ASMR: Learning Attribute-Based Person Search with Adaptive Semantic Margin Regularizer

Boseung Jeong[1*] Jicheol Park[2*] Suha Kwak[1,2]
Dept. of CSE, POSTECH[1], Graduate School of AI, POSTECH[2]
http://cvlab.postech.ac.kr/research/ASMR/

Abstract

Attribute-based person search is the task of finding person images that are best matched with a set of text attributes given as query. The main challenge of this task is the large modality gap between attributes and images. To reduce the gap, we present a new loss for learning cross-modal embeddings in the context of attribute-based person search. We regard a set of attributes as a category of people sharing the same traits. In a joint embedding space of the two modalities, our loss pulls images close to their person categories for modality alignment. More importantly, it pushes apart a pair of person categories by a margin determined adaptively by their semantic distance, where the distance metric is learned end-to-end so that the loss considers importance of each attribute when relating person categories. Our loss guided by the adaptive semantic margin leads to more discriminative and semantically well-arranged distributions of person images. As a consequence, it enables a simple embedding model to achieve state-of-the-art records on public benchmarks without bells and whistles.

1. Introduction

Person search is the task of finding people from a large set of images given a query describing their appearances. It plays critical roles in applications for public safety such as searching for criminals in videos and tracking people using multiple surveillance cameras with non-overlapping fields of view. Person search has been formulated as a fine-grained image retrieval problem focusing only on person images, where a solution should discriminate subtle appearance variations of different people and at the same time generalize well to people unseen during training.

Most of existing person search techniques need an image that exemplifies target person as query [3, 4, 5, 17, 18, 20, 24, 27, 30, 33, 38, 48, 50, 52]. However, image query is not always accessible in real world scenarios, *e.g.*, where eyewitness memory is the only evidence for finding criminals. A solution to this issue is to utilize a verbal description as query for person search [22, 23], but it suffers from the inherent ambiguity of natural language and requires complicated processes to understand the query.

To address the above issue, we study in this paper person search using text attributes as query. Specifically, a query is given as a set of predefined attributes indicating traits of target person, *e.g.*, gender, age, clothing, and accessory; we consider such a set as a *person category*, and multiple people sharing the same traits belong to the same person category. This approach is suitable for person search in the wild since attributes are cheap to collect while being less ambiguous and more tractable than natural language descriptions. The use of attributes as query, however, introduces additional challenges due to the limited descriptive capability of attributes, which leads to a large modality gap between images and person categories.

Previous work on attribute-based person search attempts to reduce the modality gap by aligning each person category and corresponding images in a joint embedding space through modality-adversarial training [2, 51] or by enhancing the expressive power of embedding vectors of person categories and images in a hierarchical manner [9]. Although these pioneer studies shed light on the important yet less explored approach to person search, there is still large room for further improvement. First, they are unstable and computationally heavy in training due to their adversarial learning strategies [2, 51], or expensive in inference due to the large dimensional embedding vectors demanding an extra network to be matched [9]. More importantly, these methods treat person categories as independent class labels of person images and ignore their relations, *e.g.*, how many attributes are different between them, although such relations can provide a rich supervisory signal for learning better representations of person categories and images.

We develop a new attribute-based person search method that overcomes these limitations. Our method learns a joint embedding space of the two different modalities through a pair of simple encoder networks, one for images and the

[*] Equal contribution

other for person categories; a person category is represented as a binary vector, each of whose dimensions indicates the presence of the corresponding attribute. When conducting person search, a person category is given as query in the form of binary vector and projected onto the joint embedding space by the person category encoder, then images whose embedding vectors are closest to that of the query in the space are retrieved.

The main contribution of this work is a new loss function, which enables our model to achieve outstanding performance with the simple architecture and retrieval pipeline. In the joint embedding space, the loss regards each person category as a semantic prototype of associated images, and encourages the images to be close to their prototype so that the two modalities are aligned. The key feature of the loss is that it determines the margin between person categories in the embedding space adaptively by their distance in the binary attribute space. Moreover, the distance is measured by weighted Hamming metric, in which weights multiplied to individual bits (*i.e.*, attributes) are optimized together with parameters of the embedding networks so that the loss focuses on more important attributes when relating person categories. This idea is implemented by Adaptive Semantic Margin Regularizer (ASMR) as a part of our loss.

The proposed loss function with ASMR allows the distributions of person images to be more discriminative and semantically well-arranged in the learned embedding space. Consequently, our method achieves the state of the art on three public benchmark datasets [8, 25, 29] without bells and whistles. Also, compared to the previous work [2, 9, 51], it is efficient since it works on an embedding space of a small dimension with no extra network, and converges very quickly in training since it does not require adversarial training. The main contribution of our work is three-fold:

- We propose a novel cross-modal embedding loss, considering semantic relations between person categories so that the embedding space becomes more discriminative and better generalizes to unseen categories.
- The straightforward architecture and retrieval pipeline of the proposed framework enable fast convergence in training and efficient person search in testing.
- Our method achieves the state of the art on three public benchmarks without bells and whistles.

