Heterogeneous Graph Driven Unsupervised Domain Adaptation of Person Re-identification

Shaochuan Lin[a], Jianming Lv[a,*], Zhenguo Yang[b], Qing Li[c], Wei-Shi Zheng[d]

[a] *South China University of Technology,Guangdong,China*
[b] *Guangdong University of Technology,Guangdong,China*
[c] *The Hong Kong Polytechnic University,Hong Kong,China*
[d] *Sun Yat-sen University,Guangdong,China*

Abstract

How to incrementally optimize a pre-trained classifier in an unlabeled target domain is a core challenging problem of domain adaptation (DA) for many visual tasks, such as Person Re-identification (re-ID). Most of the existing methods optimize the model based on pseudo labels or similarity of instance pairs, but ignoring the diverse manifold structures of unlabeled instances in the whole dataset. In this paper, we address the importance of such structural information in domain adaptation, and propose a Heterogeneous Graph driven Optimization scheme, namely H-GO, for structure based unsupervised learning. In particular, H-GO builds a heterogeneous graph of unlabeled images to consider the heterogeneous properties of images from various cameras with varied visual styles. A heterogeneous affinity propagation method is further applied to explore the graph based affinity between the instances which share similar manifold structures. Finally, a heterogeneous affinity learning procedure is taken to optimize the visual models by using the graph based affinity of instances. Comprehensive experiments are conducted on three large-scale re-ID datasets, and the results demonstrate the flexibility and the superior performance of H-GO than state-of-the-art unsupervised domain adaptation algorithms.

Keywords: unsupervised person re-ID, heterogeneous graph, domain adaptation

[*] Corresponding author

Email addresses: shaochuanlin19@gmail.com (Shaochuan Lin), jmlv@scut.edu.cn (Jianming Lv), zhengyang5-c@my.cityu.edu.hk (Zhenguo Yang), csqli@comp.polyu.edu.hk (Qing Li), wszheng@ieee.org (Wei-Shi Zheng)

1. Introduction

As a popularly researched task of computer vision, person re-ID aims to retrieve the image frames containing the same person from surveillance videos. Currently, supervised re-ID algorithms [1, 2, 3] on labeled datasets have gained impressive performance. However, as reported by the recent study [4], directly deploying a trained re-ID algorithm to an unknown unlabeled new camera network may often yield very poor performance. How to effectively optimize the trained model based on the abundant unlabeled data collected in the target domain is a quite challenging problem for domain adaptation.

Recently, some unsupervised domain adaptation algorithms of person re-ID were proposed to incrementally optimize the cross-domain transferred model. As the widely used techniques, some **pseudo label based methods** [5, 6, 7] were shown to be also effective in this task. Specifically, the research of [5] conducted clustering on unlabeled instances and assigned the pseudo label of each instance as its corresponding cluster ID. Multi-Label [6] took another way to assign each instance with multiple labels by comparing the similarity between the instance and the labeled data of the source domain. Tracklet [7] associated the successive image frames containing the same person in a camera, and assigned them with the same label for semi-supervised learning. Although these pseudo label based methods can provide definite supervised signal to optimize the visual model, how to effectively reduce the noise of assigning pseudo labels is usually quite challenging. Distinct from the hard assignment of pseudo labels, some **similarity based 'soft' methods** were proposed to optimize the model based on the similarity between images. In particular, TFusion [4] used a fusion model to teach the visual model with the comparison results between instances. ECN [8] optimized the model by measuring each image with the exemplar memories of other images.

None of above methods has considered the diverse topological structures of images in the whole dataset, which can provide abundant deep insight of instances based on structural context and can be used for better model optimization. As observed in [9], the manifold structures are ubiquitous in person image datasets, due to the diversity of the appearance in different camera views. As shown in Fig. 1(a) and 1(b), the instances

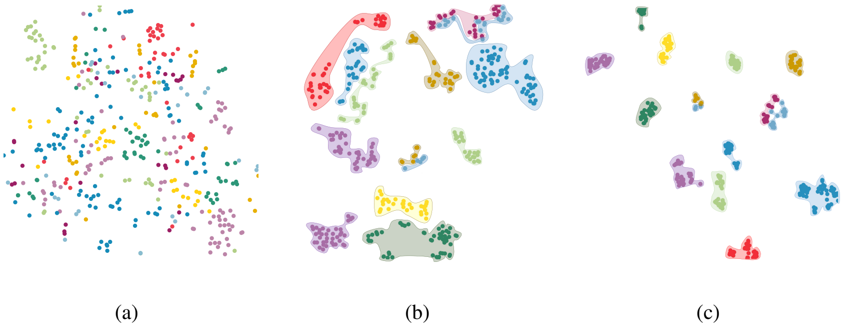


Figure 1: The t-SNE visualization of instance features generated by different visual models on Market-1501. (a) ResNet pre-trained on ImageNet. (b) ResNet pre-trained on DukeMTMC-reID. (c) ResNet pre-trained on DukeMTMC-reID and optimized on Market-1501 using H-GO.

belonging to the same class are distributed in some unpredictable manifold structures, and thus have larger Euclidean distance than those belonging to different classes. In this case, Euclidean distance cannot depict the relationship between images precisely, which increases the difficulty to classify the instances correctly in the Euclidean space. Furthermore, when transferring a model from the source domain to the target domain, the diversity of manifold structures of instances in different domain brings more unpredictable situations to the transferred model to recognize the relationship between instances correctly. As shown in Fig. 1(b), the transferred models that are pre-trained in other datasets cannot group the instances tightly.

To tackle such problem, we propose a Heterogeneous Graph driven Optimization scheme, namely **H-GO**, to explore the deep relationships beneath the manifold structures of unlabeled instances and generate better visual representation for easier classification. Specifically, H-GO models the relationship of the unlabeled images from varied cameras as a heterogenous graph, and adopts a novel heterogeneous affinity propagation method to explore the structures related affinity between instances. Furthermore, a heterogeneous affinity learning procedure is proposed to iteratively optimize the visual model to achieve better presentation of instances, as exemplified by Fig. 1(c).

The advantages of this study can be highlighted as follows:

(i) We are the first to model the structural affinity of unlabeled images as heterogeneous graph and propose the heterogeneous affinity propagation method for incremental learning. The Heterogeneous Graph driven Optimization brings a new way to effective domain adaptation.

(ii) While comparing with the pseudo label based methods [5] [6] [7], H-GO is **much softer** by learning the heterogeneous affinity between instances instead of assigning hard labels. Meanwhile, comparing with the similarity based methods [4] [8], H-GO explores **much deeper** structural affinity of instances aided by the heterogeneous affinity learning procedure.

(iii) Comprehensive experiments are conducted on three benchmarks (*i.e.*, Market-1501, DukeMTMC-reID and MSMT17). The experimental results demonstrate the superior performance of H-GO, and show its flexibility to be combined with other pseudo label based methods to significantly improve their performance.

