Deep Constrained Dominant Sets for Person Re-Identification

Leulseged Tesfaye Alemu[1] Marcello Pelillo [1, 2] Mubarak Shah [3]
[1] Ca' Foscari University of Venice [2]ECLT, Venezia [3]CRCV, University of Central Florida
{leulseged.alemu, pelillo}@unive.it, shah@crcv.ucf.edu

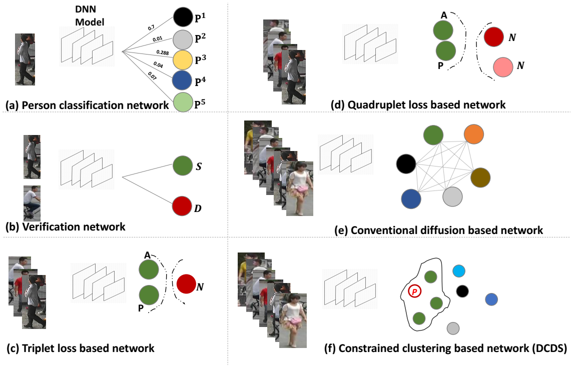
Abstract

In this work, we propose an end-to-end constrained clustering scheme to tackle the person re-identification (re-id) problem. Deep neural networks (DNN) have recently proven to be effective on person re-identification task. In particular, rather than leveraging solely a probe-gallery similarity, diffusing the similarities among the gallery images in an end-to-end manner has proven to be effective in yielding a robust probe-gallery affinity. However, existing methods do not apply probe image as a constraint, and are prone to noise propagation during the similarity diffusion process. To overcome this, we propose an intriguing scheme which treats person-image retrieval problem as a constrained clustering optimization *problem, called deep constrained dominant sets (DCDS). Given a probe and gallery images, we re-formulate person re-id problem as finding a constrained cluster, where the probe image is taken as a constraint (seed) and each cluster corresponds to a set of images corresponding to the same person. By optimizing the constrained clustering in an end-to-end manner, we naturally leverage the contextual knowledge of a set of images corresponding to the given person-images. We further enhance the performance by integrating an auxiliary net alongside DCDS, which employs a multi-scale ResNet. To validate the effectiveness of our method we present experiments on several benchmark datasets and show that the proposed method can outperform state-of-the-art methods.*

1. Introduction

Person re-identification aims at retrieving the most similar images to the probe image, from a large-scale gallery set captured by camera networks. Among the challenges which hinder person re-id tasks, include background clutter, pose, viewpoint and illumination variation can be mentioned.

Person re-id can be considered as a person retrieval problem based on the ranked similarity score, which is obtained from the pairwise affinities between the probe and the dataset images. However, relying solely on the pairwise affinities of probe-gallery images, ignoring the underlying contextual information between the gallery images often leads to an undesirable similarity ranking. To tackle this, several works have been reported, which employ similarity diffusion to estimate a second order similarity in order to capture the intrinsic manifold structure of the given affinity matrix [3], [18], [11], [4]. Similarity diffusion is a process of exploiting the contextual information between all the gallery images to provide a context sensitive similarity. Nevertheless, all these methods do not leverage the advantage of deep neural networks. Instead, they employ the similarity diffusion process as a post-processing step on the top of the DNN model. Only recently, some works have incorporated a similarity diffusion process in an end-to-end manner [20],[21],[7] in order to improve the discriminative power of a DNN model. Following [5], which applies a random walk in an end-to-end fashion for solving semantic segmentation problem, authors in [20] proposed a group-shuffling random walk network for fully utilizing the affinity information between gallery images in both the training and testing phases. Also, the authors of [21] pro-



(a) Person classification network
(d) Quadruplet loss based network
(b) Verification network
(e) Conventional diffusion based network
(c) Triplet loss based network
(f) Constrained clustering based network (DCDS)

Figure 1. Shows a variety of existing classification and similarity-based deep person re-id models. (a) Depicts a classification-based deep person re-id model, where P^i refers to the i^{th} person. (b) Illustrates a verification network whereby the similarity S and dissimilarity D for a pair of images is found. (c) A Triplet loss based DNN, where A, P, N indicate anchor, positive, and negative samples, respectively. (d) A quadruplet based DNN (e) Conventional diffusion-based DNN, which leverages the similarities among all the images in the gallery to learn a better similarity. (f) The proposed deep constrained dominant sets (DCDS), where, P indicates the constraint (probe-image); and, images in the constrained cluster, the enclosed area, indicates the positive samples to the probe image.

posed similarity-guided graph neural network (SGGNN) to exploit the relationship between several prob-gallery image similarities.

However, most of the existing graph-based end-to-end learning methods apply the similarity diffusion without considering any constraint or attention mechanism to the specific query image. Due to that the second order similarity these methods yield is highly prone to noise. To tackle this problem, one possible mechanism could be to guide the similarity propagation by providing seed (or constraint) and let the optimization process estimate the optimal similarity between the seed and nearest neighbors, while treating the seed as our attention point. To formalize this idea, in this paper, we model person re-id problem as finding *an internally coherent* and *externally incoherent* constrained cluster in an end-to-end fashion. To this end, we adopt a graph and game theoretic method called constrained dominant sets in an end-to-end manner. To the best of our knowledge, we are the first ones to integrate the well known unsupervised clustering method called dominant sets in a DNN model. To summarize, the contributions of the proposed work are: 1) for the very first time, the dominant sets clustering method is integrated in a DNN and optimized in end-to-end fashion. 2) a one-to-one correspondence between person re-identification and constrained clustering problem is established. 3) state-of-the-art results are significantly improved.

The paper is structured as follow. In section 2, we review the related works. In section 3, we discuss the proposed method with a brief introduction to dominant sets and constrained dominant sets. Finally, in section 4, we provide an extensive experimental analysis on three different benchmark datasets.