2. Related Work

2.1. Attribute-Based Person Search

A naïve approach to attribute-based person search is recognizing attributes of person images and finding images whose predicted attributes are the same with the person category given as query [21, 34, 40]. However, this approach is unreliable due to imperfection of attribute recognition. Note that attribute recognition itself is challenging since the appearance of an attribute could vary significantly and person images captured by surveillance cameras are often limited in terms of resolution and quality.

Recent methods instead learn and utilize a cross-modal embedding space where person categories and associated images are close to each other. The main issue in this direction is the large gap between the two modalities. Dong *et al.* [9] tackle the problem by capturing rich information of the two modalities through hierarchical embeddings. However, their model is computationally heavy since it computes high dimensional embeddings and deploys an extra network for matching them. Yin *et al.* [51] and Cao *et al.* [2] learn a joint embedding space where person categories and images are matched directly. To bridge the modality gap, their embedding spaces are trained in modality-adversarial manners, which however often result in unstable and tardy convergence due to the nature of the minimax optimization. Moreover, these methods share a limitation that person categories are considered as individual class labels and their nontrivial relations are ignored.

Our method also learns a cross-modal embedding space, but unlike the previous arts, it is efficient in both training and testing, and lets the learned embedding space reflect semantic relations between person categories.

2.2. Deep Metric Learning

The goal of deep metric learning is to learn an embedding space where data of the same class are grouped together and those of different classes are pushed away. Loss functions for metric learning are roughly categorized into two classes, pair-based and proxy-based losses.

Pair-based losses basically pull a pair of embedding vectors close to each other if they are of the same class and push them apart otherwise. An early example following this principle is contrastive loss [1, 6, 14], which is extended to consider higher order relations of embedding vectors by associating multiple pairs [35, 36, 37, 44, 47]. On the other hand, proxy-based losses relate embedding vectors with prototypes, each of which is a virtual embedding vector typifying each class of training data and learned as a part of embedding network. Then the losses pull together or push apart a pair of embedding vector and prototype according to their class equivalence [7, 19, 31].

Unfortunately, these losses are not proper to be applied directly to attribute-based person search for the following reasons. First, most of them are developed for uni-modal retrieval, except few examples [12, 26]. Second, they cannot take semantic relations between person categories into account since they regard the categories as independent labels whose relations are binary (*i.e.*, the same or not).

Unlike the existing losses for metric learning, our loss can handle the nontrivial inter-category relations as well as

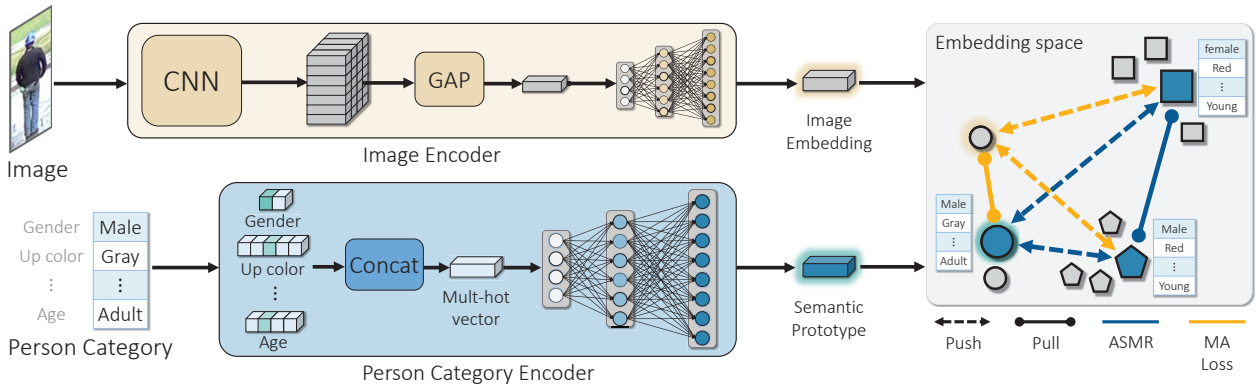


Figure 1. Overall pipeline of our method. Image is embedded by a conventional CNN followed by a MLP while the query set of attributes, called *person category*, is converted to a binary vector and encoded through a separate embedding network. In their joint embedding space, a positive pair of image embedding and semantic prototype are pulled together while a negative pair is pushed apart for cross-modal alignment. Also, a pair of semantic prototypes pushes or pulls each other by a margin determined adaptively by their semantic affinity.

those between categories and images thanks to ASMR. We believe that our loss can be applied to other tasks where inter-label relations are beyond the binary.

2.3. Cross-Modal Retrieval

Attribute-based person search is a particular example of cross-modal retrieval, which has been studied mainly for image-text or image-sound retrieval [10, 32, 41, 42, 45, 46]. Most of existing methods for cross-modal retrieval aim to learn a joint embedding space of different modalities so that a simple nearest neighbor search can find samples of the same content in the space regardless of their modalities. This idea has been implemented in general by Canonical Correlation Analysis (CCA) [15] or Generative Adversarial Networks (GANs) [13]. Specifically, methods based on CCA attempt to project samples of different modalities into a common embedding space by maximizing their correlation [10, 45, 46, 49], and those based on GANs align samples of different modalities by learning modality-adversarial embeddings [41, 42, 51]. Unfortunately, these methods cannot consider semantic relations between classes.

