CONSTRAINED DOMINANT SETS AND ITS APPLICATIONS IN COMPUTER VISION

Doctoral Dissertation of:
Alemu Leulseged Tesfaye

Supervisor:
Prof. Marcello Pelillo

Ph.D. Coordinator:
Prof. Ricardo Focardi

Abstract

I N this thesis, we present new schemes which leverage a constrained clustering method to solve several computer vision tasks ranging from image retrieval, image segmentation and co-segmentation, to person re-identification. In the last decades clustering methods have played a vital role in computer vision applications; herein, we focus on the extension, reformulation, and integration of a well-known graph and game theoretic clustering method known as Dominant Sets. Thus, we have demonstrated the validity of the proposed methods with extensive experiments which are conducted on several benchmark datasets.

We first discuss 'Dominant Sets for "Constrained" Image Segmentation,' DSCIS. In DSCIS, we present a unified model to tackle image segmentation and co-segmentation problem in both an interactive and unsupervised fashion; whereby, the proposed algorithm can deal naturally with several types of constraints and input modality, including scribbles, sloppy contours, and bounding boxes, and is able to robustly handle noisy annotations on the part of the user. Our method is based on some properties of a family of quadratic optimization problems related to dominant sets, a well-known graph-theoretic notion of a cluster which generalizes the concept of a maximal clique to edge-weighted graphs. In particular, we show that by properly controlling a regularization parameter which determines the structure and the scale of the underlying problem, we are in a position to extract groups of dominant-set clusters that are constrained to contain predefined elements.

Following, we present novel schemes for content-based image retrieval (CBIR) using constrained dominant sets (CDS). We present two different

I

CBIR methods. The first method, 'Multi-feature Fusion for Image Retrieval Using Constrained Dominant Sets,' MfFIR, fuse several hand-crafted and deep features to endow a representative similarity which better define the closeness of given query and gallery images; whereas, the second one, 'Constrained Dominant Sets for Image Retrieval,' CDSIR, exploit a constrained diffusion process to produce a robust similarity between query and gallery images. In MfFIR, we propose a computationally efficient approach to fuse several hand-crafted and deep features, based on the probabilistic distribution of a given membership score of a constrained cluster in an unsupervised manner. Towards this end, we first introduce an incremental nearest neighbor (NN) selection method, whereby we dynamically select k-NN to the query. Next, we build several graphs from the obtained NN sets and employ constrained dominant sets (CDS) on each graph G to assign edge weights which consider the intrinsic manifold structure of the graph, and detect false matches to the query. Finally, we compute the positive-impact weight (PIW) based on the dispersive degree of the characteristics vector. As a result, we exploit the entropy of a cluster membership-score distribution. In addition, the final NN set bypasses a heuristic voting scheme. Our approach presents two main advantages. Firstly, compared to the state of the art methods, it can robustly quantify the effectiveness of features for a specific query, without any supervision. Secondly, by diffusing the pairwise similarity between the nearest neighbors, our model can easily avoid the inclusion of false-positive matches in the final shortlist. On the other hand, in CDSIR, we leverage constrained dominant sets to dynamically constrain a similarity diffusion process to provide context-sensitive similarities.

Finally, we present a Deep Constrained Dominant Sets (DCDS); in which, we are able to optimize the constrained-clustering process in an end-to-end manner and leverage contextual information in the learning process. In this work, we integrate the well-known graph and game-theoretic method called CDS into a deep model and tackle the challenging computer vision problem of person re-identification. Furthermore, we reformulate the problem of person re-identification as a constrained-clustering problem and build a model that is end-to-end trainable.

Acknowledgments

I first thank the almighty God for everything. I then must express my deepest gratitude to my supervisor professor Marcello Pelillo, for his scientific guidance, encouragement and fruitful discussion during my Ph.D. journey.

I also would love to thank my co-authors: Eyasu Zemene (Josh) for being not only a co-author but also a good friend and brother; and Dr. Mubarak Shah for granting me a visiting scholar position at CRCV, University of Central Florida, whereby I solidify my computer vision knowledge further.

And, special thanks go to Yoni, Teddy, Sure, Seyum, Tinsu, Ili, Felix, Carmelo, and Ayoub; please accept my sincere gratitude for being a good friend; moreover, for the unforgettable moments, we had in Campo and Baum (coffee area). I also would like to thank the Department of Computer Science of Ca' Foscari University of Venice for financing my Ph.D. study; I thank Prof. Ricardo Focardi and Nicola Miotello for co-ordinating the Ph.D. program and for the kind assistance they have provided me without hesitation.

I thank my external reviewers: Prof. Lamberto Ballan, and Prof. Xu Yongchao for the time they spent on carefully reading the thesis and for their insightful comments and suggestions.

Last but not least, I thank my family for the unconditional love and support, especially my mother (etetye) and aunt (eneye).

Preface

Tᴴɪs dissertation is submitted in fulfillment of the requirements for the degree of doctor of philosophy at Ca' Foscari University of Venice. The thesis presents novel methods which inherit several techniques from game theory, graph theory, and deep learning. The first chapter introduces the graph and game theoretic clustering methods called Dominant Sets and its constrained variant Constrained Dominant Sets. The second chapter presents the application of constrained dominant sets to tackle the problem of interactive image segmentation and co-segmentation (in both supervised and unsupervised fashion) [99]; it has been appeared in Transactions on Pattern Analysis and Machine Intelligence (TPAMI). The third chapter discusses two distinct methods which attack the same problem of content-based image retrieval. The first method, called Constrained Dominant Sets for Image Retrieval [169], has been presented in International Conference on Pattern Recognition (ICPR); whereas, the second one called Multi-feature Fusion using Constrained Dominant Sets for Image retrieval [8] has been appeared in a journal known as Image and Vision Computing (IVC). Finally, the last chapter presents a very interesting work which integrates constrained dominant sets in a deep neural network model. It is a collaboration work with Dr. Mubarak Shah; which has been done while I was a visiting scholar at the Center for Research in Computer Vision (CRCV). This work [9] has been published in International Conference on Computer Vision (ICCV).

Introduction

Manually labeling a large amount of data, that arise due to a deluge of information from several automations and data collections, is costly. Due to this reason, unsupervised clustering has attracted a considerable attention of pattern recognition community. Clustering deals with defining classes from the data without a prior knowledge of the class labels. Cluster analysis has been applied to solve several real world problems such as anomaly detection, image segmentation, natural language processing, document grouping and recommendation systems.

Clustering methods can be roughly divided into two main categories such as partitioning and hierarchical algorithm. Partitioning based clustering methods split the dataset into k <or = n groups, where n is the number of objects in the dataset, whereas hierarchical algorithms gradually form clusters through either agglomerations or divisions. Furthermore, there has been a resurgence of interest around graph based (pairwise) methods [6, 52, 127], that cast the data to be clustered (pixels, super-pixel, edge elements, etc) as a vertices of a similairity (edge-weighted) graph, where the edges reflect neighborhood relations, and the weights denote the similarity between data. Indeed, it is natural to map the data to be clustered to the nodes of a weighted graph (the so-called similarity graph), with edge weights representing similarity relations. In cluster analysis, graph clustering is defined as a process of searching for groups of related vertices in a graph. Graph-theoretic clustering methods are of significant interest since they cast clustering as pure graph-theoretic problems for which a solid theory and powerful algorithms have been developed. As can be observed

from [43], these methods can produce highly intricate clusters, but they rarely optimize an easily specified global cost function. Graph-theoretic clustering methods basically comprises searching for certain combinatorial structures in the similarity graph, such as a minimum cut [52, 127, 150] or a minimum spanning tree [166].

In this thesis, we present different approaches that leverage techniques which are based on some properties of a family of quadratic optimization problems related to dominant sets, a well-known graph-theoretic notion of a cluster which generalizes the concept of a maximal clique to edge-weighted graphs. Moreover, we exploit the constrained version of dominant sets clustering known as constrained dominant sets clustering [172] (CDS); which is based on some properties of a family of quadratic optimization problems related to dominant sets. In particular, by properly controlling a regularization parameter which determines the structure and the scale of the underlying problem, one can extract a dominant set cluster which is constrained to contain user-provided constraints. Thus, we able to tackle several computer vision problems such as image segmentation and co-segmentation, image retrieval, and person re-identification problems.

We first discuss our novel and multi-modal scheme, which is formulated in such a way that tackles the problem of image segmentation and co-segmentation in both unsupervised and interactive manner. Image segmentation is arguably one of the oldest and best-studied problems in computer vision, being a fundamental step in a variety of real-world applications, and yet remains a challenging task [132] [50]. Besides the standard, purely bottom-up formulation, which involves partitioning an input image into coherent regions, in the past few years several variants have been proposed which are attracting increasing attention within the community. Most of them usually take the form of a "constrained" version of the original problem, whereby the segmentation process is guided by some external source of information. For example, user-assisted (or "interactive") segmentation has become quite popular nowadays, especially because of its potential applications in problems such as image and video editing, medical image analysis, etc. [21, 24, 81, 87, 102, 113, 120, 149]. Given an input image and some information provided by a user, usually in the form of a scribble or of a bounding box, the goal is to provide as output a foreground object in such a way as to best reflect the user's intent. By exploiting high-level, semantic knowledge on the part of the user, which is typically difficult to formalize, we are therefore able to effectively solve segmentation problems which would be otherwise too complex to be tackled using fully automatic segmentation algorithms. Another example of a "constrained" segmenta-

tion problem is image co-segmentation. Given a set of images, the goal here is to jointly segment same or similar foreground objects. The problem was first introduced by Rother *et al.* [33] who used histogram matching to simultaneously segment the foreground object out from a given pair of images. Recently, several techniques have been proposed which try to co-segment groups containing more than two images, even in the presence of similar backgrounds. Joulin *et al.* [4], for example, proposed a discriminative clustering framework, combining normalized cut and kernel methods and the framework has recently been extended in an attempt to handle multiple classes and a significantly larger number of images [5]. In this work (which is an extended version of [172]), we propose a unified approach to address this kind of problems which can deal naturally with various input modalities, or constraints, and is able to robustly handle noisy annotations on the part of the external source. In particular, we shall focus on interactive segmentation and co-segmentation (in both the unsupervised and the interactive versions).

Next, we present our works on CBIR. Image retrieval (CBIR) has recently attracted considerable attention within the computer vision community, especially because of its potential applications such as database retrieval, web and mobile image search. The goal of semantic image search, or content-based image retrieval (CBIR), is to search for a query image from a given image dataset. This is done by computing image similarities based on low-level image features, such as color, texture, shape and spatial relationship of images. Variation of images in illumination, rotation, and orientation has remained a major challenge for CBIR. Recently, locally constraining the diffusion process has shown its effectiveness on learning the intrinsic manifold structure of a given data. However, existing constrained-diffusion based retrieval methods have several shortcomings. For instance, manual choice of optimal local neighborhood size, do not allow for intrinsic relation among the neighbors, fix initialization vector to extract dense neighbor; which negatively affect the affinity propagation. In CDSIR, leveraging the constrained dominant sets we tackle the above issues. On the other hand, we develop a feature-fusion based image retrieval method known as Multi-feature Fusion for Image Retrieval Using Constrained Dominant Sets (MfFIR) . Multi-feature based CBIR attacks the CBIR problem by introducing an approach which utilizes multiple low-level visual features of an image. Intuitively, if the to-be-fused feature works well by itself, it is expected that its aggregation with other features will improve the accuracy of the retrieval. Nevertheless, it is quite hard to learn in advance the effectiveness of the to-be-fused features for a specific

query image. In MfFIR, we propose a computationally efficient approach to fuse several hand-crafted and deep features, based on the probabilistic distribution of a given membership score of a constrained cluster in an unsupervised manner.