2. Related Work

Unsupervised Person re-ID. Most of state-of-the-art person re-ID algorithms are in a supervised manner, by relying on sufficient labeled person pairs across cameras[3, 2]. However, the performance of these models is observed to have a serious drop when we directly transfer the model to another unlabeled dataset. To solve the domain adaptation problem, several recent works [4, 10, 7] attempted to tackle this problem by using the deep learning framework. Fan *et al.* [11] initialized the model in the source domain and fine-tuned with the pseudo-labels assigned by the K-means clustering algorithm in the target domain. Wang *et al.* [12] and Lin *et al.* [13] aligned the attributes of unlabeled data to labeled source data, and obtained an impressive improvement. Rather than utilizing the visual domain only, Lv *et al.* [4] paid attention to integrating the spatial-temporal information with visual features. Deng *et al.* [14], Wei *et al.* [15] and Liu *et al.* [16] applied Generative Adversarial Networks [17] to reduce the domain shift between different datasets. Zhong *et al.* [18, 8] proposed to address intra-domain variations of target domain with the help of camera style tansferred images. Most of the above algorithms, however, measure the relationship between a pair of images based on the similarity of their visual features, hence ignoring the global distribution and the topological structures of the images in the whole dataset.

Graph based Supervised Person re-ID. Graph is a very popular data structure to describe complex connections between instances. There were some attempts on incorporating graph into the task of person re-ID recently [19, 9, 20]. Loy *et al.* [19] formulated Laplacian-based method to obtain manifold ranking results. Bai *et al.* [9] investigated the underlying manifold structures of images through affinity graph. Zhong *et al.* [20] proposed the mutual K-NN encoding method and adopted the Jaccard distance with original distance to obtain robust ranking results. All of these methods were used as a post-processing step for person re-ID, which can not help improve the performance of visual models. From another perspective, Shen *et al.* [21] proposed a group-shuffling random walk network to leverage the affinity between gallery images for supervised training. Comparing with these graph based methods, there are two distinct features of H-GO proposed in this paper: 1) H-GO is designed for unsupervised domain adaptation and is able to utilize the unlabeled data for incremental optimization; 2) different from the homogenous graph based existing methods, H-GO models the relationship of images as heterogeneous graph and applies a novel heterogeneous affinity learning method for unsupervised learning.

3. Heterogeneous Graph Driven Optimization

3.1. Preliminary

Person re-ID aims to retrieve the surveillance videos for the image frames which contain the same person. The essence of re-ID is an image retrieval problem dedicated to person recognition. Formally, each surveillance image containing a person is denoted as I_i, which is cropped from an image frame in a surveillance video. The ID of the person in I_i is denoted as $\Gamma(I_i)$. Given a query image I_i, person re-ID is to achieve the image set: $\{I_k|\Gamma(I_k) = \Gamma(I_i), I_k \in \Omega\}$. Here Ω is the surveillance image dataset.

3.2. Framework Overview

We propose a novel Heterogeneous Graph Driven Optimization scheme (H-GO) to perceive the manifold structures of unlabeled data for continuous optimization of the

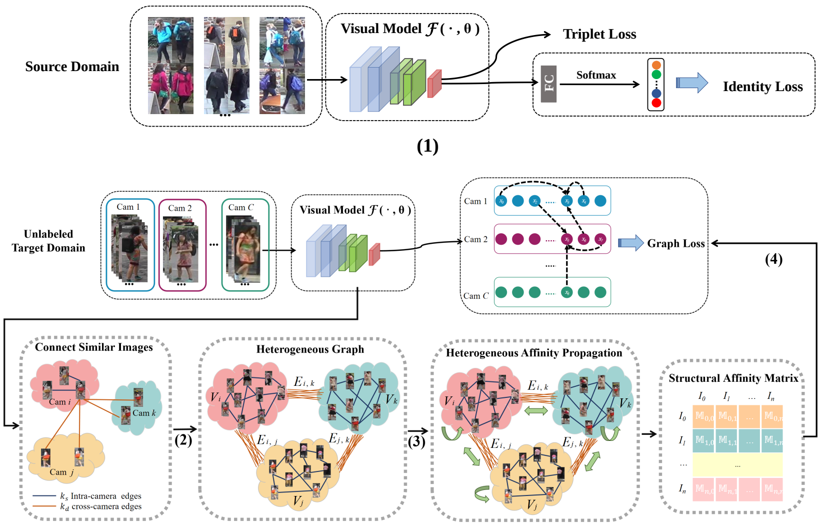


Figure 2: The framework of H-GO with four key steps:(1) Supervised learning in the source domain, where the Identity Loss and Triplet Loss are defined in Eq.(1) and Eq.(2) respectively; (2) A heterogeneous graph is built in the target domain; (3) A Heterogeneous Affinity Propagation procedure is applied on the heterogeneous graph to reveal the structural affinity between images; (4) The newly learned structural affinity of images are fed into a Heterogeneous Affinity Learning algorithm to incrementally optimize the visual model. The Graph Loss is defined in Eq. (12).

visual model. The overall framework of H-GO is shown in Fig. 2, which consists of four main stages:

- **(1): Supervised learning in the source domain.** The deep visual model based on ResNet-50 [22], which is trained in the labeled source dataset, is transferred to the target dataset to extract the visual features of images.
- **(2): Heterogeneous graph building in the target domain.** A heterogeneous graph of unlabled images in the target domain is built by considering the heterogeneous affinity of images from various cameras with diverse visual styles.
- **(3): Heterogeneous Affinity Propagation.** A Heterogeneous Affinity Propagation procedure is applied on the heterogeneous graph to reveal the structural affinity between images.
- **(4): Heterogeneous Affinity Learning.** The newly learned structural affinity of

images are fed into a Heterogeneous Affinity Learning algorithm to incrementally optimize the visual model.

By iteratively running steps (2), (3) and (4), both of the heterogeneous graph and visual model keep evolving iteratively based on the unlabeled data in the target domain. In the following subsections, we detail the design and analysis of each step of our proposed framework.

3.3. Supervised Learning in Source Domain

Given the source data $I_i^{(s)}$ and its one-hot encoding label $\Gamma(I_i^{(s)}))$, we adopt ResNet-50 [22] together with a batch normalization layer (BN) [23] to extract the visual features $\mathcal{F}(I_i, \theta)$, as shown in Fig. 2. A fully connected layer (FC in Fig. 2) is applied as the classifier, which is denoted as $\mathcal{C}^{(s)}$, to process the visual feature and output the probability of predicted identities. The whole network is optimized with respect to an identity loss and a hard-batch triplet loss [24] as follows,

$$\mathcal{L}_{identity} = -\frac{1}{N_s}\sum_{i=1}^{N_s} log(Pr(\Gamma(I_i^{(s)})|C^{(s)}(\mathcal{F}(I_i^{(s)}, \theta)))) \tag{1}$$

$$\begin{aligned}\mathcal{L}_{triplet} = \frac{1}{N_s}\sum_{i=1}^{N_s} max(0, m + ||\mathcal{F}(I_i^{(s)}, \theta) - \mathcal{F}(I_{i,p}^{(s)}, \theta)|| \\ -||\mathcal{F}(I_i^{(s)}, \theta) - \mathcal{F}(I_{i,n}^{(s)}, \theta)||)\end{aligned} \tag{2}$$

Here $Pr(\Gamma(I_i^{(s)})|C^{(s)}(\mathcal{F}(I_i^{(s)}, \theta)))$ indicates the probability predicted by the model. $||\cdot||$ is the L^2-norm distance, $I_{i,p}^{(s)}$ and $I_{i,n}^{(s)}$ indicate the hardest positive and hardest negative sample in each mini-batch for $I_i^{(s)}$, and m = 0.5 denotes the triplet distance margin. By minimizing both the identity and triplet loss, the model can be optimized to predict labels in the source dataset correctly. However, due to the visual diversity of different camera networks, the accuracy of the model usually declines seriously when testing on the target domain. In the following subsections, we will describe how to effectively utilize the abundant unlabeled data in the target domain to incrementally optimize the model.