This paper shows that the prototype-based embedding learning is fairly effective for cross-modal retrieval. Also, unlike the previous work, our method can consider relations between categories, improving performance substantially.

3. Our Method

In attribute-based person search, a set of attributes, called *person category*, describes traits of people we want to find. Given a person category as query, our method conducts person search by finding images that are closest to the query in a joint embedding space of person images and categories. It learns the embedding space through two encoders, one for images and the other for person categories; an overview of the architecture is given in Fig. 1.

The key contribution of our work is the loss function used for training the networks. In the embedding space, the loss pulls a person category and its associated images together, and at the same time, pushes apart a pair of person categories by a margin determined adaptively by their semantic dissimilarity. Our model achieves outstanding performance and converges quickly thanks to the proposed loss, and is computationally efficient due to the straightforward model architecture and retrieval pipeline.

The remainder of this section first describes details of the model architecture and its pretraining, then elaborates the proposed loss function and discusses its advantages.

3.1. Model Architecture and Its Pretraining

In the image encoder of our model, a conventional CNN extracts a feature map of input person image, which is in turn transformed to a single feature vector by Global Average Pooling (GAP) and fed to a Multi-Layer Perceptron (MLP) that produces image embedding. Meanwhile, the person category encoder is a MLP that takes person category as input and produces person category embedding. Outputs of the two encoders are all ℓ_2 normalized.

Since a person category is a set of text attributes, it has to be converted in a numerical form to be fed to its encoder. To this end, it is given in a form of binary vector, each of whose dimension indicates the presence of corresponding attribute. Suppose that attributes are grouped exclusively into a number of *attribute groups*; for example, two attributes male and female belong to the same attribute group gender. As a person can take only one attribute for each attribute group, an attribute is represented by a one-hot vector whose dimension is the same with the number of attributes in its group. The binary vector representation of a person category is then obtained by concatenating such one-hot vectors of its all attributes.

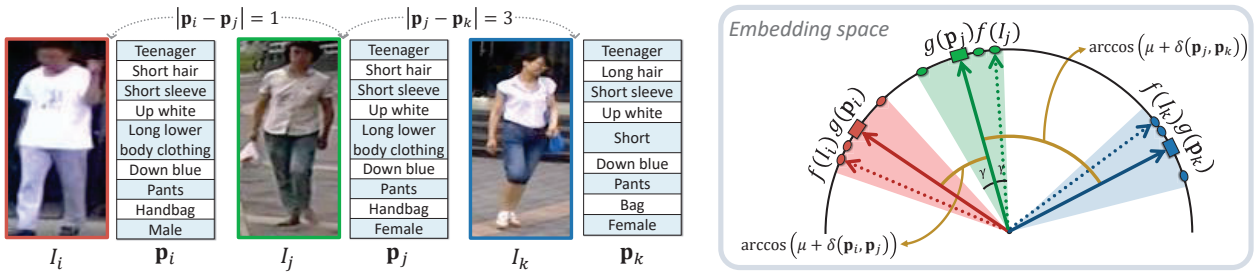


Figure 2. A conceptual illustration of the learning objective in Eq. (1). The modality alignment loss pulls images close to their person categories within the margin γ. Meanwhile, ASMR controls margins between person categories according to their semantic affinities.

Parameters of our model are initialized randomly, except those of the CNN for which we adopt ImageNet pretrained parameters. Unfortunately, the weights for ImageNet classification are suboptimal for capturing subtle appearance features of person images. We thus pretrain the image encoder for attribute classification, an auxiliary task for learning image representation more suitable to person search.[1] Specifically, we append a classification head of four Fully Connected (FC) layers on top of the GAP for each attribute group. Then each classification head is trained together with the backbone CNN for choosing the correct attribute among those in its attribute group through a multi-class classification loss; we adopt the softmax cross-entropy loss for this purpose. After the pretraining, a randomly initialized MLP replaces the attribute classification heads.

3.2. Learning Objective

The loss for our model consists of two parts. One of them is a modality alignment loss designed to group embedding vectors of images together around that of their person category for cross-modal alignment. The other is Adaptive Semantic Margin Regularizer (ASMR) that controls the margin between a pair of embedding vectors of person categories according to their semantic dissimilarity. The roles of the two components are illustrated in Fig. 2.

Let f and g be the encoders for images and person categories, respectively. Training data for learning the encoders are provided by a set of images paired with binary vectors indicating their person categories, $\mathcal{D} = \{I_i,\ \mathbf{p}_i\}_{i=1}^m$, where m is the number of training images. In addition, let $\mathcal{G}$ denote the set of embedding vectors of unique person categories in the training set. Given embedding vectors of images $\mathbf{f}_i := f(I_i)$ and those of person categories $\mathbf{g}_i := g(\mathbf{p}_i)$, the learning objective for our model is a linear combination of the two terms as follows:

$$\mathcal{L}\big(\{\mathbf{f}_i,\ \mathbf{g}_i\}_{i=1}^m\big) = \mathcal{L}_{\text{MA}}\big(\{\mathbf{f}_i,\ \mathbf{g}_i\}_{i=1}^m\big) + \lambda\, \mathcal{R}\big(\mathcal{G}\big), \quad (1)$$

where $\mathcal{L}_{\text{MA}}$ indicates the modality alignment loss, $\mathcal{R}$ means the ASMR, and λ is a weight hyper-parameter. Details of the two components are described below.