We finally discuss our work on the challenging computer vision problem of person re-identification. In this work, for the very first time, we integrate the well known graph and game theoretic clustering method called dominant sets in end-to-end manner. Thereby, we do the optimization of constrained-clustering process in the context of deep learning. 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 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 summarize, the main contributions of this thesis are:

- It leverages the constrained dominant sets to attack several computer vision problems in both classical and deep flavors.

- The proposed DSCIS has a number of interesting features which distinguishes it from existing approaches. Specifically: 1) it solves both image segmentation and co-segmentation in an interactive and unsupervised manner. 2) in the case of noiseless scribble inputs, it asks the user to provide *only* foreground pixels; 3) it turns out to be *robust* in the presence of input noise, allowing the user to draw, e.g., imperfect scribbles.

- The proposed Image retrieval methods come with several advantages.

In particular, the proposed CDSIR: 1) it constrains the diffusion process by locally extracting dense neighbors whose local neighborhood size (K) is fixed automatically; means that different neighbors can have a different value of K. 2) it has no initialization step; the dynamics, to extract the dense neighbors, can start at any point in the standard simplex 3) it turns out to be *robust* to noisy affinity matrices.

- On other hand, through MfFIR, we contribute a generic approach which can be applied not only to image retrieval but also to other computer vision problems, such as object detection and person re-identification. Furthermore, unlike existing feature-fusion methods, we propose a simple but efficient entropy-based feature effectiveness weighting system.

- Finally, for the very first time, we integrate the well-known clustering method, dominant sets, in a deep neural network (DNN) model. Moreover, we establish a one-to-one correspondence between person re-identification and constrained clustering problem.

Contents

List of Figures

List of Tables

Dominant Sets and Quadratic Optimization

Clustering or partitioning a given data based on the similarity among the data points is a fundamental task in many fields of study such as Machine Learning, Computer Vision, and Statistics. In this chapter, we discuss the well-known graph and game-theoretic pairwise data clustering scheme called Dominant Sets and its constrained variant Constrained Dominant Sets.

In the dominant set framework, the data to be clustered are represented as an undirected edge-weighted graph with no self-loops $G = (V, E, w)$, where $V = \{1, ..., n\}$ is the vertex set, $E \subseteq V \times V$ is the edge set, and $w : E \to R_+^*$ 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 $n \times n$ nonnegative, symmetric matrix $A = (a_{ij})$ defined as $a_{ij} = w(i, j)$, if $(i, j) \in E$, and $a_{ij} = 0$ otherwise. Since in G there are no self-loops, note that all entries on the main diagonal of A are zero.

For a non-empty subset $S \subseteq V$, $i \in S$, and $j \notin S$, define

$$\phi_S(i, j) = a_{ij} - \frac{1}{|S|} \sum_{k \in S} a_{ik} . \tag{1.1}$$

This quantity measures the (relative) similarity between nodes 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:

$$w_S(i) = \begin{cases} 1, & \text{if} \quad |S| = 1, \\ \sum_{j \in S \setminus \{i\}} \phi_{S \setminus \{i\}}(j, i) w_{S \setminus \{i\}}(j), & \text{otherwise} . \end{cases} \tag{1.2}$$

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 $S \setminus \{i\}$. Therefore, a positive $w_S(i)$ indicates that adding i into its neighbors in S will increase the internal coherence of the set, whereas in the presence of a negative value we expect the overall coherence to be decreased.

A non-empty subset of vertices $S \subseteq V$ such that $W(T) > 0$ for any non-empty $T \subseteq S$, is said to be a *dominant set* if:

1. $w_S(i) > 0$, for all $i \in S$,

2. $w_{S \cup \{i\}}(i) < 0$, for all $i \notin S$.

It is evident from the definition that a dominant set satisfies the two basic properties of a cluster: internal coherence and external incoherence. Condition 1 indicates that a dominant set is internally coherent, while condition 2 implies that this coherence will be destroyed by the addition of any vertex from outside. In other words, a dominant set is a maximally coherent data set.

Now, consider the following linearly-constrained quadratic optimization problem:

$$\begin{aligned} \text{maximize} \quad & f(\mathbf{x}) = \mathbf{x}' A \mathbf{x} \\ \text{subject to} \quad & \mathbf{x} \in \Delta \end{aligned} \tag{1.3}$$

where a prime denotes transposition and

$$\Delta = \left\{ \mathbf{x} \in R^n \; : \; \sum_{i=1}^{n} x_i = 1, \text{ and } x_i \geq 0 \text{ for all } i = 1 \dots n \right\}$$

is the standard simplex of R^n. In [103, 107] a connection is established between dominant sets and the local solutions of (1.3). In particular, it is
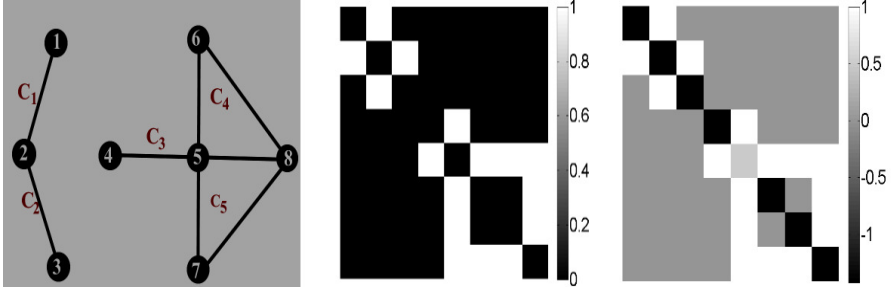
Figure 1.1: *An example graph (left), corresponding affinity matrix (middle), and scaled affinity matrix built considering vertex 5 as a user constraint (right). Notation C_i refers to the ith maximal clique.*

shown that if S is a dominant set then its "weighted characteristic vector," which is the vector $\mathbf{x}^S \in \Delta$ defined as

$$x_i^S = \begin{cases} \frac{w_S(i)}{\sum_{j \in S} w_S(j)}, & \text{if} \quad i \in S, \\ 0, & \text{otherwise} \end{cases}$$

is a strict local solution of (1.3). Conversely, under mild conditions, it turns out that if $\mathbf{x}$ is a (strict) local solution of program (1.3) then its "support"

$$\sigma(\mathbf{x}) = \{i \in V \ : \ x_i > 0\}$$

is a dominant set. By virtue of this result, we can find a dominant set by first localizing a solution of program (1.3) with an appropriate continuous optimization technique, and then picking up the support set of the solution found. In this sense, we indirectly perform combinatorial optimization via continuous optimization. A generalization of these ideas to hypergraphs has recently been developed in [119].

Note that, by construction, dominant sets capture compact structures. To deal with arbitrarily shaped clusters, path-based similarity measures can profitably be used [171]. In the work reported in this thesis, however, we did not make use of this notion.

1.1 Constrained dominant sets

Let $G = (V, E, w)$ be an edge-weighted graph with n vertices and let A denote as usual its (weighted) adjacency matrix. Given a subset of vertices $S \subseteq V$ and a parameter $\alpha > 0$, define the following parameterized family

of quadratic programs:

$$\begin{aligned}
\text{maximize} \quad & f_S^\alpha(\mathbf{x}) = \mathbf{x}'(A - \alpha \hat{I}_S)\mathbf{x} \\
\text{subject to} \quad & \mathbf{x} \in \Delta
\end{aligned} \tag{1.4}$$

where $\hat{I}_S$ is the $n \times n$ diagonal matrix whose diagonal elements are set to 1 in correspondence to the vertices contained in $V \setminus S$ and to zero otherwise, and the 0's represent null square matrices of appropriate dimensions. In other words, assuming for simplicity that S contains, say, the first k vertices of V, we have:

$$\hat{I}_S = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-k} \end{pmatrix}$$

where I_{n-k} denotes the $(n-k) \times (n-k)$ principal submatrix of the $n \times n$ identity matrix I indexed by the elements of $V \setminus S$. Accordingly, the function f_S^α can also be written as follows:

$$f_S^\alpha(\mathbf{x}) = \mathbf{x}'A\mathbf{x} - \alpha \mathbf{x}'_{\bar{S}}\mathbf{x}_{\bar{S}}$$

$\mathbf{x}_{\bar{S}}$ being the $(n-k)$-dimensional vector obtained from $\mathbf{x}$ by dropping all the components in S. Basically, the function f_S^α is obtained from f by inserting in the affinity matrix A the value of the parameter α in the main diagonal positions corresponding to the elements of $V \setminus S$.

Notice that this differs markedly, and indeed generalizes, the formulation proposed in [106] for obtaining a hierarchical clustering in that here, only a subset of elements in the main diagonal is allowed to take the α parameter, the other ones being set to zero. We note in fact that the original (non-regularized) dominant-set formulation (1.3) [103] as well as its regularized counterpart described in [106] can be considered as degenerate version of ours, corresponding to the cases $S = V$ and $S = \emptyset$, respectively. It is precisely this increased flexibility which allows us to use this idea for finding groups of "constrained" dominant-set clusters.

We now derive the Karush-Kuhn-Tucker (KKT) conditions for program (3.3), namely the first-order necessary conditions for local optimality (see, e.g., [94]). For a point $\mathbf{x} \in \Delta$ to be a KKT-point there should exist n nonnegative real constants $\mu_1, \ldots, \mu_n$ and an additional real number λ such that

$$[(A - \alpha \hat{I}_S)\mathbf{x}]_i - \lambda + \mu_i = 0$$

for all $i = 1 \ldots n$, and

$$\sum_{i=1}^{n} x_i \mu_i = 0 \, .$$

Since both the x_i's and the μ_i's are nonnegative, the latter condition is equivalent to saying that $i \in \sigma(\mathbf{x})$ implies $\mu_i = 0$, from which we obtain:

$$[(A - \alpha \hat{I}_S)\mathbf{x}]_i \begin{cases} = \lambda, & \text{if } i \in \sigma(\mathbf{x}) \\ \leq \lambda, & \text{if } i \notin \sigma(\mathbf{x}) \end{cases}$$

for some constant λ. Noting that $\lambda = \mathbf{x}'A\mathbf{x} - \alpha \mathbf{x}'_{\bar{S}}\mathbf{x}_{\bar{S}}$ and recalling the definition of $\hat{I}_S$, the KKT conditions can be explicitly rewritten as:

$$\begin{cases} (A\mathbf{x})_i - \alpha x_i &= \mathbf{x}'A\mathbf{x} - \alpha \mathbf{x}'_{\bar{S}}\mathbf{x}_{\bar{S}}, & \text{if } i \in \sigma(\mathbf{x}) \text{ and } i \notin S \\ (A\mathbf{x})_i &= \mathbf{x}'A\mathbf{x} - \alpha \mathbf{x}'_{\bar{S}}\mathbf{x}_{\bar{S}}, & \text{if } i \in \sigma(\mathbf{x}) \text{ and } i \in S \\ (A\mathbf{x})_i &\leq \mathbf{x}'A\mathbf{x} - \alpha \mathbf{x}'_{\bar{S}}\mathbf{x}_{\bar{S}}, & \text{if } i \notin \sigma(\mathbf{x}) \end{cases} \quad (1.5)$$

We are now in a position to discuss the main results which motivate the algorithm presented in this thesis. Note that, in the sequel, given a subset of vertices $S \subseteq V$, the face of Δ corresponding to S is given by: $\Delta_S = \{x \in \Delta : \sigma(x) \subseteq S\}$.

Proposition 1. *Let $S \subseteq V$, with $S \neq \emptyset$. Define*

$$\gamma_S = \max_{\mathbf{x} \in \Delta_{V \setminus S}} \min_{i \in S} \frac{\mathbf{x}'A\mathbf{x} - (A\mathbf{x})_i}{\mathbf{x}'\mathbf{x}} \quad (1.6)$$

and let $\alpha > \gamma_S$. If $\mathbf{x}$ is a local maximizer of f_S^α in Δ, then $\sigma(\mathbf{x}) \cap S \neq \emptyset$.

