Context-Aware Unsupervised Clustering for Person Search

Byeong-Ju Han*,1
bjhan@unist.ac.kr
Kuhyeun Ko*,1
khko@unist.ac.kr
Jae-Young Sim†,1,2
jysim@unist.ac.kr

[1] Department of Electrical Engineering, UNIST, Ulsan, Korea
[2] Graduate School of Artificial Intelligence, UNIST, Ulsan, Korea
*Equal contribution
†Corresponding author

Abstract

The existing person search methods use the annotated labels of person identities to train deep networks in a supervised manner that requires a huge amount of time and effort for human labeling. In this paper, we first introduce a novel framework of person search that is able to train the network in the absence of the person identity labels, and propose efficient unsupervised clustering methods to substitute the supervision process using annotated person identity labels. Specifically, we propose a hard negative mining scheme based on the uniqueness property that only a single person has the same identity to a given query person in each image. We also propose a hard positive mining scheme by using the contextual information of co-appearance that neighboring persons in one image tend to appear simultaneously in other images. The experimental results show that the proposed method achieves comparable performance to that of the state-of-the-art supervised person search methods, and furthermore outperforms the extended unsupervised person re-identification methods on the benchmark person search datasets. Code is available at https://github.com/VIP-Lab-UNIST/CUCPS_official.

1 Introduction

Person search has drawn much attention in diverse applications from private entertainment to public safety. However, it is a challenging task to detect person objects from gallery images and to recognize the identities of detected persons together. The existing methods for person search mainly try to train deep networks to embed a distinct feature for each person identity in a supervised manner by using manually annotated identity labels. Human labeling requires much time and effort and often causes incomplete annotation that usually degrades the performance of person search. A few attempts have been made to assign existing valid labels to such unlabeled persons [7, 16], however, they still follow the supervised learning framework and their performance mainly depends on the labeled data.

In this work, we define a novel problem of person search in the absence of labeled person identities and propose an end-to-end network that is trained by context-aware clustering methods to group the persons considered to have the same identity. Specifically, we exploit two specific context properties of *uniqueness* that no more than a single person has the

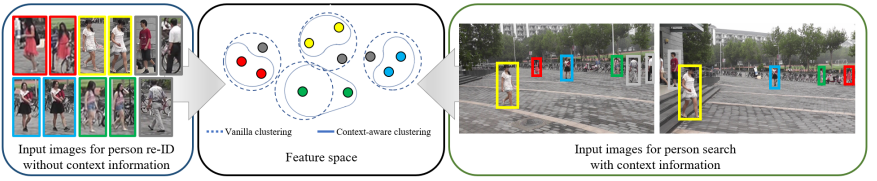


Figure 1: Comparison between the proposed clustering using contextual priors and the vanilla clustering only depending on feature similarity.

same identity to a given query person in each gallery image and *co-appearance* that multiple persons appeared in an image are highly probable to appear simultaneously in other images. Figure 1 conceptually shows the strength of the proposed clustering method using the context properties compared to a vanilla clustering method grouping the samples within a certain feature distance into the same identity. While the vanilla clustering method incorrectly predicts the identities of the persons in the gray boxes due to their similar appearances to others, the proposed clustering method removes them by the uniqueness property. Moreover, whereas the vanilla clustering method fails to group the persons in the green boxes together due to a large feature distance between two green points in the feature space, the proposed method finds such hard positive samples according to the co-appearance property.

The main contributions of this paper are summarized as follows.

1. We first introduced the problem of person search without identity labels, and proposed context-aware unsupervised clustering methods to solve this problem by using the uniqueness and co-appearance properties of the person search framework.
2. We extended the existing unsupervised person re-ID methods and showed that the proposed method outperforms them and also achieves comparable performance to the state-of-the-art supervised person search methods.

2 Related Works

This section summarizes the existing works related to the problem of training a person search method in the absence of the person identity labels. We firstly introduce the unsupervised person re-identification methods embedding distinct person features from cropped person images without the person identity labels, and then explain the supervised person search methods especially using context information of scene images.