[1] For the same reason, existing methods also take advantage of the attribute classification by adopting it as an auxiliary task [9] or using it for pretraining their models [2].

3.2.1 Modality Alignment Loss

The role of the modality alignment part is to align the two different modalities in a common embedding space. Considering each person category embedding as a semantic prototype of associated image embeddings, the cross-modal alignment is done by pulling image embeddings close to their semantic prototypes and pushing them apart from irrelevant prototypes. This idea is formulated as

$$\mathcal{L}_{\text{MA}}\big(\{\mathbf{f}_i,\ \mathbf{g}_i\}_{i=1}^m\big) = -\frac{1}{m}\sum_{i=1}^{m} \log\left(\frac{e^{\sigma\cos(a(\mathbf{f}_i,\mathbf{g}_i)+\gamma)}}{e^{\sigma\cos(a(\mathbf{f}_i,\mathbf{g}_i)+\gamma)} + \sum\limits_{\mathbf{g}_k\in\mathcal{G}\setminus\mathbf{g}_i} e^{\sigma\cos a(\mathbf{f}_i,\mathbf{g}_k)}}\right), \quad (2)$$

where $a(\cdot,\cdot)$ means the angle between its two input vectors, $\sigma > 0$ is a scale factor, and $\gamma > 0$ is a margin between image and person category embeddings. The above form resembles ArcFace loss [7], yet different in that the person category embeddings used as prototypes are not parameters but outputs of another encoder g in our loss. We empirically found that the simple joint embedding architecture trained solely with this loss is as competitive as previous arts; it can be considered as a simple yet solid baseline for attribute-based person search, and ASMR further improves the performance substantially.

3.2.2 ASMR

For accurate person search and generalization to unseen person categories, we expect from the learned embedding space that different person categories lie apart from each other clearly and their distances are larger if they are more dissimilar, *i.e.*, sharing less attributes. However, the modality alignment loss in Eq. (2) alone does not guarantee this quality of embedding space since it ignores semantic relations between them; the loss considers person categories

simply as independent class labels. One of failure cases regarding this issue is that different person categories are often located overly close to each other in the learned embedding space when images of these categories exhibit subtle appearance variations; an example is given in Fig. 6.

To address this issue, we introduce ASMR that explicitly controls the margin between a pair of person categories according to their semantic dissimilarity. The regularizer is given by

$$\mathcal{R}(\mathcal{G}) = \frac{1}{|\mathcal{G}|C_2} \sum_{i=1}^{|\mathcal{G}|-1} \sum_{j=i+1}^{|\mathcal{G}|} \left\{ s(\mathbf{g}_i, \mathbf{g}_j) - \mu - \delta(\mathbf{p}_i, \mathbf{p}_j) \right\}^2, \tag{3}$$

where $s(\cdot, \cdot)$ denotes the cosine similarity between the two input vectors and μ is the mean similarity over all pairs of person categories in the embedding space:

$$\mu = \frac{1}{|\mathcal{G}|C_2} \sum_{i=1}^{|\mathcal{G}|-1} \sum_{j=i+1}^{|\mathcal{G}|} s(\mathbf{g}_i, \mathbf{g}_j). \tag{4}$$

Also, $\delta(\mathbf{p}_i, \mathbf{p}_j)$ quantifies the semantic similarity of a pair of person categories represented as binary vectors $\mathbf{p}_i$ and $\mathbf{p}_j$, and is formulated as an inverse of weighted Hamming distance:

$$\delta(\mathbf{p}_i, \mathbf{p}_j) = \text{Sigmoid}\left(1 - \sum_k w_k |\mathbf{p}_i(k) - \mathbf{p}_j(k)|\right). \tag{5}$$

Regarding its shape, the sigmoid function lets this similarity margin respond more sensitively to pairs of more similar person categories, which in general have to be handled more carefully for accurate person search.[2] Moreover, the weight parameters w_k are trained together with those of the embedding networks, which enables ASMR to estimate importance of individual attributes and relate person categories in consideration of the importance.

ASMR enforces $s(\mathbf{g}_i, \mathbf{g}_j)$ to approximate $\mu + \delta(\mathbf{p}_i, \mathbf{p}_j)$ so that the degree of similarity between person categories in the binary vector space is reflected by their similarity in the learned embedding space. This behavior of ASMR makes distributions of embedding vectors more discriminative by enlarging the margin between person categories. Also, we believe that it helps our model avoid being biased to image information and generalize better to unseen person categories by reflecting semantic relations between person categories explicitly in the embedding space.

4. Experiments

Our method is evaluated and compared to previous work on three public benchmarks for attribute-based person search [8, 25, 29]. We also demonstrate the effect of ASMR by ablation studies and qualitative analysis.

[2]Person categories sharing more attributes are more likely to be close in the embedding space due to their similar appearances, and to affect accuracy of person search whose goal is to find samples *closest* to query.