Proof. Let $\mathbf{x}$ be a local maximizer of f_S^α in Δ, and suppose by contradiction that no element of $\sigma(\mathbf{x})$ belongs to S or, in other words, that $\mathbf{x} \in \Delta_{V \setminus S}$. By letting

$$i = \arg\min_{j \in S} \frac{\mathbf{x}'A\mathbf{x} - (A\mathbf{x})_j}{\mathbf{x}'\mathbf{x}}$$

and observing that $\sigma(\mathbf{x}) \subseteq V \setminus S$ implies $\mathbf{x}'\mathbf{x} = \mathbf{x}'_{\bar{S}}\mathbf{x}_{\bar{S}}$, we have:

$$\alpha > \gamma_S \geq \frac{\mathbf{x}'A\mathbf{x} - (A\mathbf{x})_i}{\mathbf{x}'\mathbf{x}} = \frac{\mathbf{x}'A\mathbf{x} - (A\mathbf{x})_i}{\mathbf{x}'_{\bar{S}}\mathbf{x}_{\bar{S}}} .$$

Hence, $(A\mathbf{x})_i > \mathbf{x}'A\mathbf{x} - \alpha \mathbf{x}'_{\bar{S}}\mathbf{x}_{\bar{S}}$ for $i \notin \sigma(\mathbf{x})$, but this violates the KKT conditions (1.5), thereby proving the proposition. $\qquad \square$

The following proposition provides a useful and easy-to-compute upper bound for γ_S.

Proposition 2. *Let $S \subseteq V$, with $S \neq \emptyset$. Then,*

$$\gamma_S \leq \lambda_{\max}(A_{V \setminus S}) \tag{1.7}$$

where $\lambda_{\max}(A_{V \setminus S})$ is the largest eigenvalue of the principal submatrix of A indexed by the elements of $V \setminus S$.

Proof. Let $\mathbf{x}$ be a point in $\Delta_{V \setminus S}$ which attains the maximum γ_S as defined in (1.6). Using the Rayleigh-Ritz theorem [62] and the fact that $\sigma(\mathbf{x}) \subseteq V \setminus S$, we obtain:

$$\lambda_{\max}(A_{V \setminus S}) \geq \frac{\mathbf{x}'_{\bar{S}} A_{V \setminus S} \mathbf{x}_{\bar{S}}}{\mathbf{x}'_{\bar{S}} \mathbf{x}_{\bar{S}}} = \frac{\mathbf{x}' A \mathbf{x}}{\mathbf{x}' \mathbf{x}} .$$

Now, define $\gamma_S(\mathbf{x}) = \max\{(A\mathbf{x})_i \ : \ i \in S\}$. Since A is nonnegative so is $\gamma_S(\mathbf{x})$, and recalling the definition of γ_S we get:

$$\frac{\mathbf{x}' A \mathbf{x}}{\mathbf{x}' \mathbf{x}} \geq \frac{\mathbf{x}' A \mathbf{x} - \gamma_S(\mathbf{x})}{\mathbf{x}' \mathbf{x}} = \gamma_S$$

which concludes the proof. $\qquad \square$

The two previous propositions provide us with a simple technique to determine dominant-set clusters containing user-selected vertices. Indeed, if S is the set of vertices selected by the user, by setting

$$\alpha > \lambda_{\max}(A_{V \setminus S}) \tag{1.8}$$

we are guaranteed that all local solutions of (3.3) will have a support that necessarily contains elements of S. Note that this does not necessarily imply that the (support of the) solution found corresponds to a dominant-set cluster of the original affinity matrix A, as adding the parameter $-\alpha$ on a portion of the main diagonal intrinsically changes the scale of the underlying problem. However, we have obtained extensive empirical evidence which supports a conjecture which turns out to be very useful for our interactive image segmentation application.

To illustrate the idea, let us consider the case where edge-weights are binary, which basically means that the input graph is unweighted. In this case, it is known that dominant sets correspond to maximal cliques [103]. Let $G = (V, E)$ be our unweighted graph and let S be a subset of its vertices. For the sake of simplicity, we distinguish three different situations of increasing generality.

Case 1. The set S is a singleton, say $S = \{u\}$. In this case, we know from Proposition 2 that all solutions $\mathbf{x}$ of f_α^S over Δ will have a support which contains u, that is $u \in \sigma(\mathbf{x})$. Indeed, we conjecture that there will be a

unique local (and hence global) solution here whose support coincides with the *union* of all maximal cliques of G which contain u.

Case 2. The set S is a clique, not necessarily maximal. In this case, Proposition 2 predicts that all solutions $\mathbf{x}$ of (3.3) will contain at least one vertex from S. Here, we claim that indeed the support of local solutions is the union of the maximal cliques that contain S.

Case 3. The set S is not a clique, but it can be decomposed as a collection of (possibly overlapping) maximal cliques $C_1, C_2, ..., C_k$ (maximal with respect to the subgraph induced by S). In this case, we claim that if $\mathbf{x}$ is a local solution, then its support can be obtained by taking the union of all maximal cliques of G containing one of the cliques C_i in S.

To make our discussion clearer, consider the graph shown in Fig. 1.1. In order to test whether our claims hold, we used as the set S different combinations of vertices, and enumerated all local solutions of (3.3) by multi-start replicator dynamics (see Section 1.2). Some results are shown below, where on the left-hand side we indicate the set S, while on the right hand-side we show the supports provided as output by the different runs of the algorithm.

1. $S = \emptyset$ $\Rightarrow$ $\sigma(\mathbf{x}_1) = \{5, 6, 8\},\ \sigma(\mathbf{x}_2) = \{5, 7, 8\}$
2. $S = \{2\}$ $\Rightarrow$ $\sigma(\mathbf{x}) = \{1, 2, 3\}$
3. $S = \{5\}$ $\Rightarrow$ $\sigma(\mathbf{x}) = \{4, 5, 6, 7, 8\}$
4. $S = \{4, 5\}$ $\Rightarrow$ $\sigma(\mathbf{x}) = \{4, 5\}$
5. $S = \{5, 8\}$ $\Rightarrow$ $\sigma(\mathbf{x}) = \{5, 6, 7, 8\}$
6. $S = \{1, 4\}$ $\Rightarrow$ $\sigma(\mathbf{x}_1) = \{1, 2\},\ \sigma(\mathbf{x}_2) = \{4, 5\}$
7. $S = \{2, 5, 8\}$ $\Rightarrow$ $\sigma(\mathbf{x}_1) = \{1, 2, 3\},\ \sigma(\mathbf{x}_2) = \{5, 6, 7, 8\}$

Notice that in the unconstrained case ($S = \emptyset$), the algorithm returns the two largest cliques, depending on the starting point. We refer the reader to [108] (and references therein) for a thorough analysis of the use of replicator and similar dynamics for the (unconstrained) maximum clique problem.

The previous observations can be summarized in the following general statement which does comprise all three cases. Let $S = C_1 \cup C_2 \cup \ldots \cup C_k$ ($k \geq 1$) be a subset of vertices of G, consisting of a collection of cliques C_i ($i = 1 \ldots k$). Suppose that condition (1.8) holds, and let $\mathbf{x}$ be a local solution of (3.3). Then, $\sigma(\mathbf{x})$ consists of the union of all maximal cliques containing some clique C_i of S.

7

1.2 Finding constrained dominant sets using replicator dynamics

Given an arbitrary real-valued $n \times n$ matrix $W = (w_{ij})$, consider the following continuous-time dynamical system

$$\dot{x} = x_i \left((W\mathbf{x})_i - \mathbf{x}'W\mathbf{x} \right) \tag{1.9}$$

for $i = 1 \ldots n$, where a dot signifies derivative w.r.t. time, and its discrete-time counterpart:

$$x_i(t+1) = x_i(t) \frac{C + (W\mathbf{x}(t))_i}{C + \mathbf{x}(t)'W\mathbf{x}(t)} \tag{1.10}$$

where C is a proper constant to avoid negative values in the numerator (and denominator). These are known as *replicator dynamics* in evolutionary game theory [60, 148] and it turns out that, for a large constant C, (1.10) is essentially an Euler discretization of (1.9).

It is readily seen that the standard simplex Δ is invariant under these dynamics, which means that every trajectory starting in Δ will remain in Δ for all future times. Moreover, their stationary points, i.e., the points satisfying $\dot{x} = 0$ for (1.9) and $x_i(t+1) = x_i(t)$ for (1.10), coincide and are the solutions of the equations:

$$x_i \left((W\mathbf{x})_i - \mathbf{x}'W\mathbf{x} \right) = 0 \ .$$

A stationary point $\mathbf{x}$ is said to be *asymptotically stable* if every trajectory which starts close enough to $\mathbf{x}$ will eventually converge to it.

The following result, known as the Fundamental Theorem of Natural Selection [60], provides us useful information concerning the convergence properties of replicator dynamics.

Theorem 1. *If W is symmetric ($W = W'$), then the function $\mathbf{x}(t)'W\mathbf{x}(t)$ is strictly increasing along any nonstationary trajectory, under both continuous-time (1.9) and discrete-time (1.10) replicator dynamics. Furthermore, any such trajectory converges to a stationary point. Finally, a vector $\mathbf{x} \in \Delta$ is asymptotically stable under (1.9) and (1.10) if and only if $\mathbf{x}$ is a strict local maximizer of $\mathbf{x}'W\mathbf{x}$ on Δ.*

Thanks to these properties, replicator dynamics naturally suggest themselves as a simple heuristics for finding (constrained) dominant sets [103]. In our case, the matrix W is given by

$$W = A - \alpha \hat{I}_S$$

which, in the discrete-time case, yields:

$$
x_i(t+1) \;=\; \begin{cases} x_i(t)\dfrac{C+(A\mathbf{x}(t))_i}{C+\mathbf{x}(t)'(A-\alpha\hat{I}_S)\mathbf{x}(t)} & \text{if } i \in S \\[3ex] x_i(t)\dfrac{C+(A\mathbf{x}(t))_i-\alpha x_i(t)}{C+\mathbf{x}(t)'(A-\alpha\hat{I}_S)\mathbf{x}(t)} & \text{if } i \notin S \end{cases}
\tag{1.11}
$$

Since the process cannot leave the boundary of Δ, it is customary to start the dynamics from some interior point, a common choice being the barycenter of Δ. This prevents the search from being initially biased in favor of any particular solution. By virtue of our theoretical results, we know that when the algorithm converges to a vector $\mathbf{x}$, its support $\sigma(\mathbf{x})$ will correspond to a constrained dominant set, while its positive components will reflect the membership score of the selected vertices.

Clearly, the behavior of the algorithm depends on the choice of the parameter α. Our own experience is that α might affect the convergence time (number of steps of the replicator dynamics) as well as the distribution of the membership scores of the final solution (i.e., the components of the converged vector). In particular, we observed that the membership scores assigned to the constrained dominant-set vertices become larger and larger (thereby making the scores of the other dominant-set vertices smaller and smaller) as α increases. This phenomenon, however, manifests itself more sensibly, and might become an issue, only for large values of α. No significant effect on the algorithm's performance has been observed for reasonable choices of the parameter. Accordingly, we recommend using a reasonably small value for α, close to the lower bound predicted by our theoretical results. This is what we actually did in all the experiments reported below. As for the parameter C in (1.11), its function is only to scale the matrix $A - \alpha\hat{I}_S$ properly to avoid negative values. An obvious choice would be $C = \alpha$, which is the value we used in our experiments.