3.4. Heterogeneous Graph Building

As shown in Fig. 1(b), diverse manifold structures would cause poor performance for transferred visual models, where the intra-class distance may be much larger than inter-class distnce. How to make better use of the structural information of instances in the target domain is the key of domain adaptation. Based on this observation, we establish a heterogeneous graph of the unlabeled images in the target domain to model the global structure of the whole dataset. As shown in Fig. 2, the heterogeneous graph is formally denoted as follows:

$$\begin{aligned} G &=< V, E > \\ V &= \{V_i | 0 < i \leq \mathcal{C}\} \\ E &= \{E_{i,j} | 0 < i, j \leq \mathcal{C}\} \end{aligned} \tag{3}$$

Here $V = \{V_i | 0 < i \leq \mathcal{C}\}$ indicates the image set in the target domain, where V_i is the subset of images from the camera i in the dataset. $\mathcal{C}$ is the total number of the cameras. For each image $I_i \in V$, its visual feature vector is generated by the transferred visual model and is denoted as v_i. The affinity between any pair of images $I_i, I_j \in V$ is measured as follows:

$$\mathcal{S}_{ij} = exp(-\frac{||v_i - v_j||^2}{2\sigma^2})(\sigma > 0) \tag{4}$$

Taking the form of the Gaussian kernel function aims to enhance the non-linear discriminative ability of the affinity measurement.

E in Eq. (3) denotes the edge set of the graph G, and $E_{i,j}$ indicates the subset of the edges between the image set V_i and V_j. Specifically, $E_{i,i}$ indicates the connections between the images from the same camera i. These edge sets are built in two stages. Firstly, for any image $I_t \in V_i$ from the camera i, we can rank the other images from the same camera according to their affinity, and select the top k_s ones to build the edges in $E_{i,i}$. Secondly, for any image $I_t \in V_i$, we rank the other images in $\{V_j | j \neq i\}$ according to their affinity and select the top k_d ones to build the edges. These edges form the cross-camrea edges $\{E_{i,j} | j \neq i\}$. In this way, a sparse K-NN heterogeneous graph can be built, which preserves the important adjacent structural relationship of

Algorithm 1 The Heterogeneous Affinity Learning Algorithm

```
Require: $\mathcal{F}(I_i, \theta)$: the visual feature vector of $I_i$ generated by the visual model $\mathcal{F}$; $\theta$:
    the parameters required to optimize; $\mathbb{M}$: the Deep Affinity Matrix; $\alpha$: the constant
    learning rate; $\gamma$: the convergency condition; $\mu$: the positive constant that controls
    the rate of feature update.
Ensure: optimized $\theta^*$.
function LEARNAFFINITY($\theta$, $\mathbb{M}$)
    $F_k \leftarrow \mathcal{F}(I_k, \theta)(1 \leq k \leq |V|)$ //Extract the feature vector of each image
    Calculate $W^{(\lambda)}$ based on $\mathbb{M}$ by Eq. (13)
    for $i = 1 \to |V|$ do
        Calculate $\tilde{\mathcal{L}}^{+}_{graph}$ based on $W^{(\lambda)}$ by Eq. (15)
        $\theta \leftarrow \theta - \alpha \frac{\partial \tilde{\mathcal{L}}^{+}_{graph}}{\partial \theta}$ //Optimize the parameter set $\theta$.
        $F_i \leftarrow \mu F_i + (1 - \mu)\mathcal{F}(I_i, \theta)$ //Update the feature vector
    end for
    $\theta^* \leftarrow \theta$
end function
```

the images. Note that k_s and k_d are both positive constant, which are set to control the sparsity of the graph.

3.5. Heterogeneous Affinity Propagation

After achieving the Heterogeneous graph, in order to explore the graph structure related affinity between instances, we present here a random walk based algorithm to measure the affinity by the arrival probability of random walkers in the graph. However, because of the diversity of visual styles in different cameras, the affinity $\mathcal{S}_{ij}$ between different images is heterogeneous. The images coming from the same camera tend to have larger affinity than those from different cameras. This heterogeneous property makes traditional random walk quite biased to the local neighborhood of the images from the same camera.

To echo on the above challenge, we propose a Heterogeneous Affinity Propagation algorithm by considering both the visual affinity and the camera tags of the images.

Specifically, starting from each node I_i, random walkers are initialized to walk along the edges. The probability P_{ij} of a walker at I_i to select each next hop I_j from its neighbor is determined as follows:

$$Pr_{ij} = \frac{|\{I_k | I_k \in \mathcal{N}(I_i), I_k \bigodot I_j\}|}{|\mathcal{N}(I_i)|} \cdot \frac{\mathcal{S}_{ij}}{\sum_{I_k \bigodot I_j, I_k \in \mathcal{N}(I_i)} \mathcal{S}_{ik}} \tag{5}$$

Eq. (5) means that the neighbors of I_i are firstly grouped according to their camera tags, and the probability to walk into a group is proportional to the size of the group. $\mathcal{N}(I_i)$ is the neighbor set of I_i in the graph G. $I_k \bigodot I_j$ means that I_k is taken from the same camera as I_j. The transition probability is normalized in each group as $\frac{s_{ij}}{\sum_{I_k \bigodot I_j, I_k \in \mathcal{N}(I_i)} s_{ik}}$. By adopting this heterogeneous normalization, the side-effect of the diversity between varied cameras can be erased smoothly and the long dependency of cross-camera images may have a good chance to be explored.

Based on Eq. (5), the transfer probability matrix $M^{(0)} \in \mathbb{R}^{|V| \times |V|}$ of the graph is set as: $M^{(0)}_{i,j} = Pr_{ij}(1 \leq i, j \leq |V|)$. $M^{(0)}$ indicates the affinity between adjacent nodes in the graph. Inspired by the process of manifold ranking [25], the affinity can be propagated by random walk as:

$$\begin{aligned} M^{(1)} &= \omega M^{(0)} M^{(0)} + (1 - \omega) M^{(0)} \\ M^{(2)} &= \omega M^{(1)} M^{(0)} + (1 - \omega) M^{(0)} \\ &\cdots \\ M^{(t)} &= \omega M^{(t-1)} M^{(0)} + (1 - \omega) M^{(0)} \end{aligned} \tag{6}$$

where $\omega \in [0, 1]$ is the weight to control relative contributions between the affinity update and the original affinity matrix. t denotes the t^{th} affinity propagation. As $t \to \infty$, we can have,

$$M^{(\infty)} = (1 - \omega)(I - \omega M^{(0)})^{-1} M^{(0)} \tag{7}$$

where I is the identity matrices.