Datasets	PETA	Market-1501	PA100K
# Attributes	65	27	26
# Attributes groups	17	10	15
# Train person category	1,890	508	500
# Train image	12,140	12,936	80,000
# Test person category	200	484	814
# Unseen	200	315	168
# Test image	1,181	16,483	10,000

Table 1. Statistics of three benchmarks.

4.1. Datasets

We evaluate our method and previous arts on three public datasets, PETA [8], Market-1501 Attribute [25] and PA100K [29], which are representative benchmarks for attribute-based person search. The dataset statistics are summarized in Table 1. Note that the PETA dataset follows the ordinary image retrieval setting where categories of test images are all unseen, while the other two datasets assume a more general search scenario in which both seen and unseen person categories appear in testing.

4.2. Implementation Details

Network architecture. In the image encoder, the backbone CNN is ResNet-50 [16] and the MLP consists of three FC layers. On the other hand, the person category encoder is implemented only by a MLP with three FC layers. Both of the two encoders produce 128-dimensional embedding vectors that are ℓ_2 normalized. More details of the encoders are presented in the supplementary material.

Hyper-parameters. In every experiment, our model is optimized by SGD with a momentum of 0.9 and a weight decay of $5\text{e}{-4}$ for 10 epochs; each mini-batch consists of 128 images and their person categories. The initial learning rate is set to $1\text{e}{-3}$ for the image encoder, and $1\text{e}{-2}$ for the person category encoder and the parameters of the weighted Hamming distance. Then both learning rates are decayed by a factor of 0.1 at every 5 epochs. The other hyper-parameters, λ in Eq. (1), and σ and γ in Eq. (2) are set to (4, 32, 0.1) on PETA, (6, 12, 0.2) on Market-1501 Attribute, and (5, 48, 0.1) on PA100K, respectively.

4.3. Quantitative Comparison to Previous Work

Our model is compared to the three existing methods for attribute-based person search, AAIPR [51], AIHM [9], and SAL [2]. We also report performance of related models that are not originally proposed for attribute-based person search but have been reproduced for the purpose in literature. Performance of these methods including ours is summarized in Table 2, where Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) are adopted as performance metrics following the convention.

The table shows that our model outperforms all the other methods in terms of Rank1 and mAP metrics. It clearly

Method	Dim	PETA				Market-1501 Attribute				PA100K			
		Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP
DeepMAR [21]	-	17.8	25.6	31.1	12.7	13.2	24.9	32.9	8.9	-	-	-	-
DCCAE [46]	-	14.2	22.1	30.0	14.5	8.1	24.0	34.6	9.7	21.2	39.7	48.0	15.6
2WayNet [10]	-	23.7	38.5	41.9	15.4	11.3	24.4	31.5	7.8	19.5	26.6	34.5	10.6
CMCE [22]	-	31.7	39.2	48.4	26.2	35.0	51.0	56.5	22.8	25.8	34.9	45.4	13.1
AAIPR [51]	128	39.0	53.6	62.2	27.9	40.3	49.2	58.6	20.7	-	-	-	-
AIHM [9]	3K	-	-	-	-	43.3	56.7	64.5	24.3	31.3	45.1	51.0	17.0
SAL [2]	128	47.0	66.5	74.0	41.2	49.0	**68.6**	**77.5**	29.8	-	-	-	-
SAL [2]†	128	39.0	61.5	70.0	37.2	44.4	65.7	72.5	29.4	22.7	36.5	41.6	15.0
Ours	128	**56.5**	**80.0**	**83.5**	**50.2**	**49.6**	64.9	72.5	**31.0**	**31.9**	**49.1**	**58.2**	**20.6**

Table 2. Quantitative comparison to previous arts. Dim indicates embedding dimensions of the methods based on cross-modal embeddings. **Bold** and underline denote the best and the second-best, respectively. † indicates results reproduced by the official implementation.

Method	PETA	Market-1501	PA100K
SAL [2]	202	211	957
Ours	27	18	110

Table 3. Comparison of training time (min)

surpasses AIHM [9], the state of the art in PA100K, for all available settings. This achievement is remarkable since our method is more efficient than AIHM; it works with embedding vectors of a substantially smaller dimension, and unlike AIHM, it does not require any extra network for retrieval. Moreover, our method outpaces SAL [2], the state of the art in PETA and Market-1501 Attribute, for almost all settings. Especially, it outperforms SAL on the PETA dataset by a large margin, 9.5% in Rank1 and 9.0% in mAP. On the Market-1501 Attribute dataset, it is more accurate than SAL in terms of Rank1 and mAP, although its records in Rank5 and Rank10 are slightly below those of SAL.

The key to this success of our method is two-fold. The first is MA loss. Since the loss compares each image embedding with those of all person categories in the dataset, it enables to learn more discriminative embedding space more efficiently. Meanwhile, the loss of AIHM considers images and person categories within a mini-batch only. Another cause is the person category encoder, which encodes person categories in an attribute-aware manner so that the embedding space reflects their semantic relations. On the other hand, SAL represents person categories as independent network parameters. The last yet most vital cause is ASMR, whose efficacy is validated in Sec. 4.6.