Although in the experiments reported in this thesis we used the replicator dynamics described above, we mention a faster alternative to solve linearly constrained quadratic optimization problems like ours, namely *Infection and Immunization Dynamics* (InImDyn) [117]. Each step of InImDyn has a linear time/space complexity as opposed to the quadratic per-step complexity of replicator dynamics, and is therefore to be preferred in the presence of large affinity matrices.

1.3   Summary

In this chapter, we briefly introduced the well-known graph and game theoretic clustering algorithm called Dominant Sets, and its variant Constrained Dominant Sets.

CHAPTER *2*

Dominant Sets for "Constrained" Image Segmentation

Image segmentation has come a long way since the early days of computer vision, and still remains a challenging task. Modern variations of the classical (purely bottom-up) approach, involve, e.g., some form of user assistance (interactive segmentation) or ask for the simultaneous segmentation of two or more images (co-segmentation). At an abstract level, all these variants can be thought of as "constrained" versions of the original formulation, whereby the segmentation process is guided by some external source of information. In this chapter, we propose a new approach to tackle this kind of problems in a unified way. Our work is based on some properties of a family of quadratic optimization problems related to *dominant sets*, a graph-theoretic notion of a cluster which generalizes the concept of a maximal clique to edge-weighted graphs. In particular, we show that by properly controlling a regularization parameter which determines the structure and the scale of the underlying problem, we are in a position to extract groups of dominant-set clusters that are constrained to contain predefined elements. In particular, we shall focus on interactive segmentation and co-segmentation (in both the unsupervised and the interactive versions). The

11

proposed algorithm can deal naturally with several types of constraints and input modalities, including scribbles, sloppy contours and bounding boxes, and is able to robustly handle noisy annotations on the part of the user. Experiments on standard benchmark datasets show the effectiveness of our approach as compared to state-of-the-art algorithms on a variety of natural images under several input conditions and constraints.

2.1 Introduction

Segmentation is arguably one of the oldest and best-studied problems in computer vision, being a fundamental step in a variety of real-world applications, and yet remains a challenging task [132] [50].Besides the standard, purely bottom-up formulation, which involves partitioning an input image into coherent regions, in the past few years several variants have been proposed which are attracting increasing attention within the community. Most of them usually take the form of a "constrained" version of the original problem, whereby the segmentation process is guided by some external source of information. For example, user-assisted (or "interactive") segmentation has become quite popular nowadays, especially because of its potential applications in problems such as image and video editing, medical image analysis, etc. [21,24,81,87,102,113,120,149]. Given an input image and some information provided by a user, usually in the form of a scribble or of a bounding box, the goal is to provide as output a foreground object in such a way as to best reflect the user's intent. By exploiting high-level, semantic knowledge on the part of the user, which is typically difficult to formalize, we are therefore able to effectively solve segmentation problems which would be otherwise too complex to be tackled using fully automatic segmentation algorithms.

Another example of a "constrained" segmentation problem is co-segmentation. Given a set of images, the goal here is to jointly segment same or similar foreground objects. The problem was first introduced by Rother *et al.* [33] who used histogram matching to simultaneously segment the foreground object out from a given pair of images. Recently, several techniques have been proposed which try to co-segment groups containing more than two images, even in the presence of similar backgrounds. Joulin *et al.* [4], for example, proposed a discriminative clustering framework, combining normalized cut and kernel methods and the framework has recently been extended in an attempt to handle multiple classes and a significantly larger number of images [5].

In this chapter (which is an extended version of [172]), we propose a

unified approach to address this kind of problems which can deal naturally with various input modalities, or constraints, and is able to robustly handle noisy annotations on the part of the external source. In particular, we shall focus on interactive segmentation and co-segmentation (in both the unsupervised and the interactive versions).

Although various kinds of constraints can be envisaged to encode top-down information in segmentation processes, our work is focused on what we might refer to as "first-order" (or unary) constraints, which require that one or more "seed" points be part of the extracted group. Second- or higher-order constraints, of the type discussed for example in [47, 78, 165], which include must-link constraints (pairs of points that should belong to the same cluster) and cannot-link constraints (pairs of points that should belong to different clusters), will not be treated here, although it is not difficult to adapt our framework to deal with these cases.

Our approach is based on some properties of a parameterized family of quadratic optimization problems related to dominant-set clusters, a well-known generalization of the notion of maximal cliques to edge-weighted graph which have proven to be extremely effective in a variety of computer vision problems, including (automatic) image and video segmentation [103, 107] (see [116] for a recent review). In particular, we show that by properly controlling a regularization parameter which determines the structure and the scale of the underlying problem, we are in a position to extract groups of dominant-set clusters which are constrained to contain user-selected elements. We provide bounds that allow us to control this process, which are based on the spectral properties of certain submatrices of the original affinity matrix. The resulting algorithm has a number of interesting features which distinguishes it from existing approaches. Specifically: 1) it is able to deal in a flexible manner with *both* scribble-based and boundary-based input modalities (such as sloppy contours and bounding boxes); 2) in the case of noiseless scribble inputs, it asks the user to provide *only* foreground pixels; 3) it turns out to be *robust* in the presence of input noise, allowing the user to draw, e.g., imperfect scribbles or loose bounding boxes. Experimental results on standard benchmark datasets demonstrate the effectiveness of our approach as compared to state-of-the-art algorithms on a wide variety of natural images under several input conditions. Figure 2.1 shows some examples of how our system works in both interactive segmentation (in the presence of different input annotations) and co-segmentation settings.
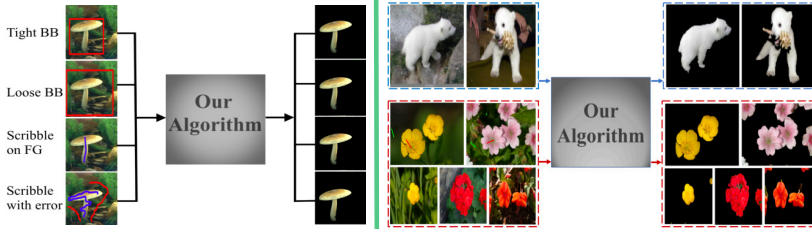
Figure 2.1:   **Left:** *An example of our interactive image segmentation method and its outputs, with different user annotation. Respectively from top to bottom, tight bounding box (Tight BB), loose bounding box (Loose BB), a scribble made (only) on the foreground object (Scribble on FG) and scribbles with errors.* **Right:** *Blue and Red dash-line boxes, show an example of our unsupervised and interactive co-segmentation methods, respectively.*

2.1.1   Background

Existing interactive segmentation interfaces fall into two broad categories, depending on whether the user annotation is given in terms of a scribble or of a bounding box, and supporters of the two approaches have both good reasons to prefer one modality against the other. For example, Wu et al. [149] claim that bounding boxes are the most natural and economical form in terms of the amount of user interaction, and develop a multiple instance learning algorithm that extracts an arbitrary object located inside a tight bounding box at unknown location. Yu et al. [164] also support the bounding-box approach, though their algorithm is different from others in that it does not need bounding boxes tightly enclosing the object of interest, whose production of course increases the annotation burden. They provide an algorithm, based on a Markov Random Field (MRF) energy function, that can handle input bounding box that only loosely covers the foreground object. Xian et al. [154] propose a method which avoids the limitations of existing bounding box methods - region of interest (ROI) based methods, though they need much less user interaction, their performance is sensitive to initial ROI.

On the other hand, several researchers, arguing that boundary-based interactive segmentation such as intelligent scissors [102] requires the user to trace the whole boundary of the object, which is usually a time-consuming and tedious process, support scribble-based segmentation. Bai et al. [16], for example, propose a model based on ratio energy function which can be optimized using an iterated graph cut algorithm, which tolerates errors in the user input. In general, the input modality in an interactive segmentation algorithm affects both its accuracy and its ease of use. Existing methods

work typically on a single modality and they focus on how to use that input most effectively. However, as noted recently by Jain and Grauman [66], sticking to one annotation form leads to a suboptimal tradeoff between human and machine effort, and they tried to estimate how much user input is required to sufficiently segment a novel input.

The co-segmentation problem has also been addressed using user interaction [35, 153]. Here, a user adds guidance, usually in the form of scribbles, on foreground objects of some of the input images. Batra *et al.* [35] proposed an extension of the (single-image) interactive segmentation algorithm of Boykov and Jolly [24]. They also proposed an algorithm that enables users to quickly guide the output of the co-segmentation algorithm towards the desired output via scribbles. Given scribbles, both on the background and the foreground, on some of the images, they cast the labeling problem as energy minimization defined over graphs constructed over each image in a group. Dong *et al.* [153] proposed a method using global and local energy optimization. Given background and foreground scribbles, they built a foreground and a background Gaussian mixture model (GMM) which are used as global guide information from users. By considering the local neighborhood consistency, they built the local energy as the local smooth term which is automatically learned using spline regression. The minimization problem of the energy function is then converted into constrained quadratic programming (QP) problem, where an iterative optimization strategy is designed for the computational efficiency.

2.2 Application to interactive image segmentation

In this section, we apply CDS to the interactive image segmentation problem. As input modalities we consider scribbles as well as boundary-based approaches (in particular, bounding boxes) and, in both cases, we show how the system is robust under input perturbations, namely imperfect scribbles or loose bounding boxes.

In this application the vertices of the underlying graph G represent the pixels of the input image (or superpixels, as discussed below), and the edge-weights reflect the similarity between them. As for the set S, its content depends on whether we are using scribbles or bounding boxes as the user annotation modality. In particular, in the case of scribbles, S represents precisely those pixels that have been manually selected by the user. In the case of boundary-based annotation instead, it is taken to contain only the pixels comprising the box boundary, which are supposed to represent the background scene. Accordingly, the union of the extracted dominant sets,

say $\mathcal{L}$ dominant sets are extracted which contain the set S, as described in the previous section and below, $\mathbf{UDS} = \mathcal{D}_1 \cup \mathcal{D}_2 ..... \cup \mathcal{D}_{\mathcal{L}}$, represents either the foreground object or the background scene depending on the input modality. For scribble-based approach the extracted set, $\mathbf{UDS}$, represent the segmentation result, while in the boundary-based approach we provide as output the complement of the extracted set, namely $\mathbf{V} \setminus \mathbf{UDS}$.

Figure 2.2 shows the pipeline of our system. Many segmentation tasks reduce their complexity by using superpixels (a.k.a. over-segments) as a preprocessing step [61, 149, 164] [144, 155]. While [149] used SLIC superpixels [1], [164] used a recent superpixel algorithm [185] which considers not only the color/feature information but also boundary smoothness among the superpixels. In this work, we used the over-segments obtained from Ultrametric Contour Map (UCM) which is constructed from Oriented Watershed Transform (OWT) using globalized probability of boundary (gPb) signal as an input [12].

We then construct a graph G where the vertices represent over-segments and the similarity (edge-weight) between any two of them is obtained using a standard Gaussian kernel

$$\mathtt{A}^{\sigma}_{ij} = \mathbb{K}_{i \neq j} exp(\|\mathbf{f}_i - \mathbf{f}_j\|^2 / 2\sigma^2)$$

where $\mathbf{f}_i$, is the feature vector of the ith over-segment, σ is the free scale parameter, and $\mathbb{K}_P = 1$ if P is true, 0 otherwise.

Given the affinity matrix A and the set S as described before, the system constructs the regularized matrix $M = A - \alpha \hat{I}_S$, with α chosen as prescribed in (1.8). Then, the replicator dynamics (1.10) are run (starting them from the simplex barycenter) until they converge to some solution vector $\mathbf{x}$. We then take the support of $\mathbf{x}$, remove the corresponding vertices from the graph and restart the replicator dynamics until all the elements of S are extracted.