Person re-ID for unlabeled data. The person re-ID methods take cropped images of detected person proposals as input and classify them into clusters according to the same person identities. To train person re-ID networks in the absence of identity labels, many unsupervised methods adopted the domain adaptation scheme that aims to make a model trained in a source dataset with identity labels perform well on a target dataset without identity labels [4, 6, 7, 14, 18, 19, 26]. Recently, unsupervised person re-ID methods have been developed without using any prior knowledge associated with other datasets. Lin et al. [12] proposed a bottom-up clustering method that first assigns different clusters to all samples and gradually merges the closest neighboring clusters together into one cluster. However, the merging criterion based on the shortest feature distance between two clusters is sensitive to outliers on cluster boundaries. Ding et al. [5] alleviated this problem by considering all the samples in each cluster to measure the feature distance between two clusters and Zeng et al. [24] maximized the shortest feature distance between positive samples and selected hard

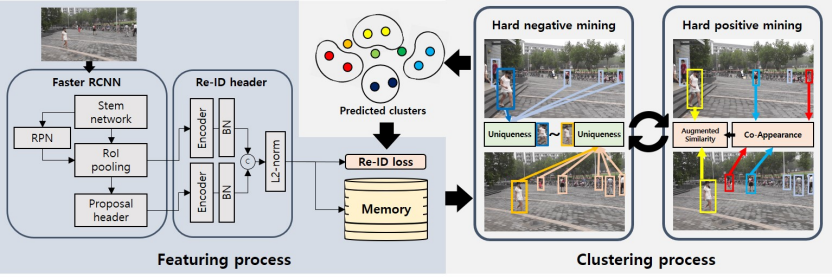


Figure 2: Overview of the proposed method.

negative samples. Wang et al. [17] used a binary vector based multi-label representation to indicate multiple samples with the same person identity. Lin et al. [13] softened the binary values of the multi-label representation according to the relation between samples to avoid hard-decision errors. While the recent unsupervised re-ID methods mainly adopt advanced distance metrics or soften labeling to reduce errors caused by hidden outliers, the proposed method analyzes the scene context to detect the outliers and create reliable clusters.

Person search using image context Person search addresses the problem to find persons from a gallery of multi-view scene images, who have the same identity to a query person. All existing methods follow a fully supervised learning framework to detect person proposals in the scene images and classify them into different clusters according to the guidance of annotated person identity labels. Unlike person re-ID, person search can investigate the context information such as uniqueness and co-appearance in the scene images. Dai et al. [7] utilized the uniqueness property to temporally assign distinct pseudo labels to each of the unlabeled persons caused by incomplete annotation. Li et al. [11] considered all possible matching pairs between a query person in a query set and a candidate person in gallery images, and found an optimal set of matching pairs by a bipartite matching algorithm, to preserve the uniqueness property. Yao et al. [23] especially proposed a new definition of similarity between persons to enforce the persons in an image not to match the same target person in gallery images. Yan et al. [21] constructed a graph whose nodes and edges are defined by detected persons in a pair of images and similarities of the initially matched person pairs, respectively, to update the matching scores according to the co-appearance property. Compared to the existing methods that adopt the image context to support the supervised training process or enhance the matching quality on the test, we first apply it on the unsupervised framework and specifically investigate that the properties of uniqueness and co-appearance complementally contribute to training a person search method in the absence of person identity labels.

3 Proposed Method

The conventional supervised person search methods train the featuring process composed of Faster R-CNN [15] and a re-ID header using person identity labels. In this work, we introduce a novel problem of person search where the bounding box labels are available for person detection but the person identity labels are not given for person re-ID. Different from the unsupervised person re-ID problem, we can exploit the contextual information of the persons located within a gallery image to address this problem. Figure 2 visualizes the overall procedures of the proposed method that contains key clustering modules of hard negative mining (HNM) and hard positive mining (HPM).

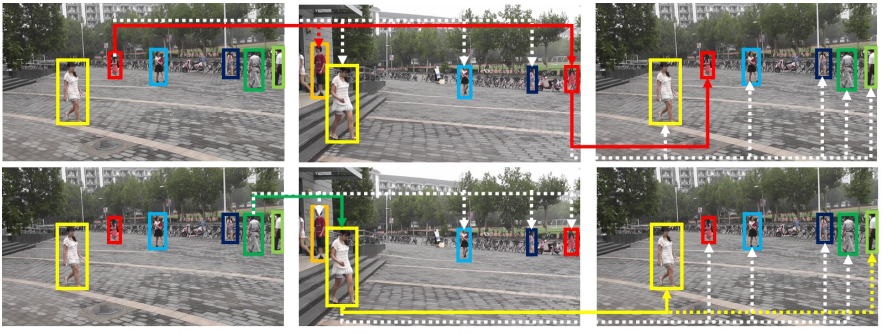


Figure 3: Behaviors of the uniqueness based HNM. The white and colored lines represent low and high feature similarities between two persons, respectively. The solid lines represent the pairs of the highest feature similarity.