$M^{(\infty)}$ indicates the affinity propagation matrix, where the element $M^{(\infty)}_{i,j}$ denotes the structural relationship between the images I_i and I_j. The i^{th} row vector $M^{(\infty)}_i$ indicates the global structure context of the image I_i. Based on $M^{(\infty)}$, a Structural

Affinity Matrix $\mathbb{M}$ is proposed to measure the structure based affinity between images, where each element $\mathbb{M}_{i,j}$ is defined as:

$$\mathbb{M}_{i,j} = exp(-\frac{||M_i^{(\infty)} - M_j^{(\infty)}||^2}{2\sigma^2}) \tag{8}$$

Here $M_i^{(\infty)}$ and $M_j^{(\infty)}$ indicate the i^{th} and j^{th} rows of $M^{(\infty)}$. Eq (8) also takes the similar form of the Gaussian kernel function like Eq (4) to enhance the non-linear discriminative ability of affinity measurement.

The bigger value of $\mathbb{M}_{i,j}$ indicates that the two images share more similar structure context and belongs to the same class with higher probability.

3.6. Heterogeneous Affinity Learning

After achieving the structure based affinity $\mathbb{M}$ learned from the graph, we propose next a Heterogeneous Affinity Learning method to utilize the derived affinity to optimize the deep visual model. The optimization goal is to pull closer the feature vectors of the images which have larger structural affinity, and push those with smaller affinity further apart. According to this objective, the basic loss function can be defined as the cross entropy of the affinity distribution and the predicted similarity, as follows:

$$\mathcal{L}_{graph} = -\sum_{i=1}^{|V|} \sum_{I_j \in \mathbb{N}(I_i)} W_{ij} log(\mathcal{P}(\mathcal{F}(I_j, \theta)|\mathcal{F}(I_i, \theta))) \tag{9}$$

Here W_{ij} is the normalized affinity shown below:

$$W_{ij} = \begin{cases} \frac{\mathbb{M}_{ij}}{\max\limits_{I_k \in \mathbb{N}(I_i)}(\mathbb{M}_{ik})}, & j \neq i \wedge I_j \in \mathbb{N}(I_i) \\ 0, & j \neq i \wedge I_j \notin \mathbb{N}(I_i) \\ 1, & j = i \end{cases} \tag{10}$$

$\mathbb{M}_{ij}$ is the structural affinity defined in Eq. (8). $\mathbb{N}(I_i)$ indicates the collection of images which have the highest structural affinity with I_i. In practical implementation, we rank the images according to the structural affinity with I_i, and select the top $k_a (k_a > 0)$ images to form the set. The function $\mathcal{F}(I_i, \theta)$ indicates the visual feature vector generated by the visual model $\mathcal{F}$, which takes I_i as input and θ as its parameters needed for optimization.

The function $\mathcal{P}(\mathcal{F}(I_j,\theta)|\mathcal{F}(I_i,\theta))$ in Eq. (9) defines the predicted probability that I_i has the same identity with another image I_j :

$$\mathcal{P}(\mathcal{F}(I_j,\theta)|\mathcal{F}(I_i,\theta)) = \frac{exp(\mathcal{F}(I_j,\theta)\cdot\mathcal{F}(I_i,\theta)/\tau)}{\sum\limits_{I_k\in\mathbb{N}(I_i)} exp(\mathcal{F}(I_k,\theta)\cdot\mathcal{F}(I_i,\theta)/\tau)} \tag{11}$$

Here, τ is a temperature parameter [26] that is designed to tune the softness of probability distribution over classes. By minimizing the loss $\mathcal{L}_{graph}$, the visual model may tend to generate similar feature vectors of the images which have higher structural affinity.

Considering the heterogeneous property of the affinity measurement of the images from various cameras with different visual styles, the optimization goal of Eq. (9) can be further extended to consider the diversity of the cameras:

$$\mathcal{L}^{+}_{graph} = -\sum_{i=1}^{|V|}\sum_{\lambda=1}^{C}\sum_{I_j\in\mathbb{N}(I_i),I_j\in V_\lambda} W^{(\lambda)}_{ij} log(\mathcal{P}^{(\lambda)}(\mathcal{F}(I_j,\theta)|\mathcal{F}(I_i,\theta))) \tag{12}$$

Here λ is the camera ID, and C is total number of cameras. $I_j \in V_\lambda$ indicates that the image I_j is captured from the camera λ. Eq. (12) groups the nodes according to the camera ID, and measure the affinity in different groups independently. Specifically, $W^{(\lambda)}$ is the group-related affinity measurement defined as:

$$W^{(\lambda)}_{ij} = \begin{cases} \frac{\mathbb{M}_{ij}}{\max\limits_{I_k\in\mathbb{N}(I_i),I_k\in V_\lambda}(\mathbb{M}_{ik})}, & j\neq i \wedge I_j\in\mathbb{N}(I_i) \\ 0, & j\neq i \wedge I_j\notin\mathbb{N}(I_i) \\ 1, & j=i \end{cases} \tag{13}$$

Meanwhile, $\mathcal{P}^{(\lambda)}$ is the extension of Eq. (11) by adding the camera specific normalization:

$$\mathcal{P}^{(\lambda)}(\mathcal{F}(I_j,\theta)|\mathcal{F}(I_i,\theta)) = \frac{exp(\mathcal{F}(I_j,\theta)\cdot\mathcal{F}(I_i,\theta)/\tau)}{\sum\limits_{I_k\in\mathbb{N}(I_i),I_k\in V_\lambda} exp(\mathcal{F}(I_k,\theta)\cdot\mathcal{F}(I_i,\theta)/\tau)} \tag{14}$$

In practical implementation, the cost to optimize the parameter set θ by directly minimizing $\mathcal{L}^{+}_{graph}$ is very expensive, because the generated feature vector of each image I_i is required to calculate the correlation with the vectors of a bunch of other images according to Eq. (12) and Eq. (14).