The reasons for the small improvement on the Market-1501 Attribute and PA100K datasets are as follows. Compared to the PETA dataset, these datasets assume a more challenging search scenario in which both seen and unseen person categories appear in testing. Further, in the Market-1501 dataset, incorrect attribute labels bind the performance; this happens because the labels are annotated not per image but per video, *e.g.*, a man labeled with "jacket" may take off his jacket in the middle of video.

In addition, compared to SAL, our model is significantly more efficient in training. SAL requires a large training time for convergence due to its adversarial learning strategy. In contrast, our method is trained simply by supervised learning with the loss function in Eq. (1). In consequence, ours using a single GPU converges more than 7.5 times faster than SAL using two GPUs as shown in Table 3.

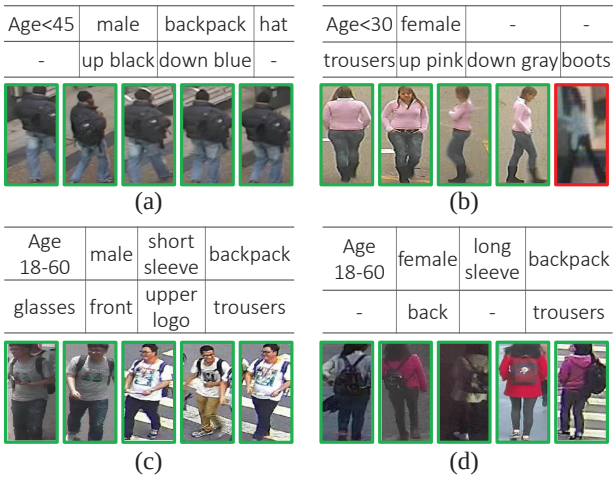


Figure 3. Top 5 retrieval results of our method on (a, b) the PETA and (c, d) PA100K datasets. Images are sorted from left to right according to their ranks. Green and red boxes indicate true and false matches, respectively. Queries are given as tables, where blanks indicate attributes that do not exist in the query.

4.4. Qualitative Analysis

Qualitative results of the proposed method are given in Fig. 3 and Fig. 4. All the presented results demonstrate that our method is insensitive to severe variations in human and camera poses. Moreover, individual examples show that our method is robust against changes in image resolution (Fig. 3(b,c,d), Fig. 4(a,b,c)), illumination (Fig. 3(c,d), Fig. 4(a,b,c)), and partial occlusions (Fig. 3(a,c,d)). It is also demonstrated that the proposed method is able to capture fine details of images for precise retrieval; examples include backpack in Fig. 3(a,c,d) and in Fig. 4(a,b), hat in Fig. 3(a) and Fig. 4(c) and glasses and clothing pattern in Fig. 3(c). More qualitative results can be found in the supplementary material.

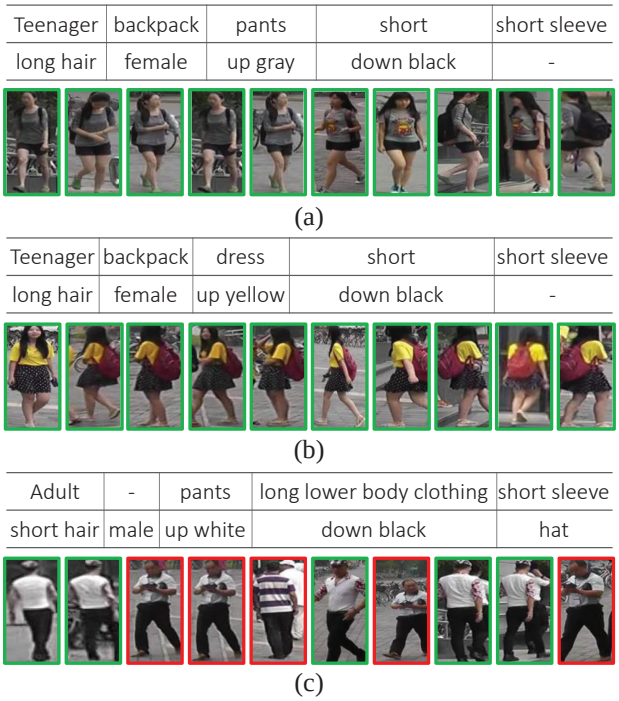


Figure 4. Top 10 retrieval results of our method on the Market-1501 Attribute dataset. Images are sorted from left to right according to their ranks. Green and red boxes indicate true and false matches, respectively. Queries are presented above their retrieved images; blanks indicate attributes that do not exist in the query.

Method		PETA	Market	PA100K
(a)	Baseline	46.5	30.4	26.0
	Baseline + pretraining	48.5	44.8	28.9
(b)	$\mathcal{L}_{\text{MA}} \rightarrow$ Proxy Anchor [19]	48.0	41.1	27.4
	$\mathcal{L}_{\text{MA}} \rightarrow$ Proxy NCA [31]	52.0	43.8	29.7
	$\mathcal{L}_{\text{MA}} \rightarrow$ CosFace [43]	50.5	45.3	24.9
	$\mathcal{L}_{\text{MA}} \rightarrow$ SphereFace [28]	52.5	45.0	23.8
Ours		**56.5**	**49.6**	**31.9**

Table 4. Performance in Rank@1 of ours and its variants on the PETA, Market-1501 Attribute, and PA100K datasets.