3.1 Uniqueness Based Hard Negative Mining

The existing unsupervised person re-ID methods [12, 13] that construct clusters of the same size regardless of person identities may produce limited performances on real datasets because each person identity has diverse numbers of samples as in the person search dataset PRW. We propose a more reliable clustering method for unsupervised person re-ID, called uniqueness based hard negative mining (HNM), that effectively reduces hard negative samples using the uniqueness property of person search.

Let $\mathcal{G} = \{I_1, I_2, \cdots, I_N\}$ be a gallery of images where N is the number of total images, and $\mathcal{X}_l$ denote a set of persons x_j^l's detected in I_l. Assume that a query is selected as the i-th person x_i^k from the k-th image I_k, and we find a positive sample of the query from the l-th image I_l with $l \neq k$. We first take the candidate persons in I_l, whose feature similarities to the query person are higher than a certain threshold δ, and initialize the set of positive samples as

$$\hat{\mathcal{C}}_l(x_i^k) = \{x_j^l \mid s(\mathbf{f}_i^k, \mathbf{f}_j^l) > \delta, \;\; x_j^l \in \mathcal{X}_l\}, \tag{1}$$

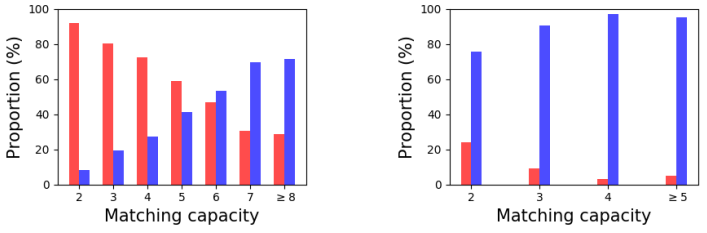
where $\mathbf{f}_j^l$ is the feature vector of x_j^l stored in the feature memory and $s(\mathbf{f}, \mathbf{g})$ is the similarity between two features of $\mathbf{f}$ and $\mathbf{g}$. We may collect $\hat{\mathcal{C}}_l(x_i^k)$'s over all the gallery images and take the union $\hat{\mathcal{C}}(x_i^k) = \bigcup_l \hat{\mathcal{C}}_l(x_i^k)$ as a set of positive samples for the query x_i^k. However, this simple thresholding scheme often includes negative samples to the clusters when relatively high similarity values are computed between two persons of different identities.

In order to reject such hard negative samples from the clusters, we exploit the uniqueness property of person search that there are no more than a single person in each image having the same identity to a given query person. In practice, we implement this constraint by winner-take-all (WTA) scheme such that only the corresponding sample

$$x_{j^*}^l = \underset{x_j^l \in \hat{\mathcal{C}}_l(x_i^k)}{\operatorname{argmax}} \, s(\mathbf{f}_i^k, \mathbf{f}_j^l), \tag{2}$$

which has the highest similarity to the query x_i^k, is remained in $\hat{\mathcal{C}}_l(x_i^k)$ and the other samples are removed from $\hat{\mathcal{C}}_l(x_i^k)$. We then have an updated candidate set as $\mathcal{C}_l(x_i^k) = \{x_{j^*}^l\}$.

To further improve the accuracy of clustering, we additionally apply the backward matching of WTA scheme from I_l to I_k to check whether the obtained candidate $x_{j^*}^l$ satisfies the



(a) PRW (b) CUHK-SYSU

Figure 4: Proportions of the image pairs containing a single true positive pair of persons (red) and multiple true positive pairs of persons (blue) according to the matching capacity. (a) PRW and (b) CUHK-SYSU datasets.

cycle consistency to x_i^k or not. If the original query x_i^k becomes the element of $\mathcal{C}_k(x_{j^*}^l)$, then we decide that the pair of x_i^k and $x_{j^*}^l$ has the same identity to each other and include them into the same cluster. Otherwise, we eliminate $x_{j^*}^l$ from $\mathcal{C}_l(x_i^k)$ due to the violation of the cyclic consistency condition. We decide that there exists no corresponding person to x_i^k in I^l when $\mathcal{C}_l(x_i^k)$ is empty. Finally, we have a positive set of the query x_i^k as $\mathcal{C}(x_i^k) = \bigcup_l \mathcal{C}_l(x_i^k)$.