2.2.1 Experiments and results

As mentioned above, the vertices of our graph represents over-segments and edge weights (similarities) are built from the median of the color of all pixels in RGB, HSV, and L*a*b* color spaces, and Leung-Malik (LM) Filter Bank [82]. The number of dimensions of feature vectors for each over-segment is then 57 (three for each of the RGB, L*a*b*, and HSV color spaces, and 48 for LM Filter Bank).

In practice, the performance of graph-based algorithms that use Gaussian kernel, as we do, is sensitive to the selection of the scale parameter
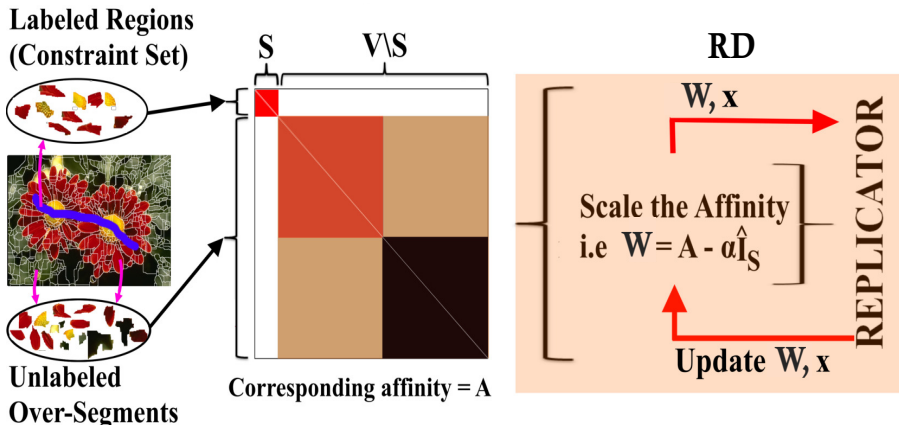
Figure 2.2: *Overview of our interactive segmentation system. **Left:** Over-segmented image (output of the UCM-OWT algorithm [12]) with a user scribble (blue label). **Middle:** The corresponding affinity matrix, using each over-segments as a node, showing its two parts: S, the constraint set which contains the user labels, and $V \setminus S$, the part of the graph which takes the regularization parameter α. **Right:** The optimization RD (Replicator Dynamics), starts from the barycenter and extracts the first dominant set and update $\mathbf{x}$ and $\mathbf{M}$, for the next extraction till all the dominant sets which contain the user labeled regions are extracted.*

σ. In our experiments, we have reported three different results based on the way σ is chosen: 1) CDS_Best_Sigma, in this case the best parameter σ is selected on a per-image basis, which indeed can be thought of as the optimal result (or upper bound) of the framework. 2) CDS_Single_Sigma, the best parameter in this case is selected on a per-database basis tuning σ in some fixed range, which in our case is between 0.05 and 0.2. 3) CDS_Self_Tuning, the σ^2 in the above equation is replaced, based on [167], by $\sigma_i * \sigma_j$, where $\sigma_i = mean(KNN(f_i))$, the mean of the K_Nearest_Neighbor of the sample f_i, K is fixed in all the experiment as 7.

Datasets: We conduct four different experiments on the well-known GrabCut dataset [120] which has been used as a benchmark in many computer vision tasks [83] [81, 135, 136, 149, 164] [112, 159]. The dataset contains 50 images together with manually-labeled segmentation ground truth. The same bounding boxes as those in [81] is used as a baseline bounding box. We also evaluated our scribbled-based approach using the well known Berkeley dataset which contains 100 images.

Metrics: We evaluate the approach using different metrics: error rate, fraction of misclassified pixels within the bounding box, Jaccard index which is given by, following [97], $J = \frac{|GT \cap O|}{|GT \cup O|}$, where GT is the ground truth and O is the output. The third metric is the Dice Similarity Coeffi-

cient (DSC), which measures the overlap between two segmented object volume, and is computed as $DSC = \frac{2*|GT \cap O|}{|GT|+|O|}$.

Annotations: In interactive image segmentation, users provide annotations which guides the segmentation. A user usually provides information in different forms such as scribbles and bounding boxes. The input modality affects both its accuracy and ease-of-use [66]. However, existing methods fix themselves to one input modality and focus on how to use that input information effectively. This leads to a suboptimal tradeoff in user and machine effort. Jain et al. [66] estimates how much user input is required to sufficiently segment a given image. In this work as we have proposed an interactive framework, figure 2.1, which can take any type of input modalities we will use four different type of annotations: bounding box, loose bounding box, scribbles - only on the object of interest -, and scribbles with error as of [16].

2.2.1.1 Scribble based segmentation

Given labels on the foreground as constraint set, we built the graph and collect (iteratively) all unlabeled regions (nodes of the graph) by extracting dominant set(s) that contains the constraint set (user scribbles). We provided quantitative comparison against several recent state-of-the-art interactive image segmentation methods which uses scribbles as a form of human annotation: [24], Lazy Snapping [87], Geodesic Segmentation [21], Random Walker [55], Transduction [42] , Geodesic Graph Cut [112], Constrained Random Walker [159].

We have also compared the performance of our algorithm againts Biased Normalized Cut (BNC) [96], an extension of normalized cut, which incorporates a quadratic constraint (bias or prior guess) on the solution **x**, where the final solution is a weighted combination of the eigenvectors of normalized Laplacian matrix. In our experiments we have used the optimal parameters according to [96] to obtain the most out of the algorithm. We also provide some qualitative comparisons with the Semi-Supervised Normalized Cut (SSNCut) algorithm recently introduced in [31], which incorporates (soft) must-link and cannot-link constraints.

Tables 2.1 and 2.2 and the plots in Figure 2.4 show the respective quantitative and the several qualitative segmentation results. Most of the results, reported on table 2.1, are reported by previous works [81, 112, 149, 159, 164]. We can see that the proposed CDS outperforms all the other approaches.

Error-tolerant Scribble Based Segmentation. This is a family of

Methods	Error Rate
BNC [96]	13.9
Graph Cut [24]	6.7
Lazy Snapping [87]	6.7
Geodesic Segmentation [21]	6.8
Random Walker [55]	5.4
Transduction [42]	5.4
Geodesic Graph Cut [112]	4.8
Constrained Random Walker [159]	4.1
CDS_Self Tuning (Ours)	**3.57**
CDS_Single Sigma (Ours)	**3.80**
CDS_Best Sigma (Ours)	2.72

Table 2.1: *Error rates of different scribble-based approaches on the Grab-Cut dataset.*

Methods	Jaccard Index
MILCut-Struct [149]	84
MILCut-Graph [149]	83
MILCut [149]	78
Graph Cut [120]	77
Binary Partition Trees [121]	71
Interactive Graph Cut [24]	64
Seeded Region Growing [2]	59
Simple Interactive O.E [51]	63
CDS_Self Tuning (Ours)	**93**
CDS_Single Sigma (Ours)	**93**
CDS_Best Sigma (Ours)	95

Table 2.2: *Jaccard Index of different approaches – first 5 bounding-box-based – on Berkeley dataset.*
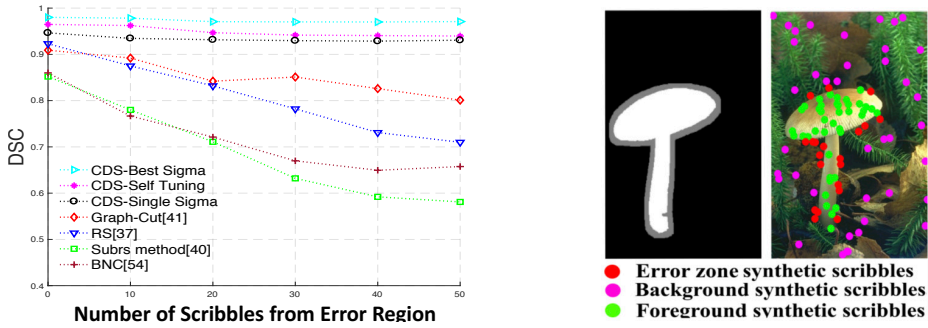
Figure 2.3: *Left: Performance of interactive segmentation algorithms, on Grab-Cut dataset, for different percentage of synthetic scribbles from the error region. **Right:** Synthetic scribbles and error region*

scribble-based approach, proposed by Bai et. al [16], which tolerates imperfect input scribbles thereby avoiding the assumption of accurate scribbles. We have done experiments using synthetic scribbles and compared the algorithm against recently proposed methods specifically designed to segment and extract the object of interest tolerating the user input errors [16, 90, 122, 129].

Our framework is adapted to this problem as follows. We give for our framework the foreground scribbles as constraint set and check those scribbled regions which include background scribbled regions as their members in the extracted dominant set. Collecting all those dominant sets which are free from background scribbled regions generates the object of interest.

Experiment using synthetic scribbles. Here, a procedure similar to the one used in [129] and [16] has been followed. First, 50 foreground pixels and 50 background pixels are randomly selected based on ground truth (see Fig. 2.3). They are then assigned as foreground or background scribbles, respectively. Then an error-zone for each image is defined as background pixels that are less than a distance D from the foreground, in which D is defined as 5 %. We randomly select 0 to 50 pixels in the error zone and assign them as foreground scribbles to simulate different degrees of user input errors. We randomly select 0, 5, 10, 20, 30, 40, 50 erroneous sample pixels from error zone to simulate the error percentage of 0%, 10%, 20%, 40%, 60%, 80%, 100% in the user input. It can be observed from figure 2.3 that our approach is not affected by the increase in the percentage of scribbles from error region.

2.2.1.2 Segmentation using bounding boxes

The goal here is to segment the object of interest out from the background based on a given bounding box. The corresponding over-segments which contain the box label are taken as constraint set which guides the segmentation. The union of the extracted set is then considered as background while the union of other over-segments represent the object of interest.

We provide quantitative comparison against several recent state-of-the-art interactive image segmentation methods which uses bounding box: Loose-Cut [164], GrabCut [120], OneCut [136], MILCut [149], pPBC and [135]. Table 2.3 and the pictures in Figure 2.4 show the respective error rates and the several qualitative segmentation results. Most of the results, reported on table 2.3, are reported by previous works [81, 112, 149, 159, 164].

Segmentation Using Loose Bounding Box. This is a variant of the bounding box approach, proposed by Yu et.al [164], which avoids the dependency of algorithms on the tightness of the box enclosing the object of interest. The approach not only avoids the annotation burden but also allows the algorithm to use automatically detected bounding boxes which might not tightly encloses the foreground object. It has been shown, in [164], that the well-known GrabCut algorithm [120] fails when the looseness of the box is increased. Our framework, like [164], is able to extract the object of interest in both tight and loose boxes. Our algorithm is tested against a series of bounding boxes with increased looseness. The bounding boxes of [81] are used as boxes with 0% looseness. A looseness L (in percentage) means an increase in the area of the box against the baseline one. The looseness is increased, unless it reaches the image perimeter where the box is cropped, by dilating the box by a number of pixels, based on the percentage of the looseness, along the 4 directions: left, right, up, and down.