Algorithm 2 Evolving of the Heterogeneous Graph

```
Require: $\mathcal{F}$: the visual model; $\theta$: the parameters required to optimize; $t_{switch}$: the
    balance control parameter.
Ensure: optimized $\theta^*$.
1:  repeat
2:      $v_i \leftarrow \mathcal{F}(I_i, \theta)(1 \leq i \leq |V|)$
3:      $G =< V, E > \leftarrow$ Heterogeneous Graph Building
4:      $\mathbb{M} \leftarrow$ Heterogeneous Affinity Propagation on $G$
5:      for $k = 1 \rightarrow t_{switch}$ do
6:          $\theta^* \leftarrow \mathrm{LearnAffinity}(\theta, \mathbb{M})$ //Heterogeneous Affinity Learning
7:          $\Delta\theta \leftarrow \theta^* - \theta$
8:          $\theta \leftarrow \theta^*$
9:      end for
10: until $(|\Delta\theta| < \gamma)$
11: $\theta^* \leftarrow \theta$
```

Inspired by [8], we adopt an alternate optimization method to reduce the optimization cost. The algorithm is presented in Algorithm 1. Specifically, we define an approximation form of $\mathcal{L}^+_{graph}$ as follows:

$$\tilde{\mathcal{L}}^+_{graph} = -\sum_{i=1}^{|V|}\sum_{\lambda=1}^{C}\sum_{I_j \in \mathbb{N}(I_i), I_j \in V_\lambda} W_{ij}^{(\lambda)} \cdot \log(\mathcal{P}^{(\lambda)}(F_j|\mathcal{F}(I_i,\theta))) \tag{15}$$

Here F_j means the visual feature of the image I_j. $\mathcal{P}^{(\lambda)}(F_j|\mathcal{F}(I_i,\theta))$ of Eq. 15 is defined as:

$$\mathcal{P}^{(\lambda)}(F_j|\mathcal{F}(I_i,\theta)) = \frac{exp(F_j \cdot \mathcal{F}(I_i,\theta)/\tau)}{\sum\limits_{I_k \in \mathbb{N}(I_i), I_k \in V_\lambda} exp(F_k \cdot \mathcal{F}(I_i,\theta)/\tau)} \tag{16}$$

As shown in the Line 2 of Algorithm 1, the visual feature of each image is extracted at the beginning. Following Line 4 of Algorithm 1, for each input image I_i, we calculate the loss according to Eq. (15), which keeps the visual features of other images as constant (F_j) and only uses the feature vector of I_i (i.e. $\mathcal{F}(I_i,\theta)$) to propagate backward the gradient to optimize the parameter set θ (Line 5 of Algorithm 1).

After the optimization of θ, the visual feature F_i is recalculated according to Line 6 of Algorithm 1. While keeping most of the visual features as constant during gradient propagation, the computing cost can be greatly reduced for parameter optimization.

3.7. Evolving of the Heterogeneous Graph

After the visual model is optimized according to the Heterogeneous Affinity Learning algorithm (Algorithm 1), the updated model can be utilized to generate the visual feature of each image and re-build the Heterogeneous Graph according to Section 3.4. As shown in Fig. 2, by running in an iterative manner, both the visual model and the heterogeneous graph evolve alternately. The detail of algorithm is shown in Algorithm 2. The hyper-parameter t_{switch} in Line 5 is adopted to tune the strength of the Heterogeneous Affinity Learning procedure in each iteration of evolution.

4. Experiments

4.1. Dataset and Experiment Settings

Dataset. Our approach [1] is evaluated on three large-scale datasets: Market-1501 [27], DukeMTMC-reID [28, 29] and MSMT17 [15]. Specifically, Market-1501 contains 12,936 training images, 3,368 query images and 19,732 gallery images, which include 1,501 identities from 6 cameras. DukeMTMC-reID is collected from 8 cameras, and contains 16,522 training images, 2,228 query images and 17,661 gallery images. As the largest Person re-ID open dataset, MSMT17 contains 126,441 images with 4,101 identities taken from 15 cameras in a campus. As the default configuration following [15], 32,621 images are selected for training, while 11,659 query images and 82,161 gallery images are prepared for testing.

Evaluation Metric. The performance is evaluated by the the Cumulative Matching Characteristic (CMC) curve and mean Average Precison (mAP). The CMC scores reflect the precision of the retrieval, while the mAP indicates the recall.

Implementation Details:

[1] Source Code: https://github.com/linshoa/H-GO

Table 1: Comparison of variation models. Heter-G and Homo-G denotes the building of Heterogeneous Graph and Homogenous Graph, respectively. Heter-AP and Homo-AP means Heterogeneous Affinity Propagation and Homogenous Affinity Propagation, respectively. L_{graph} and L^{+}_{graph} are two different ways for Heterogeneous Affinity Learning. Market-1501 and DukeMTMC-reID are abbreviate as ‘M’ and ‘D’, respectively.

	Components						D → M		M → D	
	Heter-G	Homo-G	Heter-AP	Homo-AP	L^{+}_{graph}	L_{graph}	Rank-1	mAP	Rank-1	mAP
0	✗	✗	✗	✗	✗	✗	59.1	28.7	45.1	27.5
1	✗	✓	✗	✗	✗	✓	66.2	38.0	48.3	31.6
2	✓	✗	✗	✗	✗	✓	81.0	51.3	65.8	44.8
3	✗	✓	✗	✗	✓	✗	79.8	51.0	66.0	44.8
4	✓	✗	✗	✗	✓	✗	80.9	51.1	66.1	45.0
5	✗	✓	✗	✓	✗	✓	71.7	42.0	53.9	35.5
6	✗	✓	✓	✗	✗	✓	73.5	43.1	55.9	37.3
7	✗	✓	✗	✓	✓	✗	84.4	56.8	70.7	51.5
8	✗	✓	✓	✗	✓	✗	**85.8**	**58.4**	70.5	51.1
9	✓	✗	✗	✓	✗	✓	84.1	53.3	72.7	54.0
10	✓	✗	✓	✗	✗	✓	83.6	54.0	**74.1**	**55.3**
11	✓	✗	✗	✓	✓	✗	85.0	57.6	71.1	51.6
12	✓	✗	✓	✗	✓	✗	85.0	57.6	71.5	52.0

- **Supervised learning in the source domain.** We pre-trained the backbone on ImageNet [30]. Adam optimizer is selected and the mini-batch size is set as 64 in training. We train the backbone with the learning rate of 0.00035 in total 80 epochs.

- **Training with H-GO.** During training, the input images are resized as 256 × 128 with the data augmentation of random cropping, flipping, color jittering. We use Adam optimizer with a mini-batch size of 64 for target data. We train the visual model for 8 epochs with the learning rate of 0.00035. The wight ω in Eq. (7) and the hyper-parameter of σ in Eq. (4) are empirically set as 0.99 and 1, respectively. The temperature parameter (τ in Eq. (11)) is set as 0.05. Moreover, with the help of grid search, k_s, k_d, k_a and t_{switch} are set as 2, 4, 14 and 2, respectively. All experiments are conducted on one GTX 1080Ti GPU with 80 CPU cores, 128G memory.

4.2. Ablation Studies

In this subsection, a series of ablation studies are given to demonstrate the effectiveness of each major component of H-GO on the datasets Market-1501 [27] and DukeMTMC-reID [28]. The comparison results are shown in Table 1. The baseline model is shown as 0.

Effectiveness of the Heterogeneous Graph. Without the Heterogeneous Graph based optimization, the baseline model (0 in Table 1) shows extremely poor performance, which is transferred from the source dataset to the target domain without any incremental learning. Specially, with optimization, the performance is further improved when comparing the model 0 with model 2. For instance, on DukeMTMC-reID, the mAP have raised from 27.5% to 44.8%(+17.3%). Moreover, we also compare H-GO with the variation model using the Homogenous Graph (namely Homo-G in Table 1), which selects the top-k affinity from all possible pairs of images without distinguishing from the same or different cameras. It can be observed that the mAP is improved by +13.2% on DukeMTMC-reID when comparing the Heterogeneous Graph in model 2 with the Homogenous Graph in model 1. This demonstrates the effectiveness of the Heterogeneous Graph. Due to the diversity of visual styles in different cameras, the images coming from the same camera tend to have larger affinity than cross-camera ones. Thus, in the Homogenous Graph, few cross-camera edges can be built and there is no enough cross-camera relationship which is critical for person re-ID can be learned to optimize the visual model. This is the reason why considering the heterogeneous property of the graph is important in H-GO.