Figure 3 shows the behavior of the proposed uniqueness based HNM where the persons are localized by the bounding boxes in different colors according to their identities. As shown in the first row of Figure 3, the person in the red box in I_1 is selected as a query. The two persons in the orange and red boxes are initially detected from I_2 as candidate positive samples since they wear red clothes and exhibit higher feature similarities to the query than a threshold. But we select the correct person in the red box in I_2 by WTA which has the highest similarity to the query. The matched pair is finally considered as a positive pair since the original query person in I_1 is selected to have the highest similarity to the person in the red box in I_2 via the backward matching. On the other hand, as shown in the second row of Figure 3, when selecting the person in the green box in I_1 as a query, the person in the yellow box is detected from I_2 by (2) despite not being a true positive sample since the query person does not appear in I_2. The person in the yellow box is detected from I_1 via the backward matching instead of the original query person in the green box, and therefore we remove this hard negative sample from the clusters. Note that the uniqueness based HNM cannot be applied to the unsupervised person re-ID framework due to the lack of contextual information of the associated gallery images, but can be used in the person search framework.

3.2 Co-Appearance Based Hard Positive Mining

The uniqueness based HNM for unsupervised clustering still suffers from missing some true positive samples. We also propose a hard positive mining (HPM) scheme that finds challenging positive samples having low feature similarities to the query. We utilize the contextual information of the neighboring persons to the query appeared in the same image based on the co-appearance property that multiple persons in an image are likely to appear simultaneously in other images. We investigate this property on PRW and CUHK-SYSU datasets in Figure 4. We collect all the image pairs containing at least one true positive pair of persons, and classify them into different groups according to the matching capacity that is the minimum number of persons in an image between two images of each pair. Then, for each group, we compare the proportions of the image pairs containing only a single

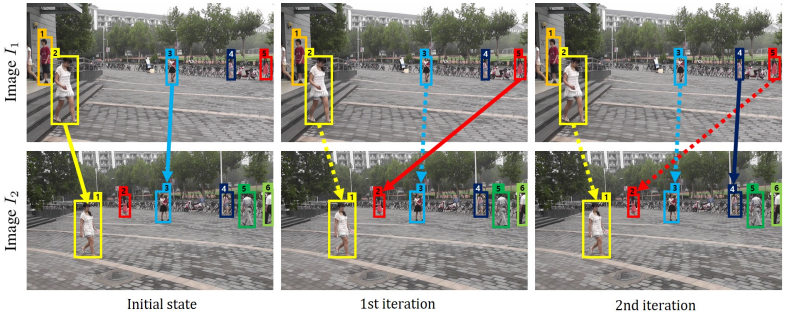


Figure 5: Behaviors of the co-appearance based HPM. The dotted arrows connect the positive pairs detected at the previous states and the solid arrows represent the positive pairs newly detected at the current iteration.

true positive pair of persons where the co-appearance property is not satisfied (red bar) and multiple true positive pairs of persons where the co-appearance property is satisfied (blue bar), respectively. The statistical results in Figure 4 show that the number of multiple true positive pairs is increased as more persons appear in an image, and especially CUHK-SYSU dataset exhibits strong co-appearance property.