2. Related Works

Person re-id is one of the challenging computer vision tasks due to the variation of illumination condition, backgrounds, pose and viewpoints. Most recent methods train DNN models with different learning objectives including verification, classification, and similarity learning [9], [36], [26], [1]. For instance, verification network (V-Net) [16], Figure 1(b), applies a binary classification of image-pair representation which is trained under the supervision of binary softmax loss. Learning accurate similarity and robust feature embedding has a vital role in the course of person re-identification process. Methods which integrate siamese network with contrastive loss are a typical showcase of deep similarity learning for person re-id [8]. The optimization goal of these models is to estimate the minimum distance between the same person images, while maximizing the distance between images of different persons. However, these methods focus on the pairwise distance ignoring the contextual or relative distances. Different schemes have tried to overcome these shortcomings. In Figure 1(c), triplet loss is exploited to enforce the correct order of relative distances among image triplets [9], [10], [36] . In Figure 1(d), Quadruplet loss [8] leverages the advantage of both contrastive and triplet loss, thus it is able to maximize the intra-class similarity while minimizing the inter-class similarity. Emphasizing the fact that these methods entirely neglect the global structure of the embedding space, [7], [20], [21] proposed graph based end-to-end diffusion methods shown in Figure 1(e).

Graph based end-to-end learning. Graph-based methods have played a vital role in the rapid growth of computer vision applications in the past. However, lately, the advent of deep convolutional neural networks and their tremendous achievements in the field has attracted great attention of researchers. Accordingly, researchers have made a significant effort to integrate, classical methods, in particular, graph theoretical methods, in end-to-end learning. Shen *et al.* [21] developed two constructions of deep convolutional networks on a graph, the first one is based upon hierarchical clustering of the domain, and the other one is based on the spectrum of graph Laplacian. Yan *et al.* [31] proposed a model of dynamic skeletons called Spatial-Temporal Graph Convolutional Networks (ST-GCN), which provides a capability to automatically learn both the spatial and temporal pattern of data. Bertasius *et al.* [5] designed a convolutional random walk (RWN), where by jointly optimizing the objective of pixelwise affinity and semantic segmentation they are able to address the problem of blobby boundary and spatially fragmented predictions. Likewise, [20] integrates random walk method in end-to-end learning to tackle person re-identification problem. In [20], through the proposed deep random walk and the complementary feature grouping and group shuffling scheme, the authors demonstrate that one can estimate a robust probe-gallery affinity. Unlike recent Graph neural network (GNN) methods [21], [15], [20], [7], Shen *et al.* [21] learn the edge weights by exploiting the training label supervision, thus they are able to learn more accurate feature fusion weights for updating node features.

Recent applications of dominant sets. Dominant sets (DS) clustering [19] and its constraint variant constrained dominant sets (CDS) [34] have been employed in several recent computer vision applications ranging from person tracking [24], [25], geo-localization [35], image retrieval [32], [2], 3D object recognition [27], to Image segmentation and co-segmentation [33]. Zemene *et al.* [34] presented CDS with its applications to interactive Image segmentation. Subsequently, authors in [33] use CDS to tackle both image segmentation and co-segmentation in interactive and unsupervised setup. Wang *et al.* [27] recently used dominant sets clustering in a recursive manner to select representative images from a collection of images and applied a pooling operation on the refined images, which survive at the recursive selection process. Nevertheless, *none of the* *above works have attempted to leverage the dominant sets algorithm in an end-to-end manner.*

In this work, unlike most of the existing graph-based DNN models, we propose a constrained clustering based scheme in an end-to-end fashion, thereby, leveraging the contextual information hidden in the relationship among person images. In addition, the proposed scheme significantly magnifies the inter-class variation of different person-images, while reducing the intra-class variation of the same person-images. The big picture of our proposed method is depicted in Figure 1(f), as can be seen, the objective here is to find a coherent constrained cluster which incorporates the given probe image P.

3. Our Approach

In this work, we cast probe-gallery matching as optimizing a constrained clustering problem, where the probe image is treated as a constraint, while the positive images to the probe are taken as members of the constrained-cluster. Thereby, we integrate such clustering mechanism into a deep CNN to learn a robust features through the leveraged contextual information. This is achieved by traversing through the global structure of the given graph to induce a compact set of images based on the given initial similarity(edge-weight).

3.1. Dominant Sets and Constrained Dominant Sets

Dominant sets is a graph theoretic notion of a cluster, which generalizes the concept of a maximal clique to edge-weighted graphs. First, the data to be clustered are represented as an undirected edge-weighted graph with no self-loops, $G = (V, E, w)$, where $V = \{1, ..., M\}$ is the vertex set, $E \subseteq V \times V$ is the edge set, and $w : E \rightarrow R_+^*$ is the (positive) weight function. Vertices in G correspond to data points, edges represent neighborhood relationships, and edge-weights reflect similarity between pairs of linked vertices. As customary, we represent the graph G with the corresponding weighted adjacency (or similarity) matrix, which is the $M \times M$ nonnegative, symmetric matrix $A = (a_{ij})$, defined as $a_{ij} = w(i, j)$, if $(i, j) \in E$, and $a_{ij} = 0$ otherwise. Note that the diagonal elements of the adjacency matrix A are always set to zero indicating that there is no self-loops in graph G. As proved in [19], one can extract a coherent cluster from a given graph by solving a quadratic program $f(\mathbf{x})$ as,

$$\begin{array}{ll} \text{maximize} & f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}, \\ \text{subject to} & \mathbf{x} \in \Delta \end{array} \tag{1}$$

where, Δ is the standard simplex of R^n. Zemene et. al [34] proposed an extension of dominant sets, which allows one to constrain the clustering process to contain intended constraint nodes P. Constrained dominant set (CDS) is an extension of dominant set which contains a parameterized regularization term that controls the global shape of the energy landscape. When the regularization parameter is zero the local solutions are known to be in one-to-one correspondence with the dominant sets. A compact constrained cluster could be easily obtained from a given graph by defining a paramertized quadratic program as,