4.5. Ablation Studies

Effects of pretraining and ASMR. We quantify the effects of our pretraining strategy and ASMR by evaluating two reduced versions of our method with and without them. To this end, we first define a baseline as the model with the same architecture as ours yet trained only with $\mathcal{L}_{\text{MA}}$ in Eq. (2); the other variant is obtained by adding the pretraining to the baseline. The results in Table 4(a) suggest that the contribution of ASMR is significant and the pretraining also helps to some extent. In detail, ASMR contributes to the performance, enhancing Rank1 by 8.0% on PETA, 4.8% on Market-1501 Attribute, and 3.0% on PA100K, respectively. These results suggest that ASMR makes the learned embedding space more discriminative and better generalized to unseen categories. Also, the pretraining improves Rank1 by 2.0% on PETA, 14.4% on Market-1501 Attribute, and 2.9% on PA100K, respectively, which clearly validates its effectiveness. Further, the table shows that the simple baseline is already comparable to the state of the art; we believe that it is a solid and unexplored baseline that future work has to consider. Finally, we again emphasize that state-of-the-art methods [2, 9] also take advantage of attribute classification, thus the comparison in Table 2 is equitable.

Method	PETA	Market-1501	PA100K
w/o $\delta(\mathbf{p}_i, \mathbf{p}_j)$	52.0	46.1	30.3
Uniform w_k	52.5	46.5	29.8
ℓ_2 normalized w_k	52.0	46.3	30.1
Ours	**56.5**	**49.6**	**31.9**

Table 5. Comparison of ASMR and its variants in Rank@1 of the search resutls on the three datasets.

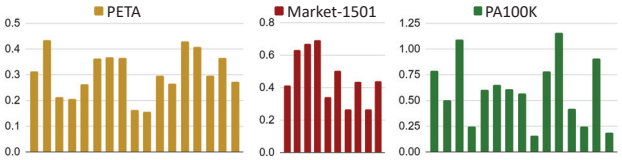


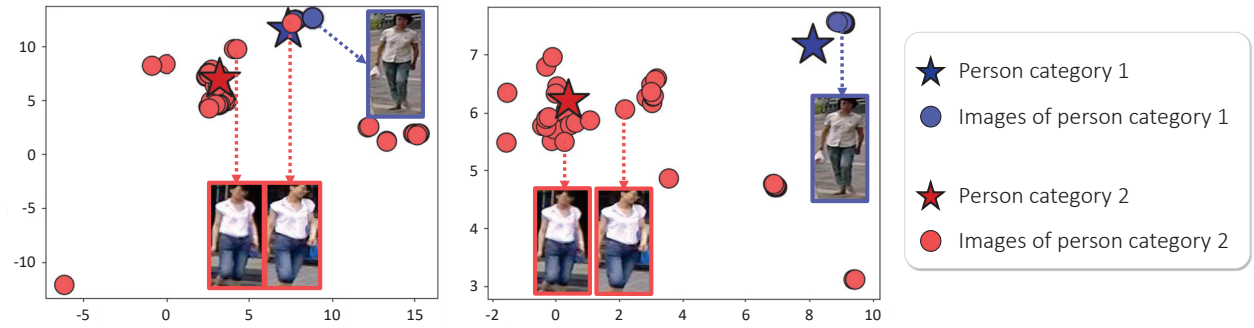
Figure 5. Visualization of w_k learned in our method on the three datasets, where each value corresponds each attribute.

Comparison to other embedding losses. To demonstrate superiority of our modality alignment loss $\mathcal{L}_{\text{MA}}$, we evaluate variants of our method that replace $\mathcal{L}_{\text{MA}}$ with Proxy Anchor [19], Proxy NCA [31], CosFace [43] and SphereFace [28], representative embedding losses using prototypes. Table 4(b) shows that our method using $\mathcal{L}_{\text{MA}}$ largely outperformed the two variants, which indicates the advantage of $\mathcal{L}_{\text{MA}}$.

4.6. In-depth Analysis on ASMR

The effect of each design points of ASMR are verified by experiments, whose results are summarized in Table 5. First of all, the large gap between ours and its variant without δ demonstrates the significant contribution of δ to the performance. Note that ASMR without δ becomes analogous to the diversity regularizer in [11], and forces person category embeddings to be uniformly distributed. This suggests that ASMR does not blindly enlarge between-category margins but controls them with consideration to semantic affinities between categories, which is vital for the outstanding performance of our work.

The role of learnable weights w_k in δ is also investigated. We observed that the performance drops when the weights are fixed by a single value (*i.e.*, uniform w_k), which suggests that the learned weights well capture the unequal importance of attributes. We also found that imposing ℓ_2 normalization to the weights does not useful, rather damages performance; Fig. 5 shows that our method learns non-uniform and positive weights with no such a constraint.



(a) w/o ASMR (b) Ours

Person category 1	Teenager	handbag	pants	long lower body clothing	short sleeve	short hair	female	up white	down blue
Person category 2	Teenager	bag	pants	short	short sleeve	long hair	female	up white	down blue

Figure 6. t-SNE visualization of a part of the joint embedding space learned for the Market-1501 Attribute dataset. Stars and circles indicate embedding vectors of person categories and their associated images, respectively, and their colors mean their person categories. The person categories are elaborated below, where attributes that are different between the two categories are colored in red.