For the sake of comparison, we conduct the same experiments as in [164]: 41 images out of the 50 GrabCut dataset [120] are selected as the rest 9 images contain multiple objects while the ground truth is only annotated on a single object. As other objects, which are not marked as an object of interest in the ground truth, may be covered when the looseness of the box increases, images of multiple objects are not applicable for testing the loosely bounded boxes [164]. Table 2.3 summarizes the results of different approaches using bounding box at different level of looseness. As can be observed from the table, our approach performs well compared to the others when the level of looseness gets increased. When the looseness $L = 0$, [149] outperforms all, but it is clear, from their definition of tight bounding

Methods	$L = 0\%$	$L = 120\%$	$L = 240\%$	$L = 600\%$
GrabCut [120]	7.4	10.1	12.6	13.7
OneCut [136]	6.6	8.7	9.9	13.7
pPBC [135]	7.5	9.1	9.4	12.3
MilCut [149]	**3.6**	-	-	-
LooseCut [164]	7.9	**5.8**	6.9	6.8
CDS_Self Tuning (Ours)	7.54	6.78	**6.35**	7.17
CDS_Single Sigma (Ours)	7.48	5.9	**6.32**	**6.29**
CDS_Best Sigma (Ours)	6.0	4.4	4.2	4.9

Table 2.3: *Error rates of different bounding-box approaches with different level of looseness as an input, on the Grab-Cut dataset. $L = 0\%$ implies a baseline bounding box as those in [81]*

box, that it is highly dependent on the tightness of the bounding box. It even shrinks the initially given bounding box by 5% to ensure its tightness before the slices of the positive bag are collected. For looseness of $L = 120$ we have similar result with LooseCut [164] which is specifically designed for this purpose. For other values of L our algorithm outperforms all the approaches.

Complexity. In practice, over-segmenting and extracting features may be treated as a pre-processing step which can be done before the segmentation process. Given the affinity matrix, we used replicator dynamics (1.10) to exctract constrained dominant sets. Its computational complexity per step is $O(N^2)$, with N being the total number of nodes of the graph. Given that our graphs are of moderate size (usually less than 200 nodes) the algorithm is fast and converges in fractions of a second, with a code written in Matlab and run on a core i5 6 GB of memory. As for the pre-processing step, the original *gPb-owt-ucm* segmentation algorithm was very slow to be used as a practical tools. Catanzaro et al. [25] proposed a faster alternative, which reduce the runtime from 4 minutes to 1.8 seconds, reducing the computational complexity and using parallelization which allow *gPb* contour detector and *gPb-owt-ucm* segmentation algorithm practical tools. For the purpose of our experiment we have used the Matlab implementation which takes around four minutes to converge, but in practice it is possible to give for our framework as an input, the GPU implementation [25] which allows the convergence of the whole framework in around 4 seconds.
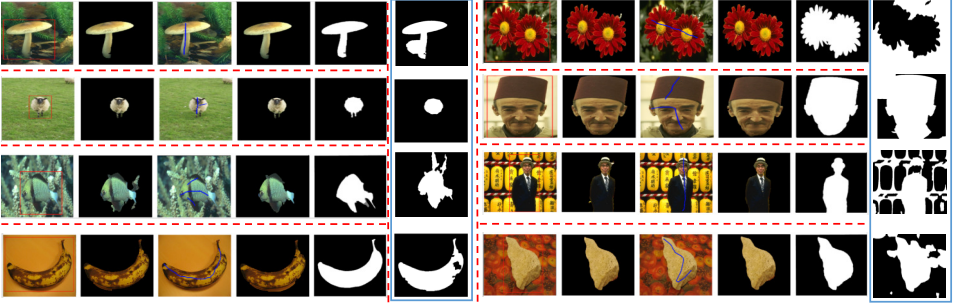
Figure 2.4: *Exemple results of the interactive segmentation algorithm tested on Grab-Cut dataset. (In each block of the red dashed line)* **Left:** *Original image with bounding boxes of [81].* **Middle left:** *Result of the bounding box approach.* **Middle:** *Original image and scribbles (observe that the scribbles are only on the object of interest).* **Middle right:** *Results of the scribbled approach.* **Right:** *The ground truth.* **Blue box:** *Results of Semi-Supervised Normalized Cuts (SSNcut) [31].*
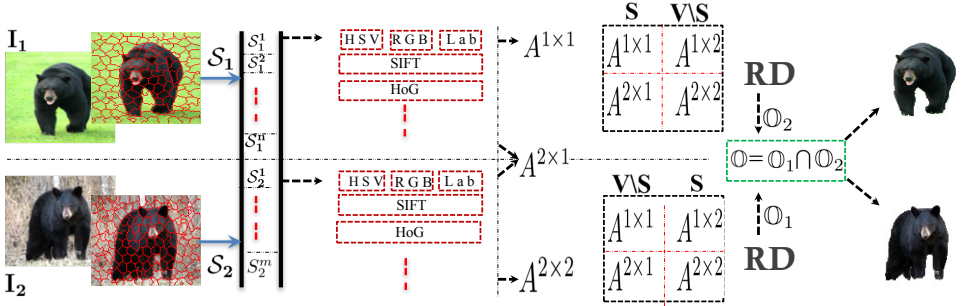
Figure 2.5: *Overview of our unsupervised co-segmentation algorithm.*

2.3 Application to co-segmentation

In this section, we describe the application of constrained dominant sets (CDS) to co-segmentation, both unsupervised and interactive. Among the difficulties that make this problem a challenging one, we mention the similarity among the different backgrounds and the similarity of object and background [131] (see, e.g., the top row of Figure 2.6). A measure of "objectness" has proven to be effective in dealing with such problems and improving the co-segmentation results [56, 131]. However, this measure alone is not enough especially when one aims to solve the problem using global pixel relations. One can see from Figure 2.6 (bottom) that the color of the cloth of the person, which of course is one of the objects, is similar to the color of the dog which makes systems that are based on objectness measure fail. Moreover the object may not be the one which we want to co-segment.

Figure 2.6: *The challenges of co-segmentation. Example image pairs: **(top left)** similar foreground objects with significant variation in background, **(top right)** foreground objects with similar background. The **bottom** part shows why user interaction is important for some cases. The **bottom left** is the image, **bottom middle** shows the objectness score, and the **bottom right** shows the user label.*

Figures 2.5 and 2.7 show the pipeline of our unsupervised and interactive co-segmentation algorithms, respectively. In figure 2.5, $\mathbf{I}_1$ and $\mathbf{I}_2$ are the given pair of images while $\mathcal{S}_1$ and $\mathcal{S}_2$ represent the corresponding sets of superpixels. The affinity is built using the objectness score of the superpixels and using different handcrafted features extracted from the superpixels. The set of nodes V is then divided into the constraint set (S) and the non-constraint set ($V \setminus S$). We run the CDS algorithm twice: first, setting the nodes of the graph that represent the first image as constraint set and $\mathbb{O}_2$ represents our output. Second we change the constraint set S with nodes that come from the second image and $\mathbb{O}_1$ represents the output. The intersection $\mathbb{O}$ refines the two results and represents the final output of the proposed unsupervised co-segmentation approach.

Our interactive co-segmentation approach, as shown in Figure 2.7, needs user interaction which guides the segmentation process putting scribbles (only) on some of the images with ambiguous objects or background. $\mathbf{I}_1, \mathbf{I}_2, ...\mathbf{I}_n$ are the scribbled images and $\mathbf{I}_{n+1}, ..., \mathbf{I}_{n+m}$ are unscribbled ones. The corresponding sets of superpixels are represented as $\mathcal{S}_1, \mathcal{S}_2, ...\mathcal{S}_n, ...\mathcal{S}_{n+1}, ...\mathcal{S}_{n+m}$. $\mathbf{A}'_{\mathbf{s}}$ and $\mathbf{A}_{\mathbf{u}}$ are the affinity matrices built using handcrafted feature-based similarities among superpixels of scribbled and unscribbled images respectively. Moreover, the affinities incorporate the objectness score of each node of the graph. $\mathcal{B}_{\mathbf{sp}}$ and $\mathcal{F}_{\mathbf{sp}}$ are (respectively) the background and foreground superpixels based on the user provided information. The CDS

algorithm is run twice over $\mathbf{A}'_\mathbf{s}$ using the two different user provided information as constraint sets which results outputs $\mathbb{O}_1$ and $\mathbb{O}_2$. The intersection of the two outputs, $\mathbb{O}$, help us get new foreground and background sets represented by $\mathcal{B}_\mathbf{s}$, $\mathcal{F}_\mathbf{s}$. Modifying the affinity $\mathbf{A}'_\mathbf{s}$, putting the similarities among elements of the two sets to zero, we get the new affinity $\mathbf{A}_\mathbf{s}$. We then build the biggest affinity which incorporates all images' superpixels. As our affinity is symmetric, $\mathbf{A}_\mathbf{us}$ and $\mathbf{A}_\mathbf{su}$ are equal and incorporates the similarities among the superpixels of the scribbled and unscribbled sets of images. Using the new background and foreground sets as two different constraint sets, we run CDS twice which results outputs $\mathbb{O}'_1$ and $\mathbb{O}'_2$ whose intersection ($\mathbb{O}'$) represents the final output.

2.3.1 Graph representation and affinity matrix

Given an image, we over-segment it to get its superpixels $\mathcal{S}$, which are considered as vertices of a graph. We then extract different features from each of the superpixels. The first features we consider are obtained from the different color spaces: RGB, HSV and CIE Lab. Given the superpixels, say size of n, of an image i, $\mathcal{S}_i$, $\mathcal{F}^i_c$ is a matrix of size $n \times 9$ which is the mean of each of the channels of the three color spaces of pixels of the superpixel. The mean of the SIFT features extracted from the superpixel $\mathcal{F}^i_s$ is our second feature. The last feature which we have considered is the rotation invariant histogram of oriented gradient (HoG), $\mathcal{F}^i_h$.

The dot product of the SIFT features is considered as the SIFT similarity among the nodes, let us say the corresponding affinity matrix is A_s. Motivated by [27], the similarity among the nodes of image i and image j ($i \neq j$), based on color, is computed from their Euclidean distance $\mathcal{D}^{i\times j}_c$ as

$$A^{i\times j}_c = max(\mathcal{D}_c) - \mathcal{D}^{i\times j}_c + min(\mathcal{D}_c)$$

The HoG similarity among the nodes, $A^{i\times j}_h$, is computed in a similar way , as A_c, from the diffusion distance. All the similarities are then min max normalized.

We then construct the $A^{i\times i}_c$, the similarities among superpixels of image i, which only considers adjacent superpixels as follows. First, construct the dissimilarity graph using their Euclidean distance considering their average colors as weight. Then, compute the geodesic distance as the accumulated edge weights along their shortest path on the graph. Assuming the computed geodesic distance matrix is $\mathcal{D}_{geo}$, the weighted edge similarity of superpixel p and superpixel q, say $e_{p,q}$, is computed as
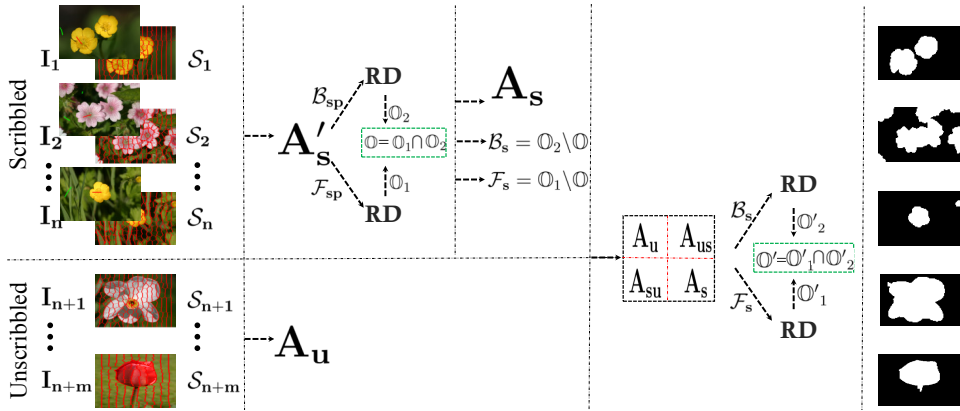
Figure 2.7: *Overview of our interactive co-segmentation algorithm.*

$$e_{p,q} = \begin{cases} 0, & \text{if } p \text{ and } q \text{ are not adjacent,} \\ max(\mathcal{D}_{geo}) - \mathcal{D}_{geo}(p,q) + min(\mathcal{D}_{geo}), & \text{otherwise} \end{cases} \quad (2.1)$$

$A_h^{i \times i}$ for HoG is computed in a similar way while and $A_s^{i \times i}$ for SIFT is built by just keeping adjacent edge similarities.