$$\begin{array}{ll} \text{maximize} & f_P^\alpha(X) = \mathbf{x}^T(A - \alpha \hat{I}_P)\mathbf{x}, \\ \text{subject to} & \mathbf{x} \in \Delta \end{array} \tag{2}$$

where, $\hat{I}_P$ refers to $M \times M$ diagonal matrix whose diagonal elements are set to zero in correspondence to the probe P and to 1 otherwise. Let $\alpha > \lambda_{max}(A_{V \backslash P})$, where $\lambda_{max}(A_{V \backslash P})$ is the largest eigenvalue of the principal submatrix of A indexed by the elements of $V \backslash P$. If $\mathbf{x}$ is a local maximizer of $f_P^\alpha(\mathbf{x})$ in Δ, then $\delta(\mathbf{x}) \cap P \neq \emptyset$, where, $\delta(\mathbf{x}) = i \in V : \mathbf{x}_i > 0$. We refer the reader to [34] for the proof. Equations 1 and 2 can be simply solved with a straightforward continuous optimization technique from evolutionary game theory called replicator dynamics, as follows:

$$x_i(t+1) = x_i(t) \frac{(A\mathbf{x}(t))_i}{\mathbf{x}(t)^T A\mathbf{x}(t)}. \tag{3}$$

3.2. Modeling Person Re-Id as a Dominant Set

Recent methods [7], [5] have proposed different models, which leverage local and group similarity of images in an end-to-end manner. Authors in [7] define a group similarity, which emphasizes the advantages of estimating a similarity of two images, by employing the dependencies among the whole set of images in a given group. In this work, we establish a natural connection between finding a robust probe-gallery similarity and constrained dominant sets. Let us first elaborate the intuitive concept of finding a coherent subset from a given set based on the global similarity of given images. For simplicity, we represent person-images as vertices of graph G, and their similarity as edge-weight w_{ij}. Given vertices V, and let $S \subseteq V$ be a non-empty subset of vertices and $i \in S$, average weighted degree of each i with regard to S is given as

$$\phi_S(i, j) = a_{ij} - \frac{1}{|S|} \sum_{k \in S} a_{ik} \ ,$$

where $\phi_S(i, j)$ measures the (relative) similarity between node j and i, with respect to the average similarity between node i and its neighbors in S. Note that $\phi_S(i, j)$ can be either positive or negative. Next, to each vertex $i \in S$ we assign a weight defined (recursively) as follows:

$$W_S(i) = \begin{cases} 1, & \text{if} \quad |S| = 1, \\ \sum_{j \in S \setminus \{i\}} \phi_{S \setminus \{i\}}(j, i) W_{S \setminus \{i\}}(j), & \text{otherwise} \end{cases} \tag{4}$$

where $W_{\{ij\}}(i) = W_{\{ij\}}(j) = a_{ij}$ for all $i, j \in V (i \neq j)$. Intuitively, $W_S(i)$ gives us a measure of the overall similarity between vertex i and the vertices of $S \setminus \{i\}$, with respect to the overall similarity among the vertices in $S \setminus \{i\}$.

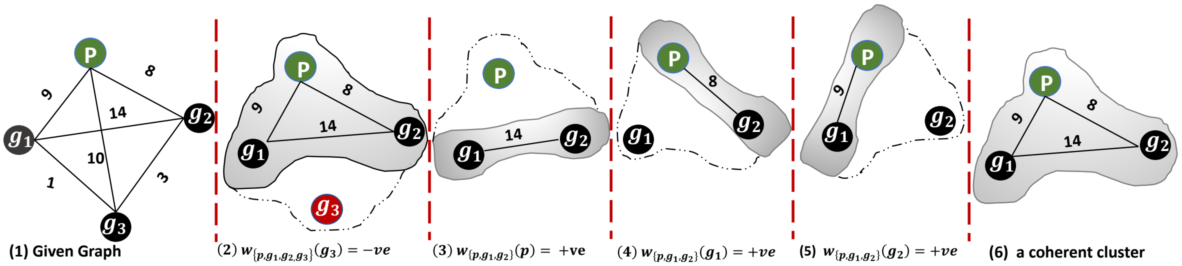

P
g_1
g_2
g_3
9
8
14
10
1
3
(1) Given Graph
(2) $w_{\{p,g_1,g_2,g_3\}}(g_3) = -ve$
(3) $w_{\{p,g_1,g_2\}}(p) = +ve$
(4) $w_{\{p,g_1,g_2\}}(g_1) = +ve$
(5) $w_{\{p,g_1,g_2\}}(g_2) = +ve$
(6) a coherent cluster

Figure 2. Let $S = \{P, g_1, g_2, g_3\}$ comprises probe, P, and gallery images g_i. As can be observed from the above toy example, the proposed method assess the contribution of each participant node $i \in S$ with respect to the subset $S \backslash i$. (1) shows graph G, showing the pairwise similarities of query-gallery images. (2-5) show the relative weight, $W_{\Gamma\}}(i)$ (Equ. 4), of each node with respect to the overall similarity between set $\Gamma \backslash i$ (shaded region) and i. (2) shows that if the Node $\{g_3\}$ is added with Node $\{P, g_1, g_2\}$ it has a negative impact on the coherency of the cluster, since $W_{p,g_1,g_2,g_3}(g_3) < 0$. (3) shows that clustering $\{P\}$ with $\{g_1\}$ and $\{g_2\}$ has a positive contribution to the compactness of set $\{P, g_1, g_2\}$. (4), similarly, shows the relative weight of g_1, $W_{p,g_1,g_2}(g_1) > 0$. (5) shows the relative weight of $g_2, W_{p,g_1,g_2}(g_2) > 0$. And, (6) is a coherent subset (dominant set cluster) extracted from the graph given in (1).