Effectiveness of Heterogeneous Affinity Propagation. We also make variant model with the heterogeneous affinity propagation as shown as model 1, model 5 and model 6 in Table 1. When adding this component in model 6, the performance is further improved comparing with model 1. For instance, on Market-1501, the mAP has significantly raised from 38.0% to 43.1%(+5.1%). What is more, drops are observed when we replace the component of heterogeneous affinity propagation by homogenous affinity propagation in model 5, $e.g.$, the mAP drops from 37.3% to 35.5%(-1.8%) on DukeMTMC-reID. This shows the effectiveness of infering the structure related deep affinity between instances.

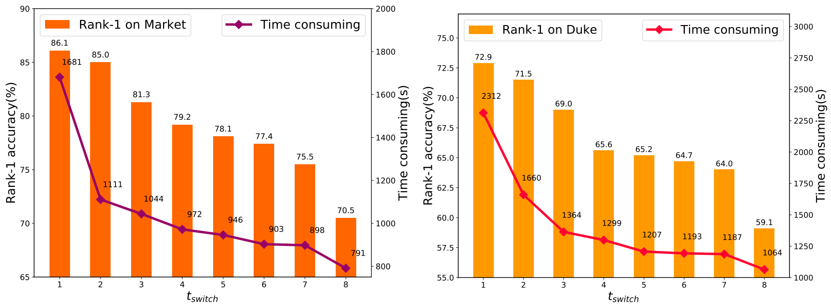


Figure 3: Training time and rank-1 accuracy of the H-GO models with different parameter t_{switch}. Market and Duke donotes Market-1501 and DukeMTMC-reID, respectively.

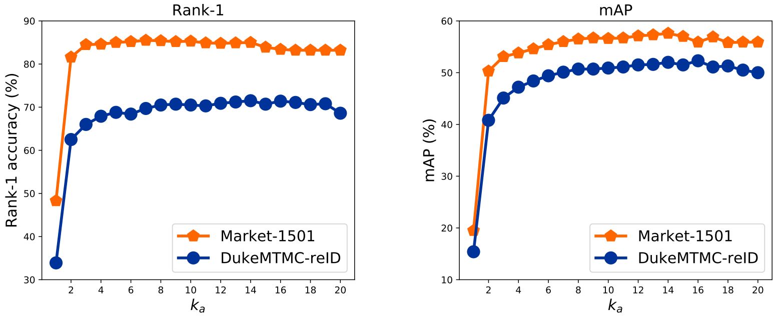
Table 2: Performance evaluation of our proposed H-GO with different affinity measure methods in Eq. 4 and Eq. 8.

Affinity Measure Methods	DukeMTMC $\rightarrow$ Market		Market $\rightarrow$ DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
Cosine-similarity	76.5	46.2	66.6	46.6
Gaussian kernel function	**85.0**	**57.6**	**71.5**	**52.0**

Effectiveness of Heterogeneous Affinity Learning. There are two kinds of Heterogeneous Affinity Learning methods proposed in Section 3.6: the basic version which takes L_{graph} (Eq. (9)) as the loss function, and the advanced model with L^{+}_{graph} (Eq. (12)) which considers the heterogeneous properties of the images taken from varied cameras. It can be observed from Table 1 that L^{+}_{graph} usually works much better than L_{graph}, especially in the cases based on homogenous graph by comparing the following pairs of models: (1,3), (5,7), and (5,8). However L^{+}_{graph} works weaker than L_{graph} by comparing model 10 and 12 on DukeMTMC-reID, since the heterogeneous property of the graph has been considered in graph building and affiniy propagation.

Effectiveness of Gaussian kernel function. We also make ablation studies to replace the Gaussian kernel function in Eq. 4 and Eq. 8 with traditional cosine-similarity, which can be seen in Table 2. It is clear that Gaussian kernel function performances better than traditional cosine-similarity, since it enhances the non-linear discriminative ability during the procedure of affinity measurement.

Effectiveness of Heterogeneous Graph Evolving. As detailed in Algorithm 2, the



(a)

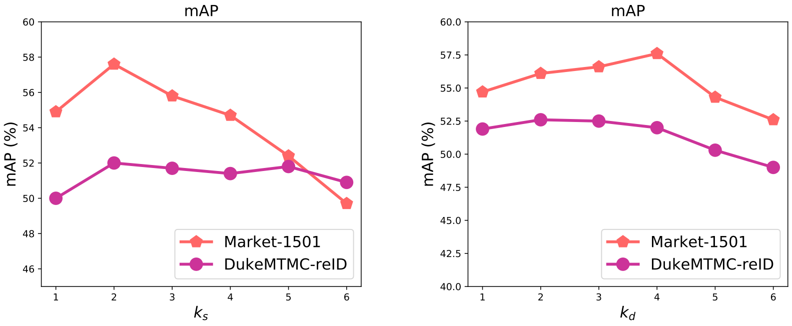
(b)

Figure 4: Parameters analysis of k_a on Market-1501 and DukeMTMC-reID.

Table 3: Performance evaluation with differtent values of τ in Eq. (16) on Market-1501 and DukeMTMC-reID.

τ	DukeMTMC $\rightarrow$ Market		Market $\rightarrow$ DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
1.0	67.0	37.1	56.6	36.0
0.5	68.3	38.1	57.7	37.4
0.1	76.3	47.8	66.1	45.3
0.06	83.8	55.9	70.7	51.5
0.05	**85.0**	**57.6**	**71.5**	**52.0**
0.04	84.0	56.0	69.1	49.3

evolving of the Heterogeneous Graph and the visual model can be run in an iterative manner. We analyze the configuration of t_{switch} which indicates the strength of the Heterogeneous Affinity Learning in each iteration of evolution. Fig. 3 shows the accuracy together with the training cost of different models with varied t_{switch}. It can be observed that larger t_{swith} may reduce the time consuming of training. This is because the graph building is relatively costly, and increasing t_{swith} in Algorithm 2 may decrease the frequency of graph building. On the other hand, increasing t_{swith} may also decrease the performance of the model, which confirms the importance of the evolution of the Heterogeneous Graph. In real deployment, t_{switch} can be configured according to the tradeoff between the performance and training cost. We set $t_{switch} = 2$ by default in the experiments, so that the training can be finished in around half an hour.



(a) (b)

Figure 5: Parameters analysis of k_s and k_d on Market-1501 and DukeMTMC-reID.

4.3. Parameter Analysis

Temperature parameter τ analysis. The sensitivity to temperature fact τ can be shown in Table 3. We can observe that the lower τ it is, the better results it will produce for the task of upsupervised person re-ID, since it enlarge variance of various probability over classes. However, extremely lower value may harm the model and be overfitting with the data, *e.g*, $\tau = 0.04$. In our experiments, we set $\tau = 0.05$.