Assuming we have I images, the final affinity A_γ (γ can be c, s or h in the case of color, SIFT or HOG respectively) is built as

$$A_\gamma = \begin{pmatrix} A_\gamma^{1 \times 1} & \cdots & A_\gamma^{1 \times I} \\ \vdots & \ddots & \vdots \\ A_\gamma^{I \times 1} & \cdots & A_\gamma^{I \times I} \end{pmatrix}$$

As our goal is to segment common foreground objects out, we should consider how related backgrounds are eliminated. As shown in the example image pair of Figure 2.6 (top right), the two images have a related background to deal with it which otherwise would be included as part of the co-segmented objects. To solve this problem we borrowed the idea from [151] which proposes a robust background measure, called boundary connectivity. Given a superpixel $\mathcal{SP}_i$, it computes, based on the background measure, the backgroundness probability $\mathcal{P}_b^i$. We compute the probability of the superpixel being part of an object $\mathcal{P}_f^i$ as its additive inverse, $\mathcal{P}_f^i = 1 - \mathcal{P}_b^i$. From the probability $\mathcal{P}_f$ we built a score affinity A_m as

$$A_m(i,j) = \mathcal{P}_f^i * \mathcal{P}_f^j$$

2.3.2 Optimization

We model the foreground object extraction problem as the optimization of the similarity values among all image superpixels. The objective function is designed to assign the object region a positive membership score and the background region zero membership score, respectively. The optimal object region is then obtained by maximizing the objective function. Let the membership score of N superpixels be $\{x_i\}_{i=1}^{N}$, the (i, j) entry of a matrix A_z is z_{ij}.

Our objective function, combining all the aforementioned terms (A_c, A_s, A_h and A_m), is thus defined, based on equation (3.3), as:

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{ \frac{1}{2}\underbrace{x_i x_j m_{ij}}_{\text{objectness score}} + \frac{1}{6}x_i x_j \underbrace{(c_{ij} + s_{ij} + h_{ij})}_{\text{feature similarity}} - \alpha x_i x_j \right\} \qquad (2.2)$$

The parameter α is fixed based on the (non-)constraint set of the nodes. For the case of unsupervised co-segmentation, the nodes of the pairs of images are set (interchangeably) as constraint set where the intersection of the corresponding results give us the final co-segmented objects.

In the interactive setting, every node i (based on the information provided by the user) has three states: $i \in FGL$, (i is labeled as foreground label), $i \in BGL$ (i is labeled as background label) or $i \in V \backslash (FGL \cup BGL)$ (i is unlabeled). Hence, the affinity matrix $A = (a_{ij})$ is modified by setting a_{ij} to zero if nodes i and j have different labels (otherwise we keep the original value).

2.3.3 Experiments and results

To evaluate the effectiveness of our approach, we conducted extensive experiments on standard benchmark datasets that are widely used to evaluate co-segmentation algorithms, namely, image pairs [59] and MSRC [104]. The image pairs dataset consists of 210 images (105 image pairs) of different animals, flowers, human objects, buses, etc. Each image pair contains one or more similar objects. Some of them are relatively simple and others include complex image pairs containing foreground objects with higher appearance variations or low-contrast objects with complex backgrounds.

The MSRC dataset contains 14 categories with 418 images in total. We evaluated our interactive co-segmentation algorithm on nine selected object classes of MSRC dataset (bird, car, cat, chair, cow, dog, flower, house, sheep), which contains 25~30 images per class. We put foreground and
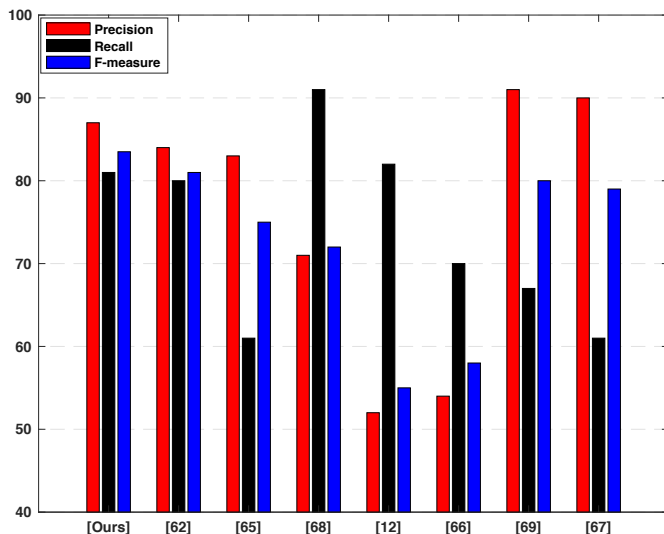
Figure 2.8: *Precision, Recall and F-Measure of our unsupervised co-segmentation algorithm and other state-of-the art approaches on the image pair dataset.*
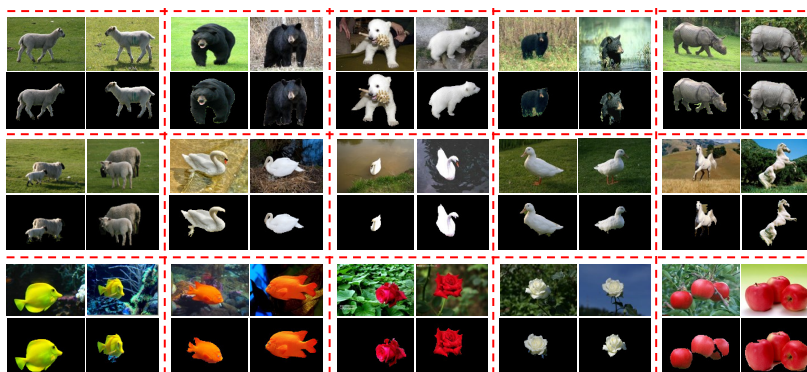
Figure 2.9: *Some qualitative results of our unsupervised method tested on the image pair dataset. Upper row: Original image Lower row: Result of the proposed unsupervised algorithm.*

background scribbles on 15~20 images per class. Each image was over-segmented to 78~83 SLIC superpixels using the VLFeat toolbox.

In order to directly compare the performance of our algorithm with state-of-the-art approaches, in the experiments reported here we used precision, recall and F-measure, which were computed based on the output mask and the ground-truth segmentation. Precision is calculated as the ratio of correctly detected objects to the number of detected object pixels, while recall is the ratio of correctly detected object pixels to the number of ground truth pixels. The F-measure is computed as customary as

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$$

where we set $\beta^2 = 0.3$ as in [56, 58, 59].

We have applied Biased Normalized Cut (BNC) [96] on co-segmentation problem on the MSRC dataset by using the same similarity matrix we used to test our method, and the comparison result of each object class is shown in Figure 2.10. As can be seen, our method significantly surpasses BNC and [153] in average F-measure. Furthermore, we have tested our interactive co-segmentation method, BNC and [153] on image pairs dataset by putting scribbles on one of the two images. As can be observed from Table 3.1, our algorithm substantially outperforms BNC and [153] in precision and F-measure (the recall score being comparable among the three competing algorithms).

We have also examined our unsupervised co-segmentation algorithm by using image pairs dataset, the barplot in Figure 2.8 shows the quantitative result of our algorithm comparing to the state-of-the-art methods [56, 80, 152]. As shown here, our algorithm achieves the best F-measure comparing to all other state-of-the-art methods. The qualitative performance of our unsupervised algorithm is shown in Figure 2.9 on some example images taken from image pairs dataset. As can be seen, our approach can effectively detect and segment the common object of the given pair of images.

Finally, to assess the robustness of our interactive co-segmentation algorithm we conducted the following experiment on the MSRC dataset. We first generated random noise-scribbles by flipping 10~20 superpixel labels from foreground to background and vice versa. We then randomly selected from 0%, to 100% erroneous superpixels from the noise-scribbles and ran our algorithm. As can be observed from Figure 2.10(left) our algorithm performs consistently well, thereby confirming the behavior observed with error-tolerant scribble-based segmentation (Section 5.1.1), and previous ex-

Metrics	$Precision$	$Recall$	$F-measure$
[153]	0.5818	0.8239	0.5971
BNC	0.6421	**0.8512**	0.6564
Ours	**0.7076**	0.8208	**0.7140**

Table 2.4: *Results of our interactive co-segmentation method on Image pair dataset putting scribbles on one of the image pairs.*
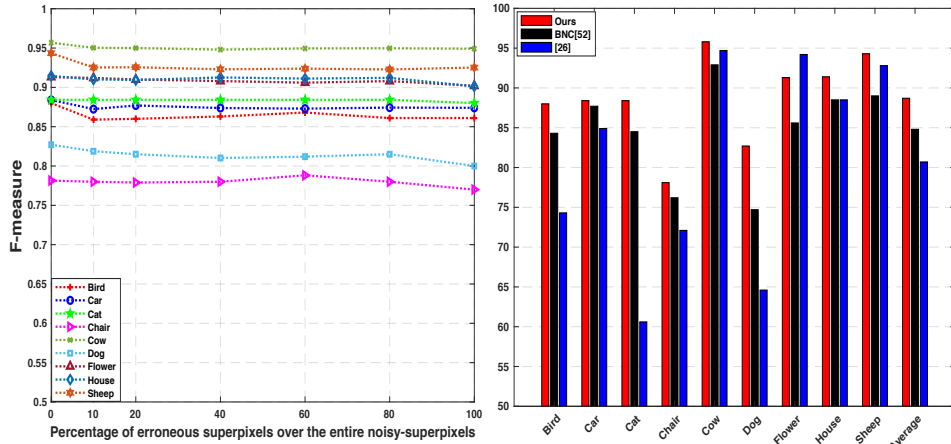
Figure 2.10: *Left: Performance of our interactive image co-segmentation framework with different percentage of erroneous superpixels. **Right:** F-Measure based performance comparison of our interactive co-segmentation method with state-of-the-art methods on MSRC dataset.*

perimental findings with dominant sets [103, 116, 119].

2.4 Summary

In this chapter, we have introduced the notion of a *constrained dominant set* and have demonstrated its applicability to problems such as interactive image segmentation and co-segmentation (in both the unsupervised and the interactive flavor). In our perspective, these can be thought of as "constrained" segmentation problems involving an external source of information (being it, for example, a user annotation or a collection of related images to segment jointly) which somehow drives the whole segmentation process.

The approach is based on some properties of a family of quadratic optimization problems related to dominant sets which show that, by properly

selecting a regularization parameter that controls the structure of the underlying function, we are able to "force" all solutions to contain the constraint elements. The proposed method is flexible and is capable of dealing with various forms of constraints and input modalities, such as scribbles and bounding boxes, in the case of interactive segmentation. Extensive experiments on benchmark datasets have shown that our approach considerably improves the state-of-the-art results on the problems addressed. This provides evidence that constrained dominant sets hold promise as a powerful and principled framework to address a large class of computer vision problems formulable in terms of constrained grouping. Indeed, we mention that they are already being used in problems such as content-based image retrieval [169], multi-target tracking [139] and image geo-localization [168].