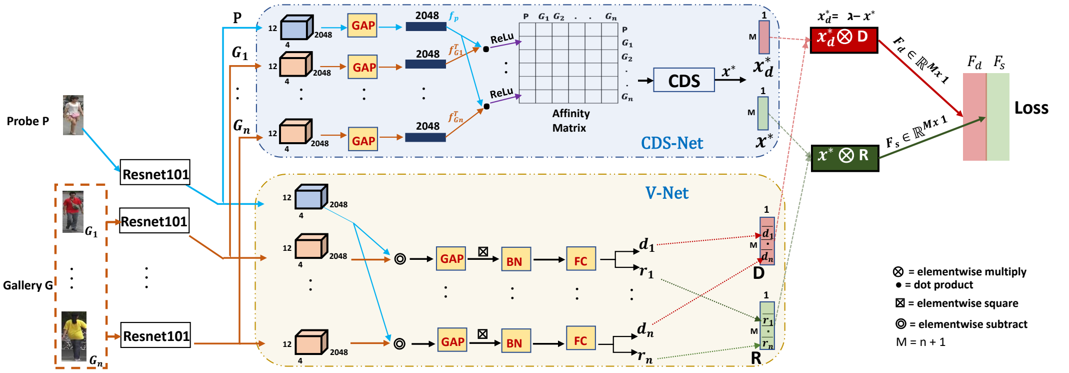


Figure 3. Workflow of the proposed DCDS. Given n number of gallery images, G, and probe image P, we first extract their Resent101 features right before the global average pooling (GAP) layer, which are then fed to CDS-Net (upper stream) and V-Net (lower stream) branches. In the CDS-branch, after applying GAP, we compute the similarity between M^2 pair of probe-gallery image features, f_p and f_{Gi}^T using their dot products, where T denotes a transpose. Followed by a ReLu operation. Thereby, we obtain $M \times M$ affinity matrix. Then, we run CDS taking the probe image as a constraint to find the solution $x^* \in \mathbb{R}^{M \times 1}$ (similarity), and the dissimilarity, x_d^*, is computed as an additive inverse of the similarity x^*. Likewise, in the lower stream (V-Net) we apply elementwise subtraction on M pair of probe-gallery features. This is followed by GAP, batch normalization (BN), and fully connected layer (FC) to obtain probe-gallery similarity score, $R \in \mathbb{R}^{M \times 1}$, and probe-gallery dissimilarity score, $D \in \mathbb{R}^{M \times 1}$. Afterwards, we elementwise multiply x^* and R, and x_d^* and D, to find the final similarity, F_s, and disimilarity, F_d, scores, respectively. Finally, to find the prediction loss of our model, we apply a cross entropy loss, the ground truth (G_t) is given as $G_t \in \mathbb{R}^{M \times 1}$.

Hence, a **positive** $W_S(i)$ indicates that adding i into its neighbors in S will raise the internal coherence of the set, whereas in the presence of a **negative** $W_S(i)$ value we expect the overall coherence to decline. In CDS, besides the additional feature, which allows us to incorporate a constraint element in the resulting cluster, all the characteristics of DS are inherited.

3.2.1 A Set of Person Images as a Constrained Cluster

We cast person re-identification as finding a constrained cluster. Given a probe and gallery images, we treat the probe image as a constraint to find a constrained cluster; where the elements of the cluster refer to the relevant images to the probe image. As customary, let us consider a given mini-batch with M number of person-images, and each mini batch with k person identities (ID), thus, each person-ID has $\Omega = M/k$ images in the given mini-batch. Note that, here, instead of a random sampling we design a custom sampler which samples k number of person IDs in each mini-batch. Let $B = \{I_{p_1}^1, ...I_{p_1}^\Omega, I_{p_2}^1, ...I_{p_2}^\Omega, ...I_{p_k}^1, ...I_{pk}^\Omega\}$ refers to the set of images in a single mini-batch. Each time when we consider image $I_{p_1}^1$ as a probe image P, images which belong to the same person id, $\{I_{p_1}^2, I_{p_1}^3...I_{p_1}^k\}$, should be assigned a large membership score to be in that cluster. In contrast, the remaining images in the mini-batch should be assigned significantly smaller membership-score to be part of that cluster. Note that our ultimate goal here is to find a constrained cluster which comprises all the images of the corresponding person given in that specific mini-batch. Thus, each participant in a given mini-batch is assigned a membership-score to be part of a cluster. Furthermore, $\sum_{i=1}^{M} z_i = 1$, where z_i denotes the membership score of each image in the cluster.

As can be seen from the toy example in Figure 2, the initial pairwise similarities between the query and gallery images hold valuable information, which define the relationships of nodes in the given graph. However, it is not straightforward to redefine the initial pairwise similarities in a way which exploit the inter-images relationship. Dominant Sets (DS) overcome this problem by defining a weight of each image g_3, p, g_1, g_2 with regard to subset $S \backslash i$ as depicted in Figure 2, $(2-5)$, respectively. As can be observed from Figure 2, adding node g_3 to cluster S degrades the coherency of cluster $S = \{p, g_1, g_2, g_3\}$, whereas the relative similarity of the remaining images with respect to set $\{p, g_1, g_2\}$ has a positive impact on the coherency of the cluster. It is evident that the illustration in Figure 2 verifies that the proposed DCDS could easily measure the contribution of each node in the graph and utilize it in an end-to-end learning process. Thereby, unlike a siamese, triplet and quadruplet based contrastive methods, DCDS consider the whole set of images in the mini-batch to measure the similarity of image pairs and enhance the learning process.

3.3. CDS Based End-to-End Learning

In this section, we discuss the integration of CDS in end-to-end learning. As can be seen from Figure 3, there are two main branches: CDS-Net and V-Net. We adopt a siamese based Resent101, V-Net, with a novel verification loss to find probe-gallery similarity (r_i) and dissimilarity (d_i) scores. Given probe and gallery images, V-Net outputs two vectors which are probe-gallery similarities R and dissimilarities D. In the CDS-Net, the elements of pairwise affinity matrix are computed first as a dot product of the global pooled features of a pair of images; that is followed by a ReLu operation. Afterward, the replicator dynamics [30] is applied, which is a discrete time solver of the parametrized quadratic program, Equ. 5 (that is equivalent to Equ. 2), whose solution corresponds to the CDS. Thus, assuming that there are M images in the given mini-batch, the replicator dynamics, Equ. 3, is recursively applied M times taking each image in the mini-batch as a constraint. Consider a graph $G = (V, E, w)$ and its corresponding adjacency matrix $A \in R^{M \times M}$, and probe $P \subseteq V$. First, a proper modification of the affinity matrix A is applied by setting the diagonal elements corresponding to the subset $V \backslash P$ to parameter α and diagonal elements corresponding to the constraint image P zero. Next, the modified adjacency matrix, B, is fed to the Replicator dynamics, by initiating the dynamics with a characteristic vector of uniform distribution x^{t_0}, such that initially all the images in the mini-batch are assigned equal membership probability to be part of the cluster. Then, to find a constrained cluster a parametrized quadratic program is defined as:

$$\begin{array}{lll} \text{maximize} & f_P^\alpha(\mathbf{x})^i = \mathbf{x}^T B \mathbf{x} & \text{where}, B = A - \alpha \hat{I}_p. \\ \text{subject to} & \mathbf{x} \in \Delta & \end{array} \tag{5}$$

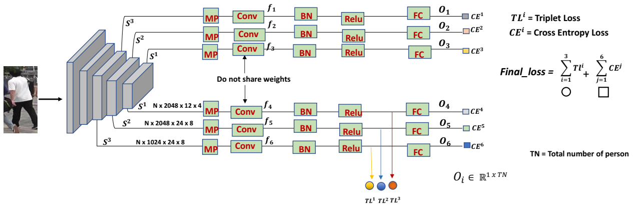


Figure 4. Illustrates the auxiliary net, which consists of two branches which are jointly trained. We first use features at different layers, S_1, S_2, S_3, and then feed these to Maxpooling (MP), Conv, BN, ReLu and FC layers for further encoding. We then compute triplet losses employing the features from the lower three streams after ReLu, shown by yellow, blue, and red circles. Next, after the final FC layer, we compute the cross-entropy loss $(CE^1, ..., CE^6)$ for each of the six different outputs, O_i, from the upper and lower stream shown by distinct colored-boxes. Note that even if the upper and lower stream apply the same operations, on S_1, S_2 and S_3, they do not share the weights; thus the encoding is different. We compute the final loss as the sum of the average of the triplet and cross entropy losses.

The solution, $\mathbf{x}_i^*$, of $f_P^\alpha(\mathbf{x})^i$ is a characteristics vector which indicates the probability of each gallery image to be included in a cluster, containing the probe image P^i. Thus, once we obtain the CDS, $\mathbf{x}_i^* = [z_{g_1}^i, z_{g_2}^i ... z_{g_M}^i]$, for each probe image, we store each solution $\mathbf{x}_i^*$, in $Y \in \mathbb{R}^{M \times M}$, as

$$Y = \begin{pmatrix} \mathbf{x}_i^* \\ \vdots \\ \mathbf{x}_M^* \end{pmatrix} = \begin{pmatrix} z_{g_1}^1 & z_{g_2}^1 & \cdots & z_{g_M}^1 \\ \vdots & & \ddots & \vdots \\ z_{g_1}^M & z_{g_2}^M & \cdots & z_{g_M}^M \end{pmatrix}.$$

Likewise, for each probe, P^i, we store the probe-gallery similarity, R, and dissimilarity, D, obtained from the V-Net (shown in Figure 3) in S' and D' as, $S' = [R^1, R^2, ... R^M]$ and $D' = [D^1, D^2, ... D^M]$. Next, we fuse the similarity obtained from the CDS branch with the similarity from the V-Net as

$$\begin{array}{l} F_s = \beta Y \otimes (1-\beta)(S'), \\ F_d = \beta Y_d \otimes (1-\beta)(D'), \quad where, \quad Y_d = \delta - Y \end{array} \tag{6}$$

δ is empirically set to 0.3 and β is a fusing parameter, which will be discussed in Experiments. We then vectorize F_s and F_d into $\mathbb{R}^{(M^2 \times 2)}$, where, the first column stores the dissimilarity score, while the second column stores the similarity score. Afterward, we simply apply cross entropy loss to find the prediction loss. The intriguing feature of our model is that it does not need any custom optimization technique, it can be end-to-end optimized through a standard back-propagation algorithm. Note that, Figure 3 illustrates the case of a single probe-gallery, whereas Equ. 6 shows the solution of M probe images in a given mini-batch.

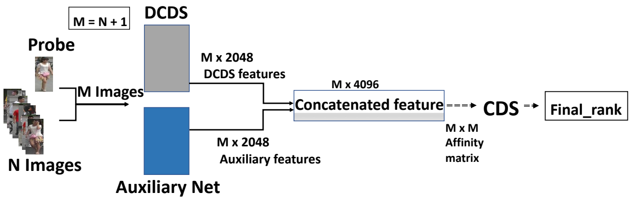


Figure 5. During testing, given a probe and gallery images, we extract DCDS and auxiliary net features and concatenate them to find a single vector. Afterward, we build $M \times M$ affinity matrix and run CDS with constraint expansion mechanism to find the final probe-gallery similarity rank.

3.4. Auxiliary Net

We integrate an auxiliary net (Figure 4) to further improve the performance of our model. The auxiliary net is trained based on the multi-scale prediction of ResNet50 [13]. It is a simple yet effective architecture, whereby we can easily compute both triplet and cross entropy loss of different layers of ResNet50 [13], hence further enhancing the learning capability. Consequently, we compute the average of both losses to find the final loss. As can be observed from Figure 4, we employ three features from different layers of ResNet50 and then we fed these three features to the subsequent layers, MP, Conv, BN, and FC layers. Next, we compute triplet and cross entropy loss for features from the ReLu and FC layers, respectively. During testing phase we concatenate the features from the DCDS and Auxiliary Net to find 4,096 dimensional feature vector. We then apply CDS to find the final ranking_score, (See Figure 5).

3.5. Constraint Expansion During Testing

We propose a new scheme (illustrated in Figure 6) to expand the number of constraints in order to guide the similarity propagation during the testing phase. Given an affinity matrix, which is constructed using the features obtained from the concatenated features (shown in Figure 5), we first collect k-NN's of the probe image. Then, we run CDS on the graph of the NNs. Next, from the resulting constrained cluster, we select the one with the highest membership score, which is used as a constraint in the subsequent step. We then use multiple-constraints and run CDS.

4. Experiments

To validate the performance of our method we have conducted several experiments on three publicly available benchmark datasets, namely CUHK03 [16], Market1501 [37], and DukeMTMC-reID [38].

4.1. Datasets and Evaluation Metrics

Datasets: CUHK03 [16] dataset comprises 14,097 manually and automatically cropped images of 1,467 identities, which are captured by two cameras on campus; in our experiments, we have used manually annotated images. Market1501 dataset [37] contains 32,668 images which are split into 12, 936 and 19,732 images as training and testing set, respectively. Market1501 dataset has totally 1501 identities which are captured by five high-resolution and one low-resolution cameras, the training and testing sets have 751 and 750 identities respectively. To obtain the person bounding boxes, Deformable part Model (DPM) [12] is utilized. DukeMTMC-reID is generated from a tracking dataset called DukeMTMC; that is captured by 8 high-resolution cameras, and person-bounding box is manually cropped; it is organized as 16,522 images of 702 person for training and 18, 363 images of 702 person for testing.

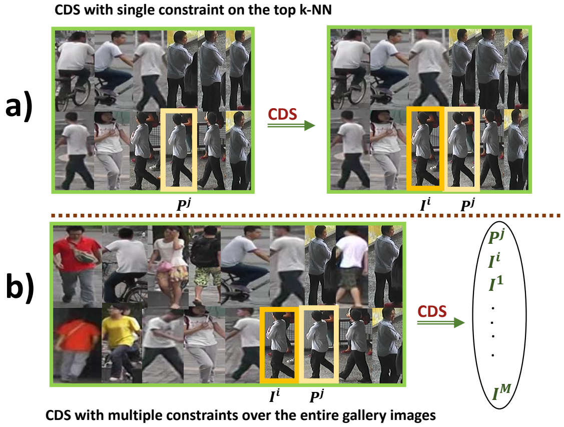

CDS with single constraint on the top k-NN
a)
CDS
P^j
I^i
P^j
b)
I^i
P^j
CDS
P^j
I^i
I^1
.
.
.
.
I^M
CDS with multiple constraints over the entire gallery images

Figure 6. a) given a constraint (probe-image) P^j, we first collect k-NNs to P^j, based on the pairwise similarities. Subsequently, we run CDS on the graph of the k-NN. Then, we choose a cluster-member with the highest membership-score, I^i. b) we re-run CDS, considering P^j and I^i as constraints, over the graph of the all set of images. Afterward, we consider the solution as our final rank.

Evaluation Metrics: Following the recent person re-id methods, we use mean average precision (mAP) as suggested in [37], and Cumulated Matching Characteristics (CMC) curves to evaluate the performance of our model. Furthermore, all the experiments are conducted using the standard single query setting [37].

4.2. Implementation Details

We implement DCDS based on ResNet101 [13] architecture, which is pretrained on imagenet dataset. We adopt the training strategy of Kalayeh *et al.* [14], and aggregate eight different person re-id benchmark datasets to train our model. In total, the merged dataset contains 89,091 images, which comprises 4,937 person-ID (detail of the eight datasets is given in the supplementary material). We first train our model using the merged dataset (denoted as multi-dataset (MD)) for 150 epochs and fine-tune it with CUHK03, Market1501, and DukeMTMC-reID dataset. To train our model using the merged dataset, we set image resolution to 450 × 150. Subsequently, for fine-tuning the model we set image resolution to 384 × 128. Mini-batch

Methods	mAP	rank-1	rank-5
SGGNN [21] ECCV18	82.8	92.3	96.1
DKPM [22] CVPR18	75.3	90.1	96.7
DGSRW [20] CVPR18	82.5	92.7	96.9
GCSL [7] CVPR18	81.6	93.5	-
CPC [28] CVPR18	69.48	83.7	-
MLFN [6] CVPR18	74.3	90.0	-
HA-CNN [17] CVPR18	75.7	91.2	-
PA [23] ECCV18	74.5	88.8	95.6
HSP [14] CVPR18	83.3	93.6	97.5
Ours	**85.8**	**94.81**	**98.1**
RA_{W-RR} [29] CVPR18	86.7	90.9	-
PA_{W-RR} [23] ECCV18	89.9	93.4	96.4
HSP_{W-RR} [14] CVPR18	90.9	94.6	96.8
Ours$_{W-RR}$	**93.3**	**95.4**	**98.3**

Table 1. A comparison of the proposed method with state-of-the-art methods on Market1501 dataset. Upper block, without re-ranking (Wo-RR). Lower block, with re-ranking method, $W-RR$, [39].

Methods	rank-1	rank-5
SGGNN [21] ECCV18	95.3	99.1
DKPM [22] CVPR18	91.1	98.3
DGSRW [20] CVPR18	94.9	98.7
GCSL [7] CVPR18	90.2	98.5
MLFN [6] CVPR18	89.2	-
CPC [28] CVPR18	88.1	-
PA [23] ECCV18	88.0	97.6
HSP [14] CVPR18	94.3	99.0
Ours	**95.8**	**99.1**

Table 2. A comparison of the proposed method with state-of-the-art methods on CUHK03 dataset.

size is set to 64, each mini-batch has 16 person-IDs and each person-ID has 4 images. We also experiment only using a single dataset for training and testing, denoted as single-dataset (SD). For data augmentation, we apply random horizontal flipping and random erasing [40]. For optimization we use Adam, we initially set the learning rate to 0.0001, and drop it by 0.1 in every 40 epochs. β, is set to 0.9.

4.3. Results on Market1501 Datasets

As can be seen from Table 1, on Market1501 dataset, our proposed method improves state-of-the-art method [14], Wo-RR, by $2.5\%, 1.21\%$, and 0.6% in mAP, rank-1 and rank-5 scores, respectively. Moreover, compared to state-of-the-art graph-based DNN method, SGGNN [21], the improvement margins are $3\%, 2.5\%$, and 2% in mAP, rank-1, and rank-5 score, respectively. Thus, our framework has significantly demonstrated its benefits over state-of-the-art graph-based DNN models. To further improve the result we have adapted a re-ranking scheme [39], and we compare our method with state-of-the art methods which use a re-ranking method as a post-processing. As it can be seen from Table 1, our method has mAP gain of 2.4% over HSP [14], and 10.5 % over SGGNN[21], 10.8 % over DGSRW.

Methods	mAP	rank-1	rank-5
SGGNN [21] ECCV18	68.2	81.1	88.4
DKPM [22] CVPR18	63.2	80.3	89.5
DGSRW [20] CVPR18	66.4	80.7	88.5
GCSL [7] CVPR18	69.5	84.9	-
CPC [28] CVPR18	59.49	76.44	-
MLFN [6] CVPR18	62.8	81.0	-
RAPR [29] CVPR18	80.0	84.4	-
PA [23] ECCV18	64.2	82.1	90.2
HSP [14] CVPR18	73.3	85.9	**92.9**
Ours	**75.5**	**87.5**	-
PA_{W-RR} [23] ECCV18	83.9	88.3	93.1
HSP_{W-RR} [14] CVPR18	84.9	**88.9**	**94.27**
Ours $_{W-RR}$	**86.1**	88.5	-

Table 3. A comparison of the proposed method with state-of-the-art methods on DukeMTMC-reID dataset.Upper block, Wo-RR. Lower block, with re-ranking method [39].

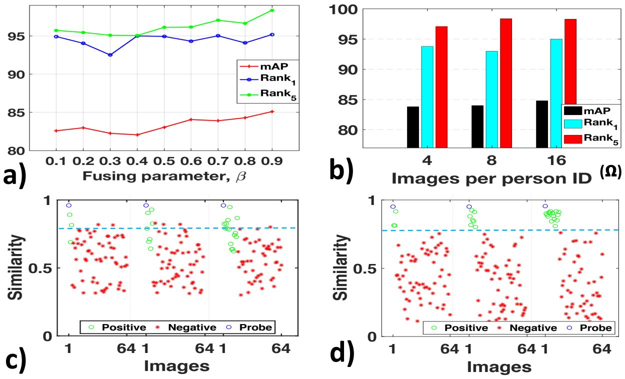


Figure 7. Illustrates different experimental analysis performed on Market1501 dataset. a) shows the impact of β. b) shows the performance of our model with varying the number of images per person in a given batch. c) and d) illustrate the similarity between the probe and gallery images obtained from the baseline and DCDS method, respectively. It can be observed that the baseline method has given larger similarity values for false positive samples (red asterisks above the blue dashed-line) and smaller similarity values for false negative samples (green circles below the blue dashed-line). On the other hand, the proposed DCDS has efficiently assigned the appropriate similarity scores to the true positive and negative samples.

4.4. Results on CUHK03 Datasets

Table 2 shows the performance of our method on CUHK03 dataset. Since most of the Graph-based DNN models report their results on the standard protocol [16], we have experimented on the standard evaluation protocol,

Methods	Market1501			CUHK03		DukeMTMC-reID		
	mAP	rank-1	rank-5	rank-1	rank-5	mAP	rank-1	rank-5
Baseline SD	72.2	86.5	94.0	87.1	94.3	61.1	77.6	87.3
Baseline MD	74.3	87.5	95.3	87.7	95.2	62.3	79.1	88.8
DCDS (SD)	81.4	93.3	97.6	93.1	98.8	69.1	83.3	89.0
DCDS (MD)	82.3	93.7	98.0	93.9	98.9	70.5	84.0	90.3
Ours (SD + Auxil Net)	83.0	93.9	**98.2**	95.4	99.0	74.4	85.6	**93.7**
Ours (MD + Auxil Net)	**85.8**	**94.1**	98.1	**95.8**	**99.1**	**75.5**	**86.1**	93.2

Table 4. Ablation studies on the proposed method. SD and MD respectively refer to the method trained on single and multiple-aggregated datasets. Baseline is the proposed method without CDS branch.

to make fair comparison. As can be observed from Table 2, our method gains a marginal improvement in the rank-1 and rank-5 scores.

4.5. Results on DukeMTMC-reID Dataset

Likewise, in DukeMTMC-reID dataset, the improvements of our proposed method is noticeable (see Table 3). Our method has surpassed state-of-the-art method [14], without applying a re-ranking (Wo-RR), by $2.2\%/1.6\%$ in mAP/rank-1 scores. Moreover, compared to state-of-the-art graph-based DNN our method outperforms DGSRW [20], SGGNN [21] and GCSL [7] by $9.1\%, 7.3\%$, and 6% in mAP, respectively. Using a reranking method [39], we report competitive results in all evaluation metrics.

4.6. Ablation Study

To investigate the impact of each component in our architecture, we perform an ablation study. Thus, we report the contributions of each module in Table 4. To make a fair comparison with the baseline and graph-based DNN models, the ablations study is conducted on SD setup.

Improvements over the baseline. As our main contribution is the DCDS, we examine its impact over the baseline method. The baseline method refers to the lower branch of our architecture that incorporates the verification network, which has also been utilized in [22], [20], [21]. On Market1501 dataset, in SD setup, DCDS provides improvements of $9.2\%, 6.8\%$ and 3.6% in mAP, rank-1, and rank-5 scores, respectively, over the baseline method; whereas in DukeMTMC-reID dataset the proposed DCDS improves the baseline method by $8.0\%, 5.5\%$ and 1.7% in mAP, rank-1, and rank-5 scores, respectively.

Comparison with graph-based deep models. We compare our method with recent graph-based-deep models, which adapt similar baseline method as ours, such as [20],[21]. As a result, on DukeMTMC-reID dataset our method surpass [20] by 9.1%(6.8%),Wo-RR, and [21] by 17.9 % (7.4 %), W-RR, in mAP (rank-1) scores. In light of this, We can conclude that incorporating a constrained-clustering mechanism in end-to-end learning has a significant benefit on finding a robust similarity ranking. In addition, experimental findings demonstrate the superiority of DCDS over existing graph-based DNN models.

Parameter analysis. Experimental results by varying several parameters are shown in Figure 7. Figure 7(a) shows the effect of β on the mAP. Thereby, we can observe that the mAP tends to increase with a larger β value. This shows that the result gets better when we deviate much from the CDS branch. Figure 7(b) shows the impact of the number of images per person-ID (Ω) in a given batch. We have experimented setting Ω to 4, 8, and 16, as can be seen, we obtain a marginal improvement when we set Ω to 16. However, considering the direct relationship between the running time and Ω, the improvement is negligible. c) and d) show probe-gallery similarity obtained from baseline and DCDS method, using three different probe-images, with a batch size of 64, and setting Ω to 4, 8 and 16.

5. Conclusion

In this work, we presented a novel insight to enhance the learning capability of a DNN through the exploitation of a constrained clustering mechanism. To validate our method, we have conducted extensive experiments on several benchmark datasets. Thereby, the proposed method not only improves state-of-the-art person re-id methods but also demonstrates the benefit of incorporating a constrained-clustering mechanism in the end-to-end learning process. Furthermore, the presented work could naturally be extended to other applications which leverage a similarity-based learning. As a future work, we would like to investigate dominant sets clustering as a loss function.

Acknowledgment

This research is partly supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA *R&D* Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] Ejaz Ahmed, Michael J. Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3908–3916, 2015. 2

[2] Leulseged Tesfaye Alemu and Marcello Pelillo. Multi-feature fusion for image retrieval using constrained dominant sets. *CoRR*, abs/1808.05075, 2018. 2

[3] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3356–3365, 2017. 1

[4] Song Bai, Zhichao Zhou, Jingdong Wang, Xiang Bai, Longin Jan Latecki, and Qi Tian. Ensemble diffusion for retrieval. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 774–783, 2017. 1

[5] Gedas Bertasius, Lorenzo Torresani, Stella X. Yu, and Jianbo Shi. Convolutional random walk networks for semantic image segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6137–6145, 2017. 1, 2, 3

[6] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2109–2118, 2018. 7

[7] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep CRF for person re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8649–8658, 2018. 1, 2, 3, 7, 8

[8] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, pages 1320–1329. IEEE Computer Society, 2017. 2

[9] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1335–1344, 2016. 2

[10] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 2

[11] Michael Donoser and Horst Bischof. Diffusion processes for retrieval revisited. In *CVPR*, pages 1320–1327. IEEE Computer Society, 2013. 1

[12] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 6

[14] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1062–1071, 2018. 6, 7, 8

[15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. 2

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159. IEEE Computer Society, 2014. 2, 6, 7

[17] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294. IEEE Computer Society, 2018. 7

[18] Chen Change Loy, Chunxiao Liu, and Shaogang Gong. Person re-identification by manifold ranking. In *IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013*, pages 3567–3571, 2013. 1

[19] Massimiliano Pavan and Marcello Pelillo. Dominant sets and pairwise clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):167–172, 2007. 2, 3

[20] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2265–2274, 2018. 1, 2, 7, 8

[21] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pages 508–526, 2018. 1, 2, 7, 8

[22] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6886–6895, 2018. 7, 8

[23] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, pages 418–437, 2018. 7

[24] Yonatan Tariku Tesfaye, Eyasu Zemene, Marcello Pelillo, and Andrea Prati. Multi-object tracking using dominant sets. *IET Computer Vision*, 10(4):289–297, 2016. 2

[25] Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, and Mubarak Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *CoRR*, abs/1706.06196, 2017. 2

[26] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 791–808, 2016. 2

[27] Chu Wang, Marcello Pelillo, and Kaleem Siddiqi. Dominant set clustering and pooling for multi-view 3d object recognition. In *BMVC*. BMVA Press, 2017. 2

[28] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1470–1478, 2018. 7

[29] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8042–8051, 2018. 7

[30] Jörgen W Weibull. *Evolutionary Game Theory*. MIT press, 1995. 5

[31] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7444–7452, 2018. 2

[32] Eyasu Zemene, Leulseged Tesfaye Alemu, and Marcello Pelillo. Constrained dominant sets for retrieval. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 2568–2573, 2016. 2

[33] Eyasu Zemene, Leulseged Tesfaye Alemu, and Marcello Pelillo. Dominant sets for ”constrained” image segmentation. *CoRR*, abs/1707.05309, 2017. 2

[34] Eyasu Zemene and Marcello Pelillo. Interactive image segmentation using constrained dominant sets. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 278–294, 2016. 2, 3

[35] Eyasu Zemene, Yonatan Tariku Tesfaye, Haroon Idrees, Andrea Prati, Marcello Pelillo, and Mubarak Shah. Large-scale image geo-localization using dominant sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(1):148–161, 2019. 2

[36] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3239–3248, 2017. 2

[37] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124. IEEE Computer Society, 2015. 6

[38] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, pages 3774–3782. IEEE Computer Society, 2017. 6

[39] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 3652–3661. IEEE Computer Society, 2017. 7, 8

[40] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017. 7