We compute a co-appearance $A^{(t)}(k,l)$ between I_k and I_l at the t-th iteration by considering the clustering results of the neighboring persons given by

$$A^{(t)}(k,l) = \sum_{x_i^k \in \mathcal{X}_k,\, x_j^l \in \mathcal{C}_l^{(t)}(x_i^k)} s(\mathbf{f}_i^k, \mathbf{f}_j^l), \tag{3}$$

where $\mathcal{C}_l^{(t)}(x)$ is the cluster of x at the t-th iteration on the l-th image. Note that $A^{(t)}(k,l)$ is obtained by the latest clusters of $\mathcal{C}_l^{(t)}(x)$ and increased as more persons in I_k are matched to the persons in I_l with larger similarities. Then, for all the pairs of $x_i^k \in \mathcal{X}_k$ and $x_j^l \in \mathcal{X}_l$, we update the original feature similarity of $s(\mathbf{f}_i^k, \mathbf{f}_j^l)$ to $s^{(t+1)}(\mathbf{f}_i^k, \mathbf{f}_j^l)$ at the next iteration as

$$s^{(t+1)}(\mathbf{f}_i^k, \mathbf{f}_j^l) = s(\mathbf{f}_i^k, \mathbf{f}_j^l) + \beta A^{(t)}(k,l), \tag{4}$$

where β denotes a weight to adjust the contribution of the co-appearance which is empirically set to 0.1. We also apply the uniqueness based HNM between I_k and I_l again using the updated feature similarities to refine the clustering results. Note that non-empty clusters are maintained since the feature similarities are increased by the same amount. However, we can check if empty clusters include new elements by detecting hard positive samples using the increased feature similarities. In this paper, HPM iterates three times maximally.

Figure 5 shows two images I_1 and I_2 where the co-appearance based HPM is applied iteratively. The person labels are annotated on the top of the bounding boxes. Let us assume that, at the initial state, the two true positive pairs of (x_2^1, x_1^2) and (x_3^1, x_3^2) are correctly clustered, respectively, whereas the other two true positive pairs of (x_5^1, x_2^2) and (x_4^1, x_4^2) are missed to be correctly clustered due to relatively low feature similarities. At the first iteration of HPM, (x_4^1, x_4^2) is still missed to be clustered as a positive pair, but (x_5^1, x_2^2) is newly detected as a positive pair since the feature similarity $s^{(1)}(\mathbf{f}_5^1, \mathbf{f}_2^2)$ between x_5^1 and x_2^2 becomes higher than the threshold δ by the co-appearance $A^{(0)}(1,2)$ associated with the persons of x_2^1 and x_3^1 neighboring to x_5^1. At the second iteration, the additional positive pair (x_5^1, x_2^2) further increases the co-appearance to $A^{(1)}(1,2)$, and therefore increases the feature similarity $s^{(2)}(\mathbf{f}_4^1, \mathbf{f}_4^2)$ accordingly such that the new pair of (x_4^1, x_4^2) is detected as a positive pair.

3.3 Loss Function

The global feature memory contains the normalized features for all the person instances in gallery images. Each feature vector $\mathbf{f}_i^k \in \mathbb{R}^{256}$ for $x_i^k \in \mathcal{X}_k$ is initialized as a zero vector and then updated after back-propagating gradients through the global feature memory by

$$\mathbf{f}_i^k \leftarrow \frac{1}{Z}(\mathbf{f}_i^k + \mathbf{g}_i^k), \tag{5}$$

where $\mathbf{g}_i^k \in \mathbb{R}^{256}$ denotes the feature vector extracted from a detected bounding box b_i^k spatially overlapping with the ground truth bounding box of x_i^k in I_k, and Z is the normalization factor such that $\|\mathbf{f}_i^k\|_2 = 1$.

The total loss function is composed of a detection loss and a re-ID loss, where we adopt a detection loss used in Faster R-CNN [15]. We propose a re-ID loss for a detected bounding box b_i^k given by

$$\mathcal{L}(b_i^k) = -\frac{1}{|\mathcal{C}(x_i^k)|} \sum_{x_j^l \in \mathcal{C}(x_i^k)} \log\left(p(x_j^l|b_i^k)\right), \tag{6}$$

that encourages b_i^k to have the same identity to $x_j^l \in \mathcal{C}(x_i^k)$ by maximizing the probability

$$p(x_j^l|b_i^k) = \frac{\exp\left(s(\mathbf{f}_j^l, \mathbf{g}_i^k)/\tau\right)}{\sum_{x \in \{\mathcal{C}(x_i^k) \cup \mathcal{D}(x_i^k)\}} \exp\left(s(\mathbf{f}, \mathbf{g}_i^k)/\tau\right)}, \tag{7}$$

where $\mathcal{D}(x_i^k)$ is the set of the negative samples in $\mathcal{C}^c(x_i^k)$ having top 1% similarities to x_i^k, and τ is a temperature coefficient empirically set to 0.1. Note that the denominator in (7) does not consider the negative samples having relatively low similarities to x_i^k, but mainly includes challenging negative samples having high similarities to x_i^k. Therefore, the proposed re-ID loss encourages increasing the probability associated with the positive samples while effectively decrease the probability for the challenging negative samples. Moreover, the probability $p(x_j^l|b_i^k)$ is equally maximized by $\frac{1}{|\mathcal{C}(x_i^k)|}$ for all $x_j^l \in \mathcal{C}(x_i^k)$ clustered to be positive samples of x_i^k, and eventually achieves good performance since HNM and HPM cooperate together to include most of the positive samples to $\mathcal{C}(x_i^k)$ successfully while excluding even challenging negative samples from $\mathcal{C}(x_i^k)$ reliably.

4 Experimental Results

4.1 Experimental Setup

Benchmark datasets. CUHK-SYSU [20] and PRW [25] have been widely used as benchmark datasets for person search. CUHK-SYSU dataset is composed of 11,206 images for training and 6,978 images for testing and provides 96,143 bounding boxes with 8,432 identities to indicate individual persons. It also provides 2900 test queries and predetermined galleries of test images. PRW dataset consists of a training set of 5,704 images and a test set of 6,112 images captured at 6 different fixed camera positions. It provides 43,110 bounding boxes of persons with 932 identities including 2,057 test queries. However, since PRW does not report a detailed description to define galleries to search persons matching to test queries, we define two different types of the gallery: *regular gallery* and *multi-view gallery* in this paper. The regular gallery consists of all the test images except the image where

(a) (b) (c)

Figure 6: Clustering results. (a) Query persons. The persons clustered by using (b) HPM and (c) HPM+HNM.

	w/o HNM	with HNM	HNM gain
w/o HPM	27.68/59.35	28.01/60.28	+0.33/+0.93
with HPM	32.87/62.86	36.61/64.85	+3.74/+1.99
HPM gain	+5.19/+3.51	+8.60/+4.57	+8.93/+5.50

Table 1: Ablation study of the effect of HNM and HPM. Each cell shows the performance in terms of mAP and Top-1 score, respectively.

the query is detected, and ignores the unlabeled persons. This is expected to be a similar setup to the gallery used in many existing methods. The multi-view gallery is quite challenging because it excludes the redundant test images captured from the same camera view to that of the query image and includes all the unlabeled persons. Note that although both datasets provide the identity labels, we do not use them for training the network to assume the unsupervised framework with the absence of identity labels.

Implementation details. We use the sequence of blocks from 'conv1' to 'conv4' of ResNet-50 [9] pre-trained on ImageNet [3] for image classification as a stem network, and use 'conv5' block of ResNet-50 as a proposal header, respectively, in Figure 2. Each encoder of the re-ID header is defined as a fully connected layer to squeeze the high dimensional feature brought from the backbone network to a feature vector of size 128. We train the proposed network using the same hyper-parameters for CUHK-SYSU and PRW datasets except for the number of training epochs and the learning rate decay. The training strategy is empirically determined based on the ablation study and related primary research. More details are described in the supplementary material.

Evaluation metrics. The performance of person search is usually evaluated by the two metrics: mean of Average Precision (mAP) and Top-k score. mAP considers both the precision and recall of the predicted results by computing the average area under the precision-recall curves. Top-k score checks whether the predicted top-k candidates best matching to a given query include at least one true positive sample or not.

4.2 Ablation Study

We conduct ablation studies about the performance of the proposed method using the multi-view gallery in PRW dataset.

Co-appearance based HPM. The proposed HPM is designed to detect challenging positive samples whose feature similarities to a given query are relatively low due to the variation of human poses and/or the changes of camera viewing directions. Each row in Figure 6 (b) shows the matched persons to a query person in Figure 6 (a) clustered by HPM. We see that HPM adds more persons in blue into the initial cluster of the first three persons by exploiting the co-appearance property. Also, Table 1 shows that the quantitative performance gains of using HPM over the naive method, which generates clusters using a constant threshold for feature similarity, are 5.19 and 3.51 in terms of mAP and Top-1 score, respectively.

Uniqueness based HNM. HNM reduces the false positive errors for clustering by detecting challenging negative samples. Figure 6 (c) shows the clustering results of using HPM and HNM together, where we see that HNM removes the unreliable persons in red from the augmented clusters shown in Figure 6 (b). As shown in Table 1, HNM also improves the performance by 3.74 and 1.99 in terms of mAP and Top-1 score, respectively, when used with HPM together, which means HNM can effectively reduce false positive errors caused

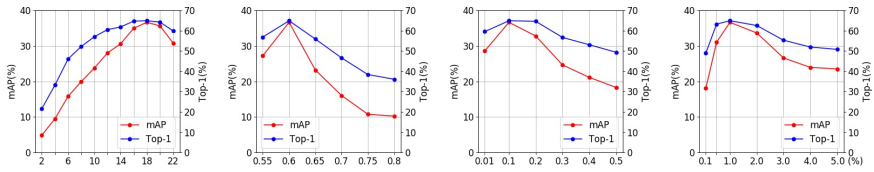


Figure 7: Effects of the hyper-parameters. The performance of the proposed method according to varying (a) the number of training iterations, (b) the threshold for feature similarity, (c) the weight β for co-appearance based update of feature similarity, and (d) the ratio of hard negative samples for training loss computation.

by HPM. Consequently, we see that each of the proposed HNM and HPM clearly demonstrates the effectiveness and achieves a remarkable synergy thanks to their complementary behaviors.

Number of training iterations. Figure 7 (a) shows the performance scores of the proposed method evaluated up to the 22nd epoch of training. We see that the performance gradually increases according to the training iteration and reaches the top score at the 18th epoch, but decreases with more epochs due to the over-fitting.

Threshold for feature similarity. The HNM module first collects candidate positive samples having higher feature similarities to a given query than a certain threshold. Finding an optimal threshold is important since inappropriately determined boundaries of clusters especially in the early training phase may cause incorrect training and eventually result in worse performance. Figure 7 (b) shows the performance scores varying this threshold from 0.55 to 0.8, where we see that the best performance is achieved with the threshold value of 0.6.

Weight β. Figure 7 (c) compares the performance of the proposed method according to different weights of β in (4) used for the iterative update of the feature similarity. We have the best performance using $\beta = 0.1$. Note that HPM with $\beta > 0.1$ clusters lots of false positive samples incorrectly that are hard to be detected by HNM. In contrast, HPM with $\beta < 0.1$ does not sufficiently exploit the co-appearance property to detect hard positive samples.

Ratio of hard negative samples. Figure 7 (d) visualizes the impact of the ratio of hard negative samples to construct $\mathcal{D}(x_i^k)$ in (7) for training loss computation. The best performance is achieved when we set the ratio to 1%. If we select more negative samples, relatively easy negative samples may have more contribution to train the network yielding decreased performance. In contrast, if we take fewer negative samples, the network may not be properly trained to encourage large feature distances between the query and hard negative samples due to the lack of diversity of negative samples.

4.3 Comparison with Existing Methods

Since there are no existing person search methods to handle the absence of person identity labels, the proposed method is compared with the existing supervised person search methods as well as the extensions of the existing unsupervised person re-ID methods that are trainable without person identity labels and any prior knowledge, respectively. For extension, an individually trained person detection network of Faster R-CNN [15] is followed by each re-ID network. Table 2 quantitatively compares the proposed method with the six state-of-the-art person search methods [1, 8, 10, 11, 22, 23] in CUHK-SYSU and the regular gallery of PRW. And Table 3 shows a more detailed performance comparison between the proposed

Method	Supervised	CUHK-SYSU		PRW	
		mAP	Top-1	mAP	Top-1
MGTS [9]	Yes	83.3	84.2	32.8	72.1
DMRNet [8]	Yes	93.2	94.2	46.9	83.3
SeqNet [13]	Yes	94.8	95.7	47.6	87.6
AlignPS [23]	Yes	94.0	94.5	46.1	82.1
PGA [19]	Yes	92.3	94.7	44.2	85.2
OR [25]	Yes	93.2	93.8	52.3	71.5
Proposed	No	**81.1**	**83.2**	**41.7**	**86.0**

Table 2: Quantitative comparison with the existing supervised person search methods on CUHK-SYSU and regular gallery of PRW.