The kind of constraints we dealt with in this thesis might be called first-order, or unary, positive constraints, as they refer to a situation whereby one wants to *include* one or more given vertices into the final cluster solution. Of course, other types of constraints can be (and have indeed been) considered when doing clustering. Using the terminology introduced above, for example, one might want to enforce (first-order) negative constraints, according to which certain vertices have to be *excluded* from the extracted cluster. This situation can easily be addressed within our framework by simply setting the initial values of the replicator dynamics corresponding to the to-be-excluded vertices to zero. Note that by combining unary (negative and positive) constraints, and by employing the simple peel-off strategy described in [103] to extract multiple clusters, it would be straightforward to generalize the proposed framework to multi-cluster versions of the problem involving multi-label seeds. Second-order constraints, better known in the literature as must-link and cannot-link constraints [31, 47, 78, 165], can also be easily incorporated. For example, a cannot-link constraint which prescribes that vertices i and j should not be part of the same cluster, can be enforced by setting $a_{ij} = 0$ in the affinity matrix. In fact, a result proven in [7] shows that, by doing so, no dominant set can contain both vertices. Similarly, must-link constraints might be enforced by setting a_{ij} to a sufficiently large value, e.g., the maximum entry in the affinity matrix (see [74] for a similar idea).

The reason why in this work we focused primarily on first-order positive constraints is that, despite their simplicity, they allow us to address in a unified manner various well-known segmentation settings which have been traditionally treated separately in the literature. Of course, the combination of various forms of pairwise or higher-order constraints might give rise to a more general and flexible segmentation framework, and it is our plan to

investigate these ideas in our future work.

Constrained Dominant Sets for Image Retrieval

3.1 Multi-features Fusion Using Constrained Dominant Sets for Image Retrieval

Aggregating different image features for image retrieval has recently shown its effectiveness. While highly effective, though, the question of how to up-lift the impact of the best features for a specific query image persists as an open computer vision problem. In this paper, we propose a computationally efficient approach to fuse several hand-crafted and deep features, based on the probabilistic distribution of a given membership score of a constrained cluster in an unsupervised manner. First, we introduce an incremental near-est neighbor (NN) selection method, whereby we dynamically select k-NN to the query. We then build several graphs from the obtained NN sets and employ constrained dominant sets (CDS) on each graph G to assign edge weights which consider the intrinsic manifold structure of the graph, and detect false matches to the query. Finally, we elaborate the computation of feature positive-impact weight (PIW) based on the dispersive degree of the characteristics vector. To this end, we exploit the entropy of a clus-
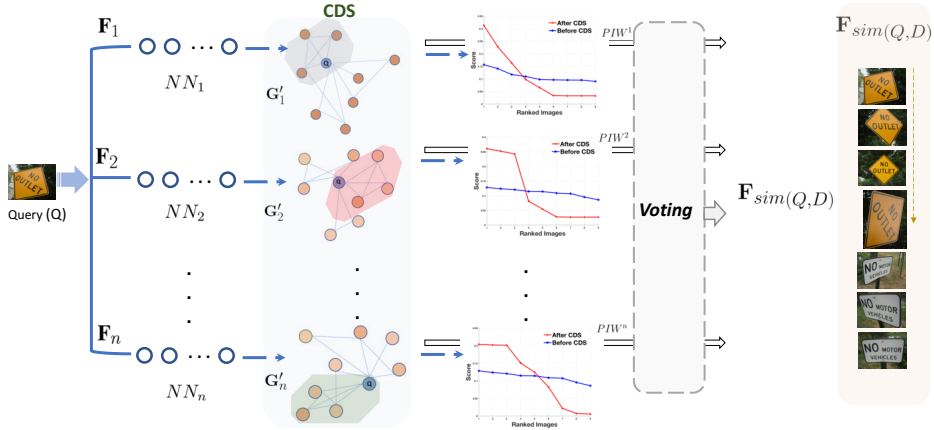
Figure 3.1: *Overview of the proposed image retrieval framework. Based on the given features, $F_1, F_2, ...F_n$, we first incrementally collect the $NN's$ to the query Q, denoted as $NN_1, NN_2, ...NN_n$. Next, for each NN we build the corresponding graph $G'_1, G'_2, ...G'_n$, and then, we apply CDS on each graph to learn the PIW of each feature, $PIW_1, PIW_2, ...PIW_n$, in the subsequent plot, the blue and red curves depict the ranked score of NN's before and after the application of CDS, respectively. Following, the final candidates, which come from each feature, pass through a voting scheme. Finally, using the obtained votes and PIW's we compute the final similarity, $F_{sim}(Q, D)$, between the query and the dataset images by equ. 3.10 .*

ter membership-score distribution. In addition, the final NN set bypasses a heuristic voting scheme. Experiments on several retrieval benchmark datasets show that our method can improve the state-of-the-art result.

3.1.1 Introduction

The goal of semantic image search, or content-based image retrieval (CBIR), is to search for a query image from a given image dataset. This is done by computing image similarities based on low-level image features, such as color, texture, shape and spatial relationship of images. Variation of images in illumination, rotation, and orientation has remained a major challenge for CBIR. Scale-invariant feature transform (SIFT) [91] based local feature such as Bag of words (BOW) [128], [65], [163], has served as a backbone for most image retrieval processes. Nonetheless, due to the inefficiency of using only a local feature to describe the content of an image, local-global feature fusion has recently been introduced.

Multi-feature based CBIR attacks the CBIR problem by introducing an approach which utilizes multiple low-level visual features of an image. Intuitively, if the to-be-fused feature works well by itself, it is expected that its

aggregation with other features will improve the accuracy of the retrieval. Nevertheless, it is quite hard to learn in advance the effectiveness of the to-be-fused features for a specific query image. Different methods have recently been proposed to tackle this problem [158], [174], [180]. Zhang et al. [174] developed a graph-based query specific fusion method, whereby local and global rank lists are merged with equal weight by conducting a link analysis on a fused graph. Zheng et al. [180] proposed a score level fusion model called Query Adaptive Late Fusion (QALF) [180], in which, by approximating a score curve tail with a reference collected on irrelevant data, they able to estimate the effectiveness of a feature as negatively related to the area under the normalized curve. Yang *et al.* [158] used a mixture Markov model to combine given graphs into one. Unlike [174] where graphs are equally weighted, [158] proposed a method to compute a weight which quantifies the usefulness of the given graph based on a naive Bayesian formulation, which depends only on the statistics of image similarity scores.

However, existing multi-feature fusion methods have different drawbacks. For instance, [180], [174], [36], [175] heavily rely on a pre-calculated and offline stored data, which turns out to be computationally expensive when new images are constantly added to the dataset. On the other hand, Ensemble Diffusion (ED) [18] requires $O(n^3)$ to perform a similarity diffusion. In addition to that, its feature-weight computation approach is not a query specific.

Inspired by [180], in this work we present a novel and simple CBIR method based on a recently introduced constrained cluster notion. Our approach presents two main advantages. Firstly, compared to the state of the art methods, it can robustly quantify the effectiveness of features for a specific query, without any supervision. Secondly, by diffusing the pairwise similarity between the nearest neighbors, our model can easily avoid the inclusion of false positive matches in the final shortlist. Towards this end, we first dynamically collect the nearest neighbors to the query, therefore, for each feature, we will have a different number of NNs. Subsequently, we set up the problem as finding a cluster from the obtained NNs, which is constrained to contain the given query image. To this end, we employ a graph-theoretic method called constrained dominant sets [172]. Here is our assumption: if the nearest neighbor to the query image is a false match, after the application of CDS its membership score to the resulting constrained cluster should be less than the fixed threshold ζ, which leads us to detect and exclude the outliers. Furthermore, we introduce the application of entropy to quantify the effectiveness of the given features based on the obtained

membership score. In contrast to [180], our method does not need any reference or external information to learn a query specific feature-weight. Fig. 3.1 shows the pipline of the proposed method.

In particular, we make the following contributions. 1) Compared to the previous work [174], [180], we propose a simple but efficient entropy-based feature effectiveness weighting system; in addition to that, we demonstrate an effective way of outlier or false nearest neighbor detection method. 2) Most importantly, our proposed model is a generic approach, which can be adapted to distinct computer vision problems, such as object detection and person re-identification. 3) We show that our unsupervised graph fusion model easily alleviates the asymmetry neighborhood problem.

This chapter is structured as follows. In section 2 we briefly survey literature relevant to our problem, followed by technical details of the proposed approach in Sec. 3. And, in Sec. 4 we show the performance of our framework on different benchmark datasets.

3.1.2 Related Work

CBIR has become a well-established research topic in the computer vision community. The introduction of SIFT feature plays a vital role in the application of BOW model on the image retrieval problem. Particularly, its robustness in dealing with the variation of images in scale, translation, and rotation provide a significant improvement in the accuracy of similar image search. Sivic et al. [128] first proposed BOW-based image retrieval method by using SIFT, in that, local features of an image are quantized to visual words. Since then, CBIR has made a remarkable progress by incorporating k-reciprocal neighbor [114], query expansion [32], [114], [140], large visual codebook [110], [14], diffusion process [158] [170], [137] and spacial verification [110]. Furthermore, several methods, which consider a compact representation of an image to decrease the memory requirement and boost the search efficiency have been proposed. Jegou et al. [69] developed a Vector of Locally Aggregated Descriptor(VLAD), whereby the residuals belonging to each of the codewords are accumulated.

While SIFT-based local features have considerably improved the result of image search, it does not leverage the discriminative information encoded in the global feature of an image, for instance, the color feature yields a better representation for smooth images. This motivates the introduction of multiple feature fusion for image retrieval. In [174], a graph-based query specific fusion model has been proposed, in which multiple graphs are combined and re-ranked by conducting a link analysis on a fused graph.

Following, [158] developed a re-ranking algorithm by fusing multi-feature information, whereby they apply a locally constrained diffusion process (LCDP) on the localized NNs to obtain a consistent similarity score.

Although the aggregation of handcrafted local and global features has shown promising results, the advent of a seminal work by A.Krizhevsky *et al.* [77] in 2012 changed the focus of the computer vision community. Since then, convolutional neural network (CNN) feature has been used as a main holistic cue in different computer vision problems, including CBIR. Despite its significant improvement on the result of image retrieval, CNN feature still can not endow the demanded accuracy on different benchmark retrieval datasets, especially without the use of fine-tuning. Thus, aggregating graphs which are built from a hand-engineered and CNN-based image features has shown improvement in the accuracy of the retrieval [128], [69], [109], [70], [39], [72].

In addition to that, Yang *et al.* [158] applied a diffusion process to understand the intrinsic manifold structure of the fused graph. Despite a significant improvement on the result, employing the diffusion process on the final (fused) graph restricts the use of the information which is encoded in the pairwise similarity of the individual graph. Instead, our proposed framework applies CDS on each graph which is built from the corresponding feature. Thus, we are able to propagate the pairwise similarity score throughout the graph. Thereby, we exploit the underutilized pairwise similarity information of each feature and alleviate the negative impact of the inherent asymmetry of a neighborhood.

3.1.3 Proposed Method

3.1.3.1 Incremental NN Selection

In this subsection, we show an incremental nearest neighbor collection method to the given query image. We start with an intuitive clustering concept that similar nodes with common features should have an approximate score distribution, while outliers, or nodes which do not belong to a similar semantic class, have different score values. Accordingly, we propose a technique to search for the transition point where our algorithm starts including the outlier nodes. To this end, we examine how distinctive two subsequent nodes are in a ranked list of neighbors. Thus, we define a criterion called neighbors proximity coefficient(NPC), which is defined as the ratio of two consecutive NNs in the given ranked list. Therefore, images are added only if the specified criterion is met, which is designed in such a way that only images that are very likely to be similar to the query image

are added. Thereby, we are able to decrease the number of false matches to the query in the k-nearest neighbors set.

Given an initial ranked list R. And then, we define top-k nearest neighbors (kNN) to query Q as

$$kNN(q,k) = \begin{cