CNN' and 'FPN' stand for Deformable Part Model [14], ResNet-50-based Faster R-CNN [34] and Feature Pyramid Network [25] respectively. They are individually trained as vanilla pedestrian detectors.

Comparison on CUHK-SYSU. As shown in Tab. 3, both NAE and NAE+ outperform all other one-step methods, including the strong counterparts QEEPS [30] and CTXGraph [47]. Note that their forward pass requires some computationally heavy operations, *e.g.* siamese attention and additional graph convolutions. In contrast, our method only needs a single forward pass, consuming less computing resources and memory. Our method is also comparable to the top two-step method 'FPN + RDLR' [16], which uses two backbones for detection and re-ID respectively. We believe the performance-boosting components of [16], *i.e.*, feature pyramid network, RoI transform layer and proxy triplet loss, could also bring improvements to our method, which is however beyond the scope of this paper.

In Fig. 6, we further evaluate the performance under larger search scopes. As is defined in [44], each query per-

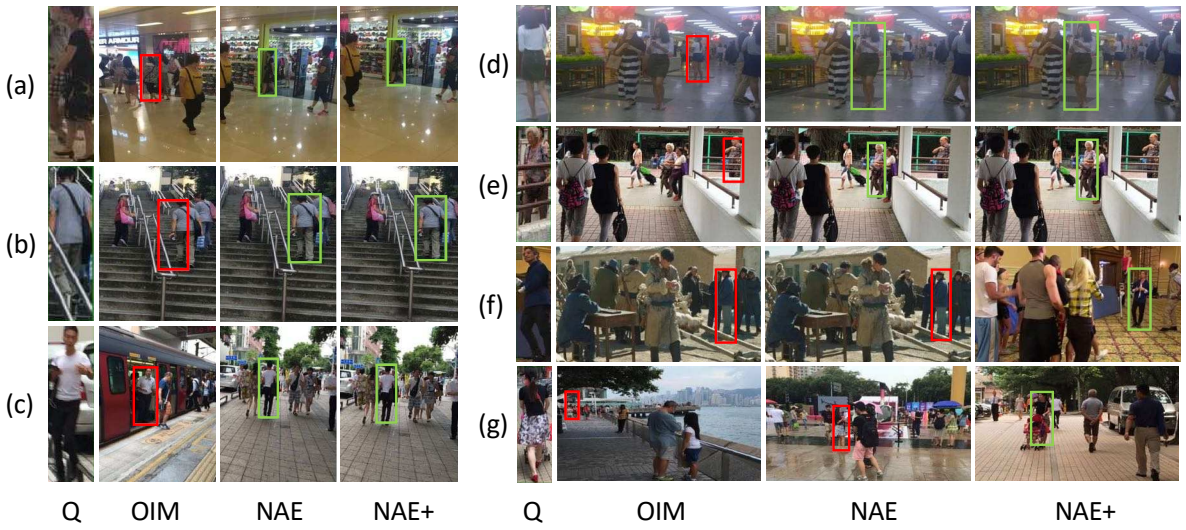


Figure 7. Top-1 search results for several hard samples. 'Q' stands for the query image, for each we show the top-1 match given by OIM-base, our NAE and NAE+. Green/red boxes denote the correct/wrong results respectively. (a)∼(e) are cases where OIM-base fails while NAE and NAE+ succeed. (f) and (g) are failure cases for both OIM-base and NAE, except for NAE+.

son is matched in galleries with an increasing size. From Fig. 6 we can see that the mAP for all methods decrease monotonically as the gallery size becomes larger. This phenomenon indicates that it is more difficult to match a person in larger scopes, which is a typical challenge in real-world applications. Our method outperforms all the one-step methods by a considerable margin, while achieving similar mAP to the two-step methods at all scopes.

Comparison on PRW. In the right column of Tab. 3, we summarize the results of our NAE and NAE+ together with other competitive methods. Our NAE method surpasses all previous methods, including both one-step and two-step ones. In particular, our NAE outperforms the second best method by a large margin of around 9 pp. w.r.t. top-1 accuracy. Compared to CUHK-SYSU, PRW consists of less training data and larger gallery size, thus it is more challenging. Our NAE method behaves better on PRW, indicating that our method is more robust with reduced training data. Moreover, the pixel-wise extension NAE+ further improves over NAE by 0.7 and 0.2 pp. w.r.t. mAP and top-1 metrics, setting the new state-of-the-art on PRW.