Analysis of the graph parameters k_s, k_d and k_a. The parameters k_s and k_d indicate the number of same camera neighbors and different numbers in the graph building, respectively. k_a indicates the heterogeneous affinity learning procedure. Fig. 5(a), Fig. 5(b) and Fig. 4 show that the models can achieve the best performance when k_s is in the range of 2∼4, k_d is in the range of 2∼4 and k_a is in the range of 10∼15. Larger size of neighborhood may even decrease the performance, because setting less similar instances as close neighbors may bring more noise to the relationship exploration procedure. In our experiments, we set $k_s = 2$, $k_d = 4$ and $k_a = 14$.

4.4. Comparisons to State-of-the-Art

Table 4 compares H-GO with state-of-the-art unsupervised domain adaptation algorithms on Market-1501 and DukeMTMC-reID. The comparisons are categorized into two groups: 1) pseudo label based methods which conduct clustering on unlabeled instances and assign the pseudo label of each image as its corresponding cluster ID

Table 4: Comparison with state-of-the-art unsupervised domain adaptation methods on Market-1501 (M) and DukeMTMC-reID (D). AHC and SP indicates agglomerative hierarchical clustering and spectral clustering, respectively.

Methods		D → M		M→ D	
		Rank-1	mAP	Rank-1	mAP
PUL[11]	pseudo label based	45.5	20.5	30.0	16.4
TFusion[4]	similarity based	63.2	21.6	57.1	19.4
HHL[18]	similarity based	62.2	31.4	46.9	27.2
ATNet[16]	similarity based	55.7	25.6	45.1	24.9
CamStyle[31]	similarity based	58.8	27.4	48.4	25.1
UCDA-CCE[32]	similarity based	60.4	30.9	47.7	31.0
PAUL[10]	similarity based	66.7	36.8	56.1	35.7
ECN[8](w/o StarGAN[33])	similarity based	58.0	27.7	39.7	23.6
ECN[8]	similarity based	75.1	43.0	63.3	40.4
PDA-Net[34]	similarity based	75.2	47.6	63.2	45.1
PCB-PAST[35]	pseudo label based	78.4	54.6	72.4	54.3
Tracklet[7]	pseudo label based	85.3	65.2	71.7	50.7
MMT(DBSCAN)[36]	pseudo label based	89.5	73.8	76.3	62.3
SpCL[37]	pseudo label based	90.3	76.7	82.9	68.8
AHC+L^{+}_{graph}	pseudo label based	67.7	40.1	56.9	36.8
KMeans+L^{+}_{graph}	pseudo label based	70.9	44.7	54.0	35.7
SP[38]+L^{+}_{graph}	pseudo label based	73.0	47.4	53.7	34.6
DBSCAN[39]+L^{+}_{graph}	pseudo label based	75.2	44.3	61.5	42.3
H-GO (Heter-G+Heter-Ap+L_{graph})	graph based	83.6	54.0	74.1	55.3
H-GO (Heter-G+Heter-Ap+L^{+}_{graph})	graph based	85.0	57.6	71.5	52.0
H-GO+Tracklet[7]	-	89.5	74.2	78.9	53.3
H-GO+MMT(DBSCAN)[36]	-	**92.3**	**80.5**	**87.9**	**70.2**
H-GO+SpCL[37]	-	90.4	77.9	82.8	69.1

on the unlabeled target domain, including PUL [11], PCB-PAST [35], Tracklet [7], MMT(DBSCAN) [36] and SpCL[37]; 2) similarity based methods which are proposed to optimize the model by using the similarity between images, including TFusion [4], HHL [18], ATNet [16], CamStyle [31], UCDA-CCE [32], PAUL [10], ECN [8] and PDA-Net [34]. Additionally, in order to exploit the flexibility of our method, we also use H-GO as pre-processing method to combine with other pseudo label based methods to significantly improve their performance and obtain new advance. Specially, Tracklet [7]+H-GO means that the combinaton of H-GO and Tracklet, which is trained based on an idea of independent per-camera identity annotation. H-GO+MMT(DBSCAN)

[36] and H-GO+SpCL [37] means that we provide the H-GO pretrained model for MMT(DBSCAN) [36] and SpCL [37], respectively.

As shown in Table 4, H-GO outperforms most existing methods by a large margin. Specially, we achieve 85.0% Rank-1 accuracy and 57.6% mAP on Market-1501, which exceeds the state-of-the-art similarity based method PDA-Net [34] by 9.8% and 10.0%, respectively. Furthermore, similar superior performance of H-GO can be observed on DukeMTMC-reID.

Besides, to further compare the heterogenous graph based method with the clustering based pseudo label methods, we also implement some variant methods by replace the heterogeneous graph with the pseudo labels generated by traditional clustering methods, as shown in Table 4. In these models, the affinity between the images with the same pseudo label is set to 1, and that of the images with different pseudo labels is set to 0. It can be observed that variation models ($e.g.$, DBSCAN[39]+L^{+}_{graph}) can achieve comparable results state-of-the-art algorithms ($e.g.$, ECN [8]), but their performance is much worse than H-GO by a large margin. This shows the superior performance of the heterogenous graph based model when compared with traditional pseudo label based methods.

Furthermore, we also verify the effectiveness to combine H-GO with other pseudo label based methods. Table 4 shows that Tracklet [7]+H-GO surpass Tracklet [7] by margins of 9.0% and 2.6% mAP on Market-1501 and DukeMTMC-reID. More importantly, we achieve state-of-the-art performances on both Market-1501 and DukeMTMC-reID with H-GO+MMT(DBSCAN) [36]. The effectiveness of H-GO is because after using the graph structure based affinity to optimize the model, the instances belonging to the same class are grouped more tightly as shown in Fig. 1(c). That means H-GO can be a very useful pre-processing tool to help the models to generate more precise pseudo labels.

We also evaluate H-GO on the MSMT17 dataset [15], one of the largest person re-ID datasets up to now, to test the generalization ability of H-GO. It shows that, H-GO+MMT(DBSCAN) [36] advances all of the state-of-the-art methods to a new level, $e.g.$, 59.6% and 54.1% rank-1 accuracy on Duke-to-MSMT17 and Market-to-MSMT17 respectively.

Table 5: Performance evaluation on the large dataset MSMT17. (*) the implementation is based on the authors' code. 'M' indicates Market-1501 and 'D' indicates DukeMTMC-reID.

