Unsupervised Domain Adaptation for Person Re-identification via Heterogeneous Graph Alignment

Minying Zhang,[1*] **Kai Liu,**[1,2*] **Yidong Li,**[2†] **Shihui Guo,**[3] **Hongtao Duan,**[1] **Yimin Long,**[1] **Yi Jin**[2]
[1] Alibaba Group [2] Beijing Jiaotong University [3] Xiamen University
minying.zmy@alibaba-inc.com, {liukai18, ydli}@bjtu.edu.cn, guoshihui@xmu.edu.cn,
{hongtao.dht, yimin.lym}@alibaba-inc.com, yjin@bjtu.edu.cn

Abstract

Unsupervised person re-identification (re-ID) is becoming increasingly popular due to its power in real-world systems such as public security and intelligent transportation systems. However, the person re-ID task is challenged by the problems of data distribution discrepancy across cameras and lack of label information. In this paper, we propose a coarse-to-fine heterogeneous graph alignment (HGA) method to find cross-camera person matches by characterizing the unlabeled data as a heterogeneous graph for each camera. In the coarse-alignment stage, we assign a projection for each camera and utilize an adversarial learning based method to align coarse-grained node groups from different cameras into a shared space, which consequently alleviates the distribution discrepancy between cameras. In the fine-alignment stage, we exploit potential fine-grained node groups in the shared space and introduce conservative alignment loss functions to constrain the graph aligning process, resulting in reliable pseudo labels as learning guidance. The proposed domain adaptation framework not only improves model generalization on target domain, but also facilitates mining and integrating the potential discriminative information across different cameras. Extensive experiments on benchmark datasets demonstrate that the proposed approach outperforms the state-of-the-arts.

Introduction

Person re-identification(Re-ID) aims to identify the same people across non-overlapping camera views. Most of these studies focus on supervised learning (Zheng, Gong, and Xiang 2012; Liao et al. 2015; Li et al. 2014; Ahmed, Jones, and Marks 2015; Sun et al. 2017). In real-world applications, the performance of the supervised learning method quite depends on the quality of labeled data. Therefore, some works attempt to take advantage of abundant unlabelled data and apply unsupervised learning (Peng et al. 2016; Kodirov et al. 2016; Wang et al. 2018; Lin et al. 2019; Fan et al. 2018; Deng et al. 2018; Zhong et al. 2018a). However, the existing unsupervised methods tend to be less effective and unstable due to the lack of supervised information. The main issues are: 1) the data distribution of different camera views varies significantly due to the variation of viewpoint, illumination, image resolution, and background noise; and 2) existing loss functions of person Re-ID are mostly designed for supervised learning, which cannot be applied directly due to the lack of labeled data.

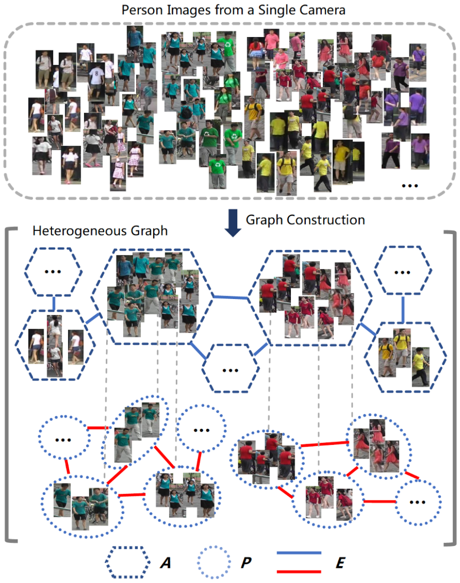

Figure 1: Illustration of the heterogeneous graph for a single camera. Person images are distributed with latent structures. We construct heterogeneous graphs in each camera-specific sub-domain to exploit the potential distribution structure. There are two types of node (coarse-grained appearance node A and fine-grained person node P) and one type of edge (distance E) in one graph for each camera. As shown in this figure, images in a coarse-grained appearance node A mostly look similar in color, while images of a fine-grained person node P come from the same person ID.

The studies in (Wei et al. 2018; Deng et al. 2018; Liu et al. 2019; Huang et al. 2020) make effort on addressing the first issue by treating person Re-ID as an unsupervised domain adaptation (UDA) problem However, these methods

[*] Equal contribution. This work was done when Kai Liu was an intern at Alibaba Group.

[†] Corresponding author.

only consider the feature distribution difference across domains while ignoring the difference in distribution structures across cameras, which leads to the performance drop of a source model in the target domain(Yang and Yuen 2019). The authors in (Qi et al. 2019) proposed a camera-aware domain adaptation to map the images of different cameras into a shared subspace. However, merely aligning the camera-level domain can not effectively improve the Re-ID performance. Without proper constraints, the distribution structure in each camera is easily corrupted, which makes the optimization harder to converge and ultimately affects the Re-ID performance.

In addition, to make up for the lack of labeled data and apply loss functions in a supervised learning manner, some domain transfer methods (Lv et al. 2018; Fan et al. 2018; Song et al. 2018; Lin et al. 2019; Fu et al. 2019; Zhao et al. 2020) utilize pseudo labels as supervised information. These methods normally apply unsupervised clustering method, such as DBSCAN (Ester et al. 1996), to group the unlabeled target dataset into independent clusters. However, compared with manually annotated labels, pseudo labels are less accurate and unstable. The performance highly depends on the clustering quality, reflecting what extent are the pseudo labels consistent with ground truth labels. Moreover, the difference in distribution structure across cameras further increases the difficulty of a perfect clustering result.