Timing. We compare the speed of different methods in Tab. 4. Since different methods are implemented on different GPUs, we show the Tera-Floating Point Operation per-second (TFLOPs) beside each GPU for fair comparison. Our NAE and NAE+ are implemented in PyTorch [33] without bells and whistles. We test the models with an input image size as 900×1500, which is the same as MGTS and QEEPS [30]. We can see from Tab. 4 that our method is around 2 times faster than the two-step method MGTS [3]. Our method is also 2 times faster than QEEPS, which is the current state-of-the-art one-step method. Finally, our NAE and NAE+ methods cost 83 and 98 milliseconds per-frame respectively on a V100 GPU. The fast speed of our method reveals its great potential for real-world applications.

5. Conclusion

In this paper, we propose an embedding decomposing method to deal with the contradictory objective problem of person search. Person embeddings are disintegrated into norm and angle, which are used to measure the detection confidence and identity similarity accordingly. In this way, the detection and re-ID sub-tasks both get higher performance, which in result improves the person search accuracy. We further extend our method from region-level to pixel-level in order to extract more fine-grained information. Thorough experiments on two standard benchmarks confirm the advantages of our method in terms of both accuracy and speed.

Acknowledgement

The authors would like to thank the AC and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Science Fund of China (Grant Nos. U1713208, 61702262), Funds for International Co-operation and Exchange of the National Natural Science Foundation of China (Grant No. 61861136011), "111" Program B13022, Natural Science Foundation of Jiangsu Province, China (Grant No. BK20181299), Young Elite Scientists Sponsorship Program by CAST (2018QNRC001), and Science and Technology on Parallel and Distributed Processing Laboratory (PDL) Open Fund (WDZC20195500106).

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G. Hauptmann. Rcaa: Relational context-aware agents for person search. In *ECCV*, 2018.

[3] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *ECCV*, 2018.

[4] 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 *CVPR*, 2016.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*.

[7] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *PR*, 48(10):2993–3003, Oct 2015.

[8] Piotr Dollar, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *TPAMI*, 36(8):1532–1545, Aug 2014.

[9] Piotr Dollar, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. In *BMVC*, 2009.

[10] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. *arXiv preprint arXiv:1807.00537*, 2018.

[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[12] Pedro F Felzenszwalb, Ross B Girshick, David Mcallester, and Deva Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 32(9):1627–1645, Sept 2009.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[14] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *CVPR*, 2015.

[15] Yandong Guo and Lei Zhang. One-shot face recognition by promoting underrepresented classes. *arXiv preprint arXiv:1707.05574*, 2017.

[16] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *ICCV*, 2019.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[20] Martin Kostinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[21] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *ECCV*, 2018.

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[23] Xiang Li, Wei Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 2015.

[24] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[26] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *ICCV*, 2017.

[27] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 26(7):3492–3506, July 2017.

[28] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.

[29] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, 2016.

[30] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *CVPR*, 2019.

[31] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012.

[32] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013.

[33] 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. In *NIPS-W*, 2017.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, June 2017.

[35] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.

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

[37] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[38] Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Zhifeng Li, Wei Liu, and Tong Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *ECCV*, 2018.

[39] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM'MM*, 2017.

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

[41] Wangmeng Xiang, Jianqiang Huang, Xianbiao Qi, Xian-Sheng Hua, and Lei Zhang. Homocentric hypersphere feature embedding for person re-identification. *arXiv preprint arXiv:1804.08866*, 2018.

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

[43] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[44] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.

[45] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[46] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *ACM'MM*, 2014.

[47] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *CVPR*, 2019.

[48] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *TIP*, 28(6):2860–2871, June 2019.

[49] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[50] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[51] Shanshan Zhang, Christian Bauckhage, and Armin B. Cremers. Informed haar-like features improve pedestrian detection. In *CVPR*, 2014.

[52] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *CVPR*, 2016.

[53] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. Towards reaching human performance in pedestrian detection. *TPAMI*, 40(4):973–986, April 2018.

[54] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, 2015.

[55] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017.

[56] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *CVPR*, 2018.

[57] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[58] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

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

[60] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[61] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, 2017.