Methods	M → MSMT17			D → MSMT17		
	Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP
PTGAN[15]	10.2	24.4	2.9	11.8	27.4	3.3
ECN[8]	25.3	42.1	8.5	30.2	46.8	10.2
Tracklet[7]	44.1	63.9	18.6	44.1	63.9	18.6
MMT(DBSCAN)[36]	50.1	69.3	24.0	52.9	71.3	25.1
SpCL[37]	51.6	69.7	25.4	53.1	70.5	26.5
SpCL*[37]	48.4	66.5	23.6	48.8	66.7	23.4
H-GO (Heter-G+Heter-Ap+L^{+}_{graph})	25.1	41.2	9.3	36.2	53.1	13.6
H-GO (Heter-G+Heter-Ap+L_{graph})	29.1	44.3	11.0	41.4	58.1	16.2
H-GO (Homo-G+Heter-Ap+L^{+}_{graph})	34.4	51.4	13.9	42.4	59.9	17.4
H-GO+Tracklet[7]	47.7	66.4	21.8	52.4	71.2	24.0
H-GO+MMT(DBSCAN)[36]	**54.1**	**72.4**	**28.0**	**59.6**	**76.5**	**31.2**
H-GO+SpCL[37]	49.1	66.5	23.6	50.3	67.8	24.4

5. Conclusion

In this paper, we propose a novel Heterogeneous Graph driven Optimization scheme (H-GO) for structure based unsuperised learning. In particular, H-GO builds a heterogeneous graph of unlabeled images to consider the heterogeneous properties of images from various cameras with varied visual styles. A heterogeneous affinity propagation method is further applied to explore the graph based affinity between the instances which share similar manifold structures. Finally, a heterogeneous affinity learning procedure is taken to optimize the visual models by using the graph based affinity of instances. Comprehensive experiments are conducted on three large-scale re-ID datasets, and the results demonstrate the flexibility and the superior performance of H-GO than state-of-the-art unsupervised domain adaptation algorithms. In the future, we will further study how the diversity of visual styles affect the optimized parameter setting in H-GO, and thus be able to make the algorithm more self-adaptive to the dynamic change in varied domains.

6. Acknowledgements

The work described in this paper was supported by the grants from Natural Science Foundation of China (No. 61876065), Natural Science Foundation of Guangdong Province, China (No.2018A0303130022), and Science and Technology Program of Guangzhou, China (No.201904010200).

References

[1] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: ECCV, 2018, pp. 501–518. `doi:10.1007/978-3-030-01225-0\_30`.

[2] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Acm Multimedia, 2018, pp. 274–282. `doi:10.1145/3240508.3240552`.

[3] F. Zheng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, A coarse-to-fine pyramidal model for person re-identification via multi-loss dynamic training, in: CVPR 2019, 2019.

[4] J. Lv, W. Chen, Q. Li, C. Yang, Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns, in: CVPR, 2018. `doi:10.1109/CVPR.2018.00829`.

[5] Y. Lin, X. Dong, L. Zheng, Y. Yan, Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, in: AAAI, 2019. `doi:10.1609/aaai.v33i01.33018738`.

[6] H. Yu, W. Zheng, A. Wu, X. Guo, S. Gong, J. Lai, Unsupervised person re-identification by soft multilabel learning, in: CVPR, 2019. `doi:10.1109/CVPR.2019.00225`.

[7] X. Zhu, X. Zhu, M. Li, V. Murino, S. Gong, Intra-camera supervised person re-identification: A new benchmark, in: 2019 IEEE/CVF International Conference

on Computer Vision Workshops, ICCV Workshops 2019, 2019, pp. 1079–1087. doi:10.1109/ICCVW.2019.00138.

[8] Z. Zhong, L. Zheng, Z. Luo, S. Li, Y. Yang, Invariance matters: Exemplar memory for domain adaptive person re-identication, in: CVPR, 2019. doi:10.1109/CVPR.2019.00069.

[9] S. Bai, X. Bai, Q. Tian, Scalable person re-identification on supervised smoothed manifold, in: CVPR, 2017, pp. 3356–3365. doi:10.1109/CVPR.2017.358.

[10] Q. Yang, H. Yu, A. Wu, W. Zheng, Patch-based discriminative feature learning for unsupervised person re-identification, in: CVPR, 2019. doi:10.1109/CVPR.2019.00375.

[11] H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: Clustering and fine-tuning, TOMCCAP. (2018) 83:1–83:18doi:10.1145/3243316.

[12] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: CVPR, 2018, pp. 2275–2284. doi:10.1109/CVPR.2018.00242.

[13] S. Lin, H. Li, C. Li, A. C. Kot, Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification, in: BMVC, 2018, p. 9.

[14] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in: CVPR, 2018. doi:10.1109/CVPR.2018.00110.

[15] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer GAN to bridge domain gap for person re-identification, in: CVPR, 2018, pp. 79–88. doi:10.1109/CVPR.2018.00016.

[16] J. Liu, Z. Zha, D. Chen, R. Hong, M. Wang, Adaptive transfer network for cross-domain person re-identification, in: CVPR, 2019. doi:10.1109/CVPR.2019.00737.

[17] J. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: ICCV, 2017, pp. 2242–2251. doi:10.1109/ICCV.2017.244.

[18] Z. Zhong, L. Zheng, S. Li, Y. Yang, Generalizing a person retrieval model hetero- and homogeneously, in: ECCV, 2018. doi:10.1007/978-3-030-01261-8_11.

[19] C. C. Loy, C. Liu, S. Gong, Person re-identification by manifold ranking, in: ICIP, 2013, pp. 3567–3571. doi:10.1109/ICIP.2013.6738736.

[20] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: CVPR, 2017, pp. 3652–3661. doi:10.1109/CVPR.2017.389.

[21] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, X. Wang, Deep group-shuffling random walk for person re-identification, in: CVPR, 2018, pp. 2265–2274. doi:10.1109/CVPR.2018.00241.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[23] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: ICML, 2015, pp. 448–456.

[24] H. Alexander, B. Lucas, L. Bastian, In defense of the triplet loss for person re-identification, in: CVPR, 2017.

[25] Z. Denny, W. Jason, G. Arthur, B. Olivier, S. Bernhard, Ranking on data manifolds, in: NIPS, NIPS Edition, 2003.

[26] G. E. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015.

[27] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: ICCV, 2015, pp. 1116–1124. doi:10.1109/ICCV.2015.133.

[28] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: ECCV, 2016, pp. 17–35. doi:10.1007/978-3-319-48881-3_2.

[29] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: ICCV, 2017, pp. 3774–3782. doi:10.1109/ICCV.2017.405.

[30] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: A large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[31] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: A novel data augmentation method for person re-identification, IEEE Trans. Image Processing. (2019). doi:10.1109/TIP.2018.2874313.

[32] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, Y. Gao, A novel unsupervised camera-aware domain adaptation framework for person re-identification, in: ICCV, 2019. doi:10.1109/ICCV.2019.00817.

[33] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: CVPR, 2018, pp. 8789–8797. doi:10.1109/CVPR.2018.00916.

[34] Y. Li, C. Lin, Y. Lin, Y. F. Wang, Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation, in: ICCV, 2019, pp. 7918–7928. doi:10.1109/ICCV.2019.00801.

[35] X. Zhang, J. Cao, C. Shen, M. You, Self-training with progressive augmentation for unsupervised cross-domain person re-identification, in: ICCV, 2019, pp. 8221–8230. doi:10.1109/ICCV.2019.00831.

[36] Y. Ge, D. Chen, H. Li, Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification, in: ICLR 2020, 2020.

[37] Y. Ge, F. Zhu, D. Chen, R. Zhao, H. Li, Self-paced contrastive learning with hybrid memory for domain adaptive object re-id, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[38] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416. doi:10.1007/s11222-007-9033-z.

[39] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, 1996, pp. 226–231.