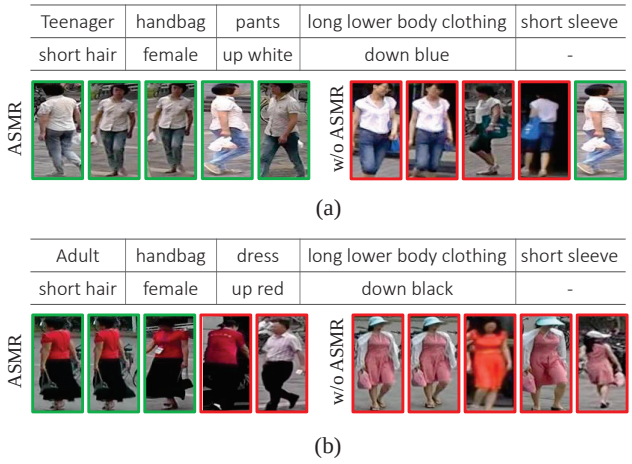


Figure 7. Top 5 retrieval results of our method and its variant without ASMR on the Market-1501 Attribute dataset. Images are sorted from left to right according to their ranks. Green and red boxes indicate true and false matches, respectively.

Finally, we present more detailed qualitative analysis on the effect of ASMR to explain how it works and to validate its contribution. Fig. 6 compares joint embedding spaces learned by our model and its reduced version without ASMR. We adopt t-SNE [39] to visualize their embedding spaces, and focus only on two particular person categories sharing many of their attributes for a clear analysis. As shown in Fig. 6(a), some images of person category 2, whose appearances are quite similar to those of person category 1, are located overly close to person category 1 in the embedding space learned without the regularizer; such images will lead to failures in person search. This happens since the model is biased towards the image modality immoderately if no constraint is imposed for person category embeddings. In contrast, Fig. 6(b) shows that our final model with the regularizer enlarges the margin between the two categories according to their semantic dissimilarity so that they are well discriminated in the embedding space.

The effectiveness of ASMR is further validated by comparing retrieval results of the models with and without the regularizer in Fig. 7. The results suggest that the model without the regularizer often fails when images of different person categories are overly similar as in Fig. 7(a) and/or some attributes of query are about fine details of images like hat and age in Fig. 7(b). Our method with ASMR handles these issues effectively thanks to the improved discriminability by ASMR.

5. Conclusion

We have presented an efficient and effective framework for attribute-based person search. The main contribution of our work is a novel loss function based on ASMR for learning cross-modal embeddings: It aligns a person category and associated images in a common embedding space, and at the same time, arranges person categories according to their semantic affinities in the space. We demonstrated by experiments that the proposed loss allows a simple embedding model to achieve state-of-the-art performance. Considering its brevity and outstanding performance, our work will be a solid baseline for attribute-based person search.

Acknowledgement: This work was supported by the NRF grant, the IITP grant, and R&D program for Advanced Integrated-intelligence for IDentification, funded by Ministry of Science and ICT, Korea (No.2019-0-01906 Artificial Intelligence Graduate School Program–POSTECH, NRF-2021R1A2C3012728–30%, NRF-2018R1A5A1060031–20%, NRF-2018M3E3A1057306–30%, IITP-2020-0-00842–20%).

References

[1] Jane Bromley, Isabelle Guyon, Yann Lecun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proc. Neural Information Processing Systems (NeurIPS)*, 1994.

[2] Yu-Tong Cao, Jingya Wang, and Dacheng Tao. Symbiotic adversarial learning for attribute-based person search. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[3] 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 *Proc. European Conference on Computer Vision (ECCV)*, 2018.

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

[5] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[6] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proc. ACM Multimedia Conference (ACMMM)*, 2014.

[9] Qi Dong, Shaogang Gong, and Xiatian Zhu. Person search by text attribute query as zero-shot learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[10] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] M. Hayat *et al.* Gaussian affinity for max-margin class imbalanced learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[12] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proc. British Machine Vision Conference (BMVC)*, 2017.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[14] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[15] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 2004.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[21] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015.

[22] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

[23] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[25] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019.

[26] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia*, 19(6):1234–1244, 2016.

[27] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 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] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

[31] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Le-

ung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

[32] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[33] Jicheol Park, Boseung Jeong, Jongju Shin, Juyoung Lee, and Suha Kwak. Learning discriminative part features through attentions for effective and scalable person search. In *IEEE International Conference on Image Processing (ICIP)*, 2020.

[34] Walter J Scheirer, Neeraj Kumar, Peter N Belhumeur, and Terrance E Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[36] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2016.

[37] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[38] 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 *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[39] L.J.P van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9: 2579–2605, 2008.

[40] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *Proc. Winter Conference on Applications of Computer Vision (WACV)*, 2009.

[41] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proc. ACM Multimedia Conference (ACMMM)*, 2017.

[42] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[43] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[44] Jiang Wang, Yang Song, T. Leung, C. Rosenberg, Jingbin Wang, J. Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[45] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[46] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proc. International Conference on Machine Learning (ICML)*, 2015.

[47] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

[49] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

[51] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[52] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.