In this paper, we propose a novel coarse-to-fine heterogeneous graph alignment (HGA) method to tackle the above problems. As shown in Fig. 1, given the feature set from a backbone network, we first construct a heterogeneous graph for each camera, which consists of two types of nodes and one type of edges. In coarse-grained alignment, an adversarial training scheme is adopted to coarsely align the appearance nodes of each camera, which consequently alleviates the distribution bias between cameras. In fine-grained alignment, we introduce conservative alignment loss functions to exploit potential discriminative information in the shared space and align the person nodes of each camera with careful consideration, which generate reliable pseudo labels as learning guidance.

To summarize, our main contributions are as follows:

- We present a heterogeneous graph alignment (HGA) method to solve the unsupervised domain adaptation person Re-ID problem. By constructing and aligning the heterogeneous graph of each camera in a coarse-to-fine manner, our method significantly improves the model generalization on unlabeled target datasets.
- We propose multiple loss functions to learn a graph-aligned feature space, in which features of the same person ID are aligned and structural information of each camera's heterogeneous graph is preserved. This preservation could further help HGA mine potential discriminative information by avoiding overfitting on node alignment and inaccurate pseudo labels.
- We conduct extensive experiments and ablation studies on three standard benchmarks, which demonstrates the effectiveness and superiority of our proposed HGA method.

Related Work

Many unsupervised domain adaptive person ReID methods are proposed to exploit the full potential of abundant unlabeled person images. Most of them focus on two key issues: 1) the data distribution discrepancy between domains, and 2) the lack of label information in the target domain. Accordingly, recent studies in cross-domain person re-ID can be classified into distribution aligning methods (Wei et al. 2018; Deng et al. 2018; Wang et al. 2018; Liu et al. 2019) and clustering-based adaptation methods (Fan et al. 2018; Song et al. 2018).

Distribution aligning methods try to reduce the distribution gap between source domain and target domain in a shared feature space. Researchers (Zhong et al. 2018b,a; Deng et al. 2018; Bak, Carr, and Lalonde 2018; Zhong et al. 2019; Zhai et al. 2020; Zou et al. 2020) adopt GAN-based methods to transfer the source images into target-domain style and use generated images to train a model. However, generated images still have a large gap compared with real images. Some camera-aware domain adaptation methods (Yang et al. 2020) are developed by reducing the camera-level sub-domains divergence. CAMEL (Yu, Wu, and Zheng 2017) proposes to learn view-specific projections to deal with view-specific interference. Some researchers (Qi et al. 2019) develop a camera-aware domain adaptation to reduce the feature discrepancy across cameras. However, merely aligning the distribution across domains is insufficient since it ignores the difference in distribution bias across cameras. Instead, this paper proposes an unsupervised graph alignment method to explore both cross-domain distribution and structure information.

Clustering-based adaptation is another straight-forward approach to learn a re-ID model. To make up for the lack of labeled data, some works (Fan et al. 2018; Wu et al. 2019a; Li et al. 2019; Yu et al. 2019; Song et al. 2018; Fu et al. 2019; Zhang et al. 2019; Wang and Zhang 2020; Zhao et al. 2020) exploit unlabeled target data and adopt clustering methods to generate pseudo-labels. The basic idea is exploiting the similarity of unlabeled samples by feature clustering and generating pseudo labels for supervised information. To improve quality of pseudo labels, some works try to utilize potential relation information of data to promote the matching reliability (Yu et al. 2019; Wu et al. 2019b). However, the clustering result suffers from the data distribution discrepancy. Meanwhile, without properly constrain, the valuable structure information can be easily destroyed when the learning proceeds.

With the help of heterogeneous graph, we attempt to promote the domain adaptation and the quality of pseudo-labels at the same time. Considering the variance-bias dilemma (Geman, Bienenstock, and Doursat 1992) in the neural network based methods, we construct heterogeneous graphs with two types of nodes to represent the feature space of each camera. Using low-bias coarse-grained nodes, the proposed HGA learns camera-specific projection matrices to eliminate camera-level distribution deviation. With the low-variance fine-grained nodes, HGA learns to match the same person, meanwhile exploiting the camera-specific structural information to ensure the accuracy of the pseudo-labels. To

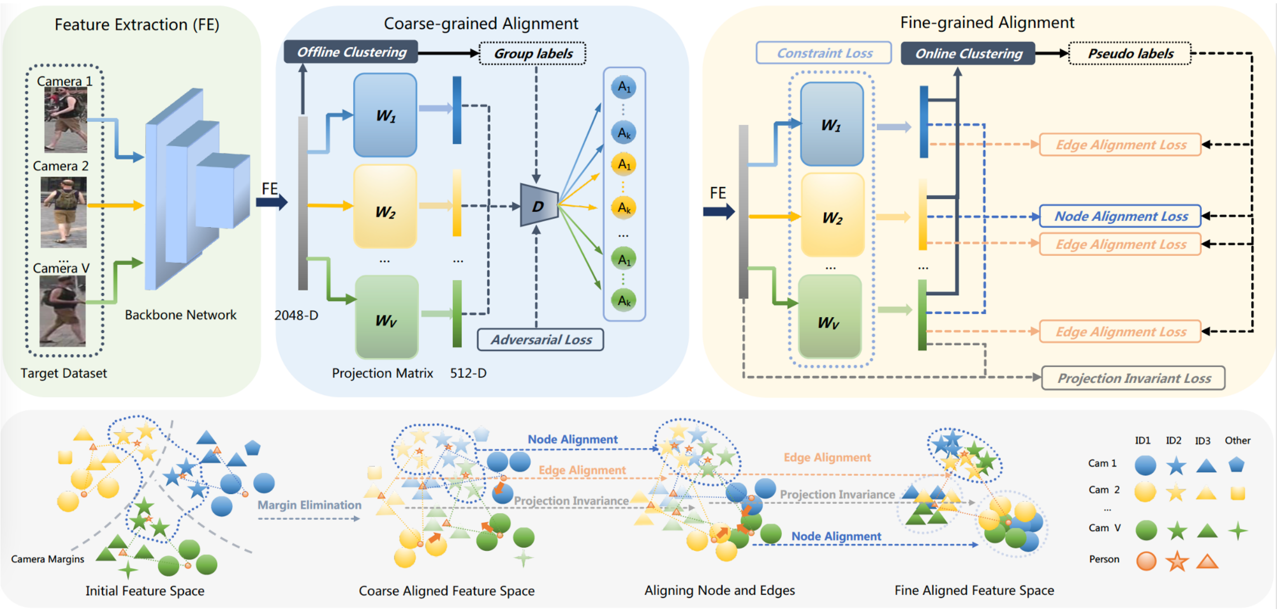

Figure 2: Illustration of our proposed framework HGA. We first extract features of images from a backbone network. Due to camera-specific variations, the initial feature space has severe camera margins: stars (ID2) from different cameras (colors) are far away from each other. We perform the coarse-to-fine HGA to learn a robust aligned feature space. In the coarse-grained alignment, an adversarial learning scheme is adopted to coarsely align the appearance nodes of each camera, which consequently alleviates the distribution bias between cameras. During fine-grained alignment, we adopt the node alignment loss to pull nodes of the same pseudo person ID closer. Also, an edge alignment loss and a projection invariant loss are used to keep the graph distribution structure of each camera unified and retained. A constraint loss is used to alleviate the inconsistency caused by projection matrices. In the aligned feature space, the camera-specific bias is alleviated, while distribution structures in the original feature space are preserved.

our best knowledge, this is the first work that addresses the person Re-ID with heterogeneous graph alignment, which helps retain and integrate the discriminative information across cameras and results in superior Re-ID performance.

Proposed Method

Problem Formulation

Suppose we have a surveillance camera network that consists of V cameras. The target domain has N unlabeled images $X = \{x_i\}_{i=1}^{N}$ in total and N_v images for v-th camera. We use a CNN backbone network ϕ to extract initial features. HGA learns camera-specific projection matrices $\{W_v\}_{v=1}^{V}$ for cameras rather than a universal one to better alleviate the effect of cross-camera divergence while retaining the discriminative information. Hereafter, we use $\sim$ to denote the representation in the shared space which is projected by corresponding $W_v (i = 1, 2, \cdots, V)$. Such as we denote $\tilde{\phi} = W_v \phi$ as an end-to-end function which contains both the backbone ϕ along with the camera-specific projection matrix W_v to map v-th camera original feature into a shared space. The goal of our method is to make full use of the hidden inherent relationship in different granularity and generate reliable labels to train and update $\tilde{\phi}$.

Heterogeneous Graph Definition. We first construct a heterogeneous graph for each camera. For example, the heterogeneous graph $G_v = (A_v, P_v, E_v)$ for v-th camera consists of coarse-grained appearance group nodes A_v, fine-grained person nodes P_v and edges E_v between nodes. To avoid cross-camera bias, we use less but larger appearance clusters to coarsely align the graph. And here we adopt k-means (Hartigan 1975) on the image set of each camera to obtain A_v which consists K appearance group nodes. To reduce cross-camera variance, we utilize HDBSCAN (McInnes, Healy, and Astels 2017) to obtain smaller and more clusters as person nodes P_v to refine the alignment. We use $\mathcal{E}(\cdot)$ as the Euclidean distance function to calculate the edge between two nodes. The overview of proposed HGA approach is shown in Fig. 2.

Coarse-grained Alignment

The task of coarse-grained alignment (CA) is to align the appearance nodes of each heterogeneous graph and reduce the distribution discrepancy at camera-level sub-domains. Similar appearance nodes from V different cameras are grouped together by solving the maximum V-dimensional matching problem(Hazan, Safra, and Schwartz 2003). Thus, similar appearance nodes will have an identical group ID. With this information, we adopt an adversarial learning method to update the projection matrices. The generator in our work is the projection matrices $\{W_v\}_{v=1}^{V}$. The discriminator is designed as a classifier D to simultaneously discriminate camera IDs and appearance group IDs. During the training process of discriminator D, we fix all generators $W_1, W_2, \cdots, W_V$.

The discriminator is optimized by a cross-entropy loss defined on the $K * V$ classes (K coarse-grained nodes for each one of V cameras) in target domain as:

$$\mathcal{L}_D(D|\{W_v\}_{v=1}^V) = \frac{1}{N}\sum_{i=1}^{N} - \log(D(W_{v_i}\phi(x_i), v_i, a_i)), \tag{1}$$

where v_i and a_i denote the camera ID and the appearance group ID that the i-th image x_i belongs to, respectively. $D(W_{v_i}\phi(x_i), v_i, a_i)$ denotes the prediction score for mapped feature $W_{v_i}\phi(x_i)$ respect to the a_i-th appearance group of the v_i-th camera class.

When training the generator W_v, we fix the weights of discriminators D. The projection of the v-th camera W_v is optimized to fool the discriminator in predicting the wrong camera ID, such as u. But, to avoid arbitrary feature distribution, D needs to keep the ability in judging the correct appearance group ID. Thus, we achieve this by minimizing the following objective function:

$$\begin{aligned}\mathcal{L}_{W_v}(W_v|D) =& \frac{1}{N_v}\sum_{i=1}^{N_v}(- \log(1 - D(W_v\phi(x_{v,i}), v, a_i)) \\ &- \frac{1}{V-1}\sum_{u \neq v}\log(D(W_v\phi(x_{v,i}), u, a_i))).\end{aligned} \tag{2}$$

The $x_{v,i}$ denotes an image in the v-th camera, a_i is the appearance group which $x_{v,i}$ belongs to. To train this coarse-grained alignment, we follow the standard generative adversarial networks (GANs) training procedure, which alternately optimizes the discriminator D and all projections $\{W_v\}_{v=1}^V$ by minimizing $\mathcal{L}_D$ and $\mathcal{L}_{W_v}$, respectively.

Fine-grained Alignment

The coarse-grained alignment step learns a series of mappings $\{W_v\}_{v=1}^V$ that coarsely align the appearance nodes of each heterogeneous graph. However, the objective of person Re-ID is to cluster the same person together and push different person far away from each other. Here we present conservative alignment loss functions which include node alignment loss, edge alignment loss, projection invariant loss and projection constraint loss. The conservative alignment loss functions are proposed with the consideration of not only aligning the same person together but also preserving the distribution structure of each heterogeneous graph, which could help the proposed approach exploit more potential cross-camera information.

Node Alignment Loss. With the coarse-aligned projections, HGA maps the data into a shared feature space where the camera bias is eliminated. Thus, we can perform person-level data alignment in a simpler way.

We utilize unsupervised clustering algorithm (McInnes, Healy, and Astels 2017) in the coarsely aligned space to obtain a undefined number of clusters as pseudo persons, so that each image can be assigned with a pseudo label according to the cluster it belongs to. Suppose there are M clusters in total, we define the projected person node $\tilde{p_i}$ by calculating the average feature of images X_i which belong to i-th cluster, $\tilde{p_i} = avg(\tilde{\phi}(X_i)), i = (1, 2, \cdots, M)$. The task of Node Alignment Loss (NAL) is to align the same person features of different camera views. Specifically, we focus on gathering the same person samples together by optimizing the distance relationship inside the 'person' node:

$$\mathcal{L}_{node} = \sum_{i=1}^{M}\sum_{v=1}^{V}\omega_{v,i}\left\|\tilde{p}_{v,i} - \tilde{p}_i\right\|_2^2. \tag{3}$$

Here $\tilde{p}_{v,i} = avg(\tilde{\phi}(X_{v,i}))$ is the average feature of images $X_{v,i}$ which belong to the i-th person and v-th camera.

To avoid overfitting on unreliable pseudo labels, we consider that the model should not have full confidence on the pseudo labels. Thus, we impose a soft constraint on the NAL to exploit the potential association information between person nodes. Specifically, $\omega_{v,i} = 1 - \frac{\mathcal{E}(\tilde{p}_{v,i}\ ,\ \tilde{p}_i)}{\sum_{u=1}^{V}\mathcal{E}(\tilde{p}_{u,i}\ ,\ \tilde{p}_i)}$ is a descending function of the distance between $\tilde{p}_{v,i}$ and $\tilde{p}_i$.

Edge Alignment Loss. As previously discussed, the distribution structure of different cameras should also be aligned to match persons across cameras in the shared space. And the distribution structural information in the camera-specific sub-domains is recorded in the edges of graphs. Therefore, the alignment of distribution structures can be modeled as the alignment between edges. The edge between the i-th and the j-th person nodes of the v-th camera is defined as $\tilde{e}_v^{i,j} = \mathcal{E}(\tilde{p}_{v,i}, \tilde{p}_{v,j}) = \left\|\tilde{p}_{v,i} - \tilde{p}_{v,j}\right\|_2$, $i, j = (1, 2, \cdots, M)$.

To align graph from different views, we unify the edges based on the following criteria: edges in a same person group are minimized, while edges between different person groups are approached to a distance ℓ. Specifically, we set $\ell = max(\frac{1}{V}\sum_{v=1}^{V}\tilde{e}_v^{i,j}, \epsilon)$ as the mean of the edges between the i-th and the j-th person node in all camera views, and ϵ is set to 0.001.

$$\mathcal{L}_{edge} = \sum_{i=j}\left\|\tilde{e}_v^{i,j}\right\|_2^2 + \sum_{i \neq j}\left\|\tilde{e}_v^{i,j} - \ell\right\|_2^2. \tag{4}$$

Projection Invariant Loss. Without proper constraints, nodes and edges can be aligned arbitrarily and the camera-specific projection matrix will be updated independently. Therefore, the structural information in each heterogeneous graph is likely to be corrupted during the learning process. To avoid arbitrary alignment and retain discriminative information in the original feature space, we propose a Projection Invariant Loss (PIL) to retain the structural information, which is defined as:

$$\mathcal{L}_{invariant} = \sum_{i \neq j}\sum_{v=1}^{V}\left\|\sigma(\tilde{e}^{i,j}, \tilde{\mu}) - \sigma(e_v^{i,j}, \mu_v)\right\|_2^2, \tag{5}$$

where $\tilde{e}^{i,j}$ is an edge between person node i and j in the shared space, and $e_v^{i,j}$ denotes the corresponding edge in the v-th camera original space without the projection matrix W_v. And $\sigma(e, \mu) = 1/(1 + exp(\mu - e))$ is a logistic-like function. $\tilde{\mu}$ and μ_v are the mean values of edges in shared space and v-th camera original space, respectively.

Algorithm 1: The proposed HGA framework.

```
Input: Unlabeled target data: $X$; Training epochs $T_c$ and
       $T_f$ for coarse-grained and fine-grained alignment,
       respectively; Number of appearance groups $K$;
       Minimal samples $S_{min}$ for HDBSCAN.
Output: Parameters of the trained network $\phi$ and all
        projection matrices $\{W_v\}_{v=1}^V$.
Pre-train the network $\phi$ on source dataset.
Extract feature set $\mathbf{F} = \phi(X)$ on data $X$.
Obtain groups for each camera by using K-means on $\mathbf{F}$.
Establish groups matching across $V$ cameras by solving the
  maximum $V$-dimensional matching.
for Epoch = 1, ..., $T_c$ do
    Update discriminator $D$ by minimizing Eq.(1)
    Update projections $\{W_v\}_{v=1}^V$ by minimizing Eq.(2).
for Epoch = 1, ..., $T_f$ do
    Extract feature set $\mathbf{F} = \phi(X)$ on data $X$.
    Map feature set $F$ thorough camera-specific projection
      matrix $W$ to obtain projection feature set $\tilde{\mathbf{F}} = W\mathbf{F}$.
    Obtain pseudo labels using HDBSCAN($\tilde{\mathbf{F}}$, $S_{min}$).
    Update the network $\phi$ and projection matrices
      $\{W_v\}_{v=1}^V$ by minimizing $\mathcal{L}_{total}$ Eq.(7).
```

Constrained by PIL, the structural information of each camera in the original space is preserved when features of each camera are projected into a shared space.

Projection Constraint Loss. Since the transformations are with respect to person images from different cameras, they are inherently correlated and homogeneous. Therefore, we adopt a cross-camera consistency term to balance between the ability to capture discriminative information and the capability to alleviate view-specific bias. We also propose a soft-orthogonal constraint to maintain the transformation matrix close to an orthogonal matrix as training proceeds, which helps to preserve the individual characteristic and stabilize the learning process. Together, the Projection Constraint Loss (PCL) is defined as:

$$\mathcal{L}_{constraint} = \sum_{u,v} \left\| W_u - W_v \right\|_F^2 + \sum_{v=1}^{V} \left\| W_v^T W_v - I \right\|_F^2, \tag{6}$$

where W_v is the specific transformation for the v-th camera view. $W_v^T W_v$ is a convariance matrix, and I represents the identity matrix.

Finally, the graph-aligned feature representations are obtained by minimizing equation Eq.(3), Eq.(4), Eq.(5) and Eq.(6) jointly. The total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{node} + \mathcal{L}_{edge} + \mathcal{L}_{invariant} + \eta \mathcal{L}_{constraint}, \tag{7}$$

where η is a trade-off parameter, which is set to 0.1 according to the best evaluation results. Alg. 1 concludes the proposed learning method.

Experiments

In this section, we conduct sufficient ablation studies to prove the effectiveness of each component in HGA. Then, we compare the performance of proposed HGA with other state-of-the-art unsupervised domain adaptation person re-ID methods to show superiority.

Methods	Duke→ Market		Market→ Duke	
	mAP	R1	mAP	R1
Same	19.5	39.7	13.4	28.2
Random	21.4	42.8	17.7	34.9
CAMEL	25.2	49.1	20.8	43.8
CA	**31.5**	**54.1**	**29.2**	**49.6**
Same+FA	23.5	52.5	22.9	41.5
Random+FA	32.8	55.7	28.9	38.8
CAMEL+FA	42.3	61.7	39.3	58.8
HGA(CA+FA)	**70.3**	**89.5**	**67.1**	**79.4**

Table 1: The comparison of different initialization schemes for W. Same denotes setting all W into a same matrix. Random means each W_i is randomly initialized into different matrices. CAMEL denotes the initialization method from (Yu, Wu, and Zheng 2017), we re-implement their work and conduct experiments in the same settings. Hereafter, Duke→Market represents that we use Duke as source domain and Market as target domain and vice versa.

Datasets and Evaluation Metrics

We evaluate our method on three person re-ID benchmark datasets, $i.e.$ Market1501 (Zheng et al. 2015), DukeMTMC-ReID (Ristani et al. 2016; Zheng, Zheng, and Yang 2017) and MSMT17 (Wei et al. 2018), which are considered as large scale in the community. Performance is evaluated by the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP).

Implementation Details

We adopt ResNet-50 (He et al. 2016) without the last classification layer as backbone network to conduct all experiments. The backbone is pretrained on ImageNet (Deng et al. 2009) and then further trained on the source dataset. All codes are implemented on Pytorch. During training, all images are resized to 256×128. Random flipping, random cropping and random erasing are used for data augmentation. We adopt stochastic gradient descent with a momentum of 0.9 to optimize the model. As shown in Alg. 1, the training process lasts for 100 epochs, including 20 epochs for coarse-grained alignment and another 80 epochs for fine-grained alignment. The learning rate is initialized to 0.01 and divided by 10 for every 40 epochs. We set the batch size equal to 128 for both training and testing. To guarantee that each batch contains images from all views, we first compute the distribution of the numbers of samples in each camera, and then compose a batch of images from all cameras with corresponding proportions. The shape of each projection matrix is 512 × 2048. For testing, we use 512-d projected features as final representations of testing images to compute their similarity.

Ablation Study

The Impact of Coarse-grained Alignment. To evaluate the effectiveness and stability of proposed coarse-

Methods	Duke→ Market		Market→ Duke	
	mAP	R1	mAP	R1
CA	31.5	54.1	29.2	49.6
CA+NAL	55.4	73.6	49.8	62.4
CA+EAL	59.1	78.1	54.7	66.8
CA+NAL+EAL	63.5	81.5	61.4	74.6
CA+NAL+EAL+PIL	69.5	87.1	66.2	79.1
CA+FA	**70.3**	**89.5**	**67.1**	**79.4**

Table 2: Ablation studies of different loss functions in Fine-grained Alignment.

grained alignment (CA), we compare our adversarial learning scheme with various initialization strategies of projection matrices. Specifically, we implement three initialization schemes for W including Same, Random and CAMEL.

As shown in Table 1, CA clearly outperforms other three strategies when tested on Duke→Market and Market→Duke. This demonstrates the effectiveness of the proposed adversarial learning scheme for coarse-grained alignment. Moreover, when the fine-grained alignment (FA) is added to these methods, the proposed HGA further outperforms other methods by a larger margin. Specifically, with fine-grained alignment, CAMEL only improves mAP by +17.1%(+18.5%) on Duke→Market(Market→ Duke), meanwhile the HGA shows a significant performance gain on mAP by +38.8% (+37.9%) on Market(DukeMTMC). This indicates that CA has its own superiority in aligning appearance node of different cameras and promoting the subsequent fine-grained alignment, and also proves that the adversarial learning of each W_v has successfully made the projected features indistinguishable for discriminator D to decide which camera it comes from.

The Impact of Losses in Fine-grained Alignment. The effect of loss functions in FA is listed in Table 2. Firstly, only with the proposed NAL, we improve the performance by 19.5% and 12.8% at Rank-1 accuracy compared with the results from CA when tested on Duke→Market and Market→Duke, respectively. Secondly, we observe that containing only the proposed EAL, the mAP and Rank-1 accuracy increase by 27.6% and 24.0% for Duke→Market, while 25.5% and 17.2% for Market→Duke. This improvement demonstrates that both NAL and EAL are beneficial to model generalization. Thirdly, we combine NAL and EAL together to jointly optimize model. It is clear that we achieve better results on both Duke→Market and Market→Duke than using each loss function alone. This shows that NAL and EAL are complementary and to each other in fine-grained graph alignment. Then, we continue employing PIL and further improving mAP and Rank-1 by 6.0% and 5.6% for Market→Duke, and 4.8% and 4.5% for Market→Duke. Note that containing only PIL could not help much in improving performance. This indicates that PIL is effective to keep the person graph structures untouched in each appearance node and useful when the nodes and edges of person graphs are aligned by NAL and PIL. Finally, with the constraints of PCL, the proposed model (CA+FA) further gains a performance boost by 2.4% and 0.3% on Rank-1 for Duke→Market and Market→Duke, respectively.

Comparison with State-of-the-art Methods

Results on Market1501. Table 3 reports comparisons on the task of Duke→Market. BOW(Zheng et al. 2015) and LOMO(Liao et al. 2015) directly apply hand-crafted features to evaluate re-ID performance, and both methods have poor performance due to the lack of training. Compared to GAN-based re-ID methods, the proposed HGA treats adversarial learning in a simpler way and only demands a coarse graph alignment. HGA achieves mAP = 70.3% and Rank-1 = 89.5%, which significantly exceeds the GAN-based re-ID methods by a large margin. This indicates the proposed method can make use of the unlabeled data more effectively. Note that CAMEL and UCDA-CCE (Qi et al. 2019) also treat camera-level sub-domains as a unique characteristic, thus share certain similarity to our work. However, our method significantly outperforms them, which demonstrates that the proposed HGA has its own superiority in exploiting the intrinsic distinctions among identities and indeed alleviates camera-specific bias through aligning nodes and edges of the same person in the heterogeneous graph.

Results on DukeMTMC-reID. The similar improvement can also be observed when we test our method on the task of Market→Duke. As shown in Table 3, our proposed HGA achieves mAP = 67.1% and rank-1 accuracy = 80.4%, which is superior to all previous UDA methods. Specifically, compared to the best UDA method DG-Net++(Zou et al. 2020), the proposed method is 3.3% and 1.5% higher on mAP and Rank-1 accuracy. Note that NRMT(Zhao et al. 2020) performs better on Duke→Market task. But we outperform them on Market→Duke task by a large margin and also have a lead in Rank-1 accuracy on Duke→Market task, which proves our method has a better performance consistency on both two tasks. Therefore, the advantage of the proposed HGA domian adaptation approach for person re-ID can be confirmed.

Results on MSMT17. In addition, we further evaluate the proposed HGA approach on MSMT17 dataset, which is larger and more challenging. As shown in Table 4, the proposed method clearly outperforms five existing UDA methods including PTGAN(Wei et al. 2018), ECN(Zhong et al. 2019), SSG (Fu et al. 2019), MMCL (Wang and Zhang 2020), DAAM (Huang et al. 2020), NRMT (Zhao et al. 2020), DG-Net++ (Zou et al. 2020). Our proposed HGA achieves 25.5% and 55.1% in terms of mAP and Rank-1 accuracy when the model is trained on Market-1501, which outperforms previously best method DG-Net++ by 6.7% in Rank-1 accuracy. This further verifies the effectiveness and generalization of our proposed method.

Parameter Analysis

The Number of Appearance Groups. We evaluate the learned model when K is set to 3,4,...,10 in 12 respectively, and the results are shown in Fig. 3. With K increasing in a range from 3 to 5, the improvement is increasingly significant. And we also observe that the proposed method is very stable and does not fluctuate greatly when K is greater than 5. The best result is obtained when K is set to 6. This shows

Methods	Reference	DukeMTMC-reID→ Market-1501				Market-1501→ DukeMTMC-reID			
		mAP	R1	R5	R10	mAP	R1	R5	R10
LOMO	CVPR 15	8.0	27.2	41.6	49.1	4.8	12.3	21.3	26.6
Bow	ICCV 15	14.8	35.8	52.4	60.3	8.3	17.1	28.8	34.9
SPGAN	CVPR 18	22.8	51.5	70.1	76.8	22.3	41.1	56.6	63.0
PTGAN	CVPR 18	-	38.6	-	66.1	-	27.4	-	50.7
UCDA-CCE	ICCV 19	34.5	64.3	-	-	36.7	55.4	-	-
ECN	CVPR 19	43.0	75.1	87.6	91.6	40.4	63.3	75.8	80.4
CR-GAN	ICCV 19	54.0	77.7	89.7	92.7	48.6	68.9	80.2	84.7
PUL	TOMM 18	20.5	45.5	60.7	66.7	16.4	30.0	43.4	48.5
CAMEL	ICCV 17	26.3	54.5	-	-	-	-	-	-
TJ-AIDL	CVPR 18	26.5	58.2	74.8	81.1	23.0	44.3	59.6	65.0
CASCL	ICCV 19	35.6	64.7	80.2	85.6	30.5	51.5	66.7	71.7
MAR	CVPR 19	40.0	67.7	81.9	-	48.0	67.1	79.8	-
PDA-Net	ICCV 19	47.6	75.2	86.3	90.2	45.1	63.2	77.0	82.5
PAST	ICCV 19	54.6	78.4	-	-	54.3	72.4	-	-
SSG	ICCV 19	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
MMCL	CVPR 20	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0
AD-Cluster	CVPR 20	68.3	86.7	94.4	96.5	54.1	72.6	82.5	85.5
DAAM	AAAI 20	67.8	86.4	-	-	63.9	77.6	-	-
DG-Net++	ECCV 20	61.7	82.1	90.2	92.7	63.8	78.9	87.8	90.3
NRMT	ECCV 20	**71.7**	87.8	**94.6**	**96.5**	62.2	77.8	86.9	89.5
HGA	This Paper	70.3	**89.5**	93.6	95.5	**67.1**	**80.4**	**88.7**	**90.3**

Table 3: Comparison of proposed HGA approach with state-of-arts unsupervised domain adaptation person Re-ID methods on Market-1501 and DukeMTMC-re-ID dataset.

Methods	Market-1501→ MSMT17			
	mAP	R1	R5	R10
PTGAN	2.9	10.2	-	24.4
ECN	8.5	25.3	36.3	42.1
SSG	13.2	31.6	-	49.6
MMCL	15.1	40.8	51.8	56.7
DAAM	20.8	44.5	-	-
NRMT	19.8	43.7	56.5	62.2
DG-Net++	22.1	48.4	60.9	**66.1**
HGA	**25.5**	**55.1**	**61.2**	65.5
Methods	**DukeMTMC-reID→ MSMT17**			
	mAP	R1	R5	R10
PTGAN	3.3	11.8	-	27.4
ECN	10.2	30.2	41.5	46.8
SSG	13.3	32.2	-	51.2
MMCL	16.2	43.6	54.3	58.9
DAAM	21.6	46.7	-	-
NRMT	20.6	45.2	57.8	63.3
DG-Net++	22.1	48.8	60.9	65.9
HGA	**26.8**	**58.6**	**64.7**	**69.2**

Table 4: Comparison of proposed HGA approach with state-of-arts unsupervised domain adaptive person Re-ID methods on MSMT17 dataset.

that appearance groups do exist in the feature space and our intuition on constructing a heterogeneous graph for person re-ID is valid.

The Parameters of HDBSCAN. In addition, we analyse how the number of minimum samples (S_{min}) for each cluster in HDBSCAN clustering affects the Re-ID results. We test the impact of 5, 10, 15, 20, 25 minimum samples on the performance of our HGA framework for Duke→Market task. As shown in Fig. 3, we can see that setting S_{min} to 15 yields superior accuracy. Meanwhile, different S_{min} has large impact on the Re-ID accuracy. We believe that the best setting of S_{min} is dependent on the true data distribution of the target dataset. For example, the average number of images for each person in Market-1501 is near 16, which is almost consistent with our experimental results.

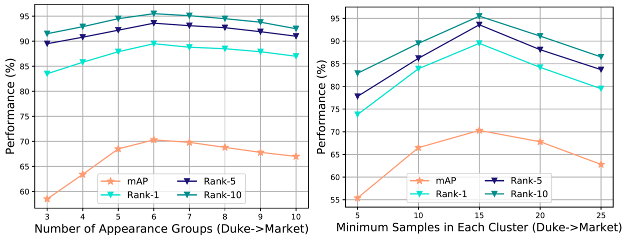

Figure 3: Analysis of hyper parameters on Duke→Market task. Left: The impact of K in K-means clustering. Right: The impact of the minimum samples S_{min} at each cluster in HDBSCAN clustering.

Conclusion

In this work, we propose Heterogeneous Graph Alignment (HGA), which can exploit discriminative information by constructing and aligning the heterogeneous graph of each camera, to tackle the challenging unsupervised domain adaptation person Re-ID. With the proposed coarse-to-fine learning scheme, HGA achieves the graph alignment in different granularity, resulting in aligning features of the same person and preserving the distribution structure of each graph. This preservation can further help HGA to find more potential discriminative information. Extensive experiments demonstrate that the performance of our approach outperforms the state-of-the-arts.

Acknowledgements

This work is supported in part by the Fundamental Research Funds for the Central Universities of China under Grant 2019YJS032 and in part by the National Natural Science Foundation of China under Grant U1934220, 6167208.

References

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3908–3916.

Bak, S.; Carr, P.; and Lalonde, J.-F. 2018. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 189–205.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. 248–255.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. 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*, 994–1003.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*.

Fan, H.; Zheng, L.; Yan, C.; and Yang, Y. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14(4): 1–18.

Fu, Y.; Wei, Y.; Wang, G.; Zhou, X.; Shi, H.; and Huang, T. S. 2019. Self-similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-identification.

Geman, S.; Bienenstock, E.; and Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural computation* 4(1): 1–58.

Hartigan, J. A. 1975. *Clustering algorithms*. John Wiley & Sons, Inc.

Hazan, E.; Safra, S.; and Schwartz, O. 2003. On the complexity of approximating k-dimensional matching. In *Approximation, Randomization, and Combinatorial Optimization.. Algorithms and Techniques*, 83–97. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. 770–778.

Huang, Y.; Peng, P.; Jin, Y.; Li, Y.; Xing, J.; and Ge, S. 2020. Domain Adaptive Attention Learning for Unsupervised Person Re-Identification. In *AAAI*, 11069–11076.

Kodirov, E.; Xiang, T.; Fu, Z.; and Gong, S. 2016. Person Re-Identification by Unsupervised ℓ_1 Graph Learning. In *European conference on computer vision*, 178–195. Springer.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 152–159.

Li, Y.-J.; Lin, C.-S.; Lin, Y.-B.; and Wang, Y.-C. F. 2019. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 7919–7929.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. 2197–2206.

Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; and Yang, Y. 2019. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8738–8745.

Liu, J.; Zha, Z.-J.; Chen, D.; Hong, R.; and Wang, M. 2019. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7202–7211.

Lv, J.; Chen, W.; Li, Q.; and Yang, C. 2018. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7948–7956.

McInnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2(11): 205.

Peng, P.; Xiang, T.; Wang, Y.; Pontil, M.; Gong, S.; Huang, T.; and Tian, Y. 2016. Unsupervised cross-dataset transfer learning for person re-identification. 1306–1315.

Qi, L.; Wang, L.; Huo, J.; Zhou, L.; Shi, Y.; and Gao, Y. 2019. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 8080–8089.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking.

Song, L.; Wang, C.; Zhang, L.; Du, B.; Zhang, Q.; Huang, C.; and Wang, X. 2018. Unsupervised Domain Adaptive Re-Identification: Theory and Practice. *arXiv preprint arXiv:1807.11334* .

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2017. Beyond Part Models: Person Retrieval with Refined Part Pooling. *arXiv preprint arXiv:1711.09349* .

Wang, D.; and Zhang, S. 2020. Unsupervised Person Re-identification via Multi-label Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10981–10990.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2275–2284.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer GAN to bridge domain gap for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, J.; Liao, S.; Wang, X.; Yang, Y.; Li, S. Z.; et al. 2019a. Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 886–891. IEEE.

Wu, J.; Yang, Y.; Liu, H.; Liao, S.; Lei, Z.; and Li, S. Z. 2019b. Unsupervised graph association for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 8321–8330.

Yang, B.; and Yuen, P. C. 2019. Cross-Domain Visual Representations via Unsupervised Graph Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5613–5620.

Yang, F.; Li, K.; Zhong, Z.; Luo, Z.; Sun, X.; Cheng, H.; Guo, X.; Huang, F.; Ji, R.; and Li, S. 2020. Asymmetric Co-Teaching for Unsupervised Cross-Domain Person Re-Identification. In *AAAI*, 12597–12604.

Yu, H.-X.; Wu, A.; and Zheng, W.-S. 2017. Cross-view asymmetric metric learning for unsupervised person re-identification.

Yu, H.-X.; Zheng, W.-S.; Wu, A.; Guo, X.; Gong, S.; and Lai, J.-H. 2019. Unsupervised Person Re-identification by Soft Multilabel Learning.

Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; and Tian, Y. 2020. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9021–9030.

Zhang, X.; Cao, J.; Shen, C.; and You, M. 2019. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 8222–8231.

Zhao, F.; Liao, S.; Xie, G.-S.; Zhao, J.; Zhang, K.; and Shao, L. 2020. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *European Conference on Computer Vision*, 526–544. Springer.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. 1116–1124.

Zheng, W.-S.; Gong, S.; and Xiang, T. 2012. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence* 35(3): 653–668.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717* .

Zhong, Z.; Zheng, L.; Li, S.; and Yang, Y. 2018a. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–188.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification.

Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; and Yang, Y. 2018b. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing* 28(3): 1176–1190.

Zou, Y.; Yang, X.; Yu, Z.; Kumar, B.; and Kautz, J. 2020. Joint disentangling and adaptation for cross-domain person re-identification. *arXiv preprint arXiv:2007.10315* .