Method	CUHK-SYSU		PRW		
	mAP	Top-1	mAP	Top-1	Gallery
DET+BUC [12]	74.8	77.4	26.0	83.6	Regular
DET+MLC [17]	69.2	73.7	25.4	84.7	
Proposed	**81.1**	**83.2**	**41.7**	**86.0**	
DET+BUC [12]	-	-	18.6	53.0	Multi -view
DET+MLC [17]	-	-	17.1	50.8	
Proposed	-	-	**36.6**	**64.9**	

Table 3: Quantitative comparison with the extended unsupervised person re-ID methods on CUHK-SYSU and regular and multi-view galleries of PRW.

Figure 8: Challenging cases for the proposed unsupervised clustering method. The leftmost figure shows a query person, and the searched persons are highlighted in green and red corresponding to the true and false matching, respectively.

method and the two extended unsupervised state-of-the-art re-ID methods [12, 17], evaluated in CUHK-SYSU and both galleries of PRW.

Though trained without the labels of person identity, the quantitative performance of the proposed method achieves more than 85% of the top scores of the state-of-the-art supervised methods on both datasets in terms of mAP and Top-1, respectively, as shown in Table 2. Furthermore, the proposed method yields significantly better performance than that of the extended unsupervised person re-ID methods in both datasets. Particularly, in Table 3 where the lower block shows the performances evaluated in the challenging multi-view gallery of PRW, we see that the proposed method providing the best mAP score of 36.6 almost improves the performance twice compared to the second-best method with mAP score of 18.6 on the multi-view gallery. Note that the proposed method and the extensions of BUC and MLC generate person clusters mainly employ the appearances of the detected persons in unsupervised manners, and therefore they often provide non-negligible similarities between different persons exhibiting similar looks, especially in PRW dataset that contains many different persons wearing similar cloths as shown in Figure 8. In such cases we have high Top-1 scores but relatively low mAP scores. Due to the space limit, the qualitative results are reported in the supplementary material.

5 Conclusion

We addressed a novel person search problem without using the labels of person identities in this work. We investigated contextual properties of the person search framework, and proposed two unsupervised clustering methods of the uniqueness based HNM and the co-appearance based HPM to classify unlabeled person samples. We conducted comparative experiments including ablation studies of the proposed method. Experimental results demonstrated that the proposed method provides comparable performance to the existing state-of-the-art supervised person search methods and outperforms the extended unsupervised person re-ID methods.

Acknowledgments

This work was supported by NRF of Korea within the Ministry of Science and ICT (MSIT) under Grant 2020R1A2B5B01002725, and by Institute of Information & communications Technology Planning & Evaluation (IITP) through MSIT under Grant 20200013360011001 (Artificial Intelligence graduate school support (UNIST)) and 20170006670021001.

References

[1] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search by separated modeling and a mask-guided two-stream cnn model. *IEEE Transactions on Image Processing*, 2020.

[2] Ju Dai, Pingping Zhang, Huchuan Lu, and Hongyu Wang. Dynamic imposter based online instance matching for person search. *Pattern Recognition*, 100:107120, 2020.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[4] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] Guodong Ding, Salman H Khan, Zhenmin Tang, J Zhang, and F Porikli. Dispersion based clustering for unsupervised person re-identification. In *Proceedings of the British Machine Vision Conference*, 2019.

[6] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14:1–18, 2018.

[7] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE Conference on International Conference on Computer Vision*, 2019.

[8] Chuchu Han, Zhedong Zheng, Changxin Gao, Nong Sang, and Yi Yang. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] Hanjae Kim, Sunghun Joung, Ig-Jae Kim, and Kwanghoon Sohn. Prototype-guided saliency feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[11] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[12] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[13] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[14] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *Proceedings of the IEEE Conference on International Conference on Computer Vision*, 2019.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[16] Wei Shi, Hong Liu, Fanyang Meng, and Weipeng Huang. Instance enhancing loss: Deep identity-sensitive feature embedding for person search. In *Proceedings of the IEEE Conference on International Conference on Image Processing*, 2018.

[17] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[18] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[19] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[20] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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

[22] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[23] Hantao Yao and Changsheng Xu. Joint person objectness and repulsion for person search. *IEEE Transactions on Image Processing*, 2021.

[24] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

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

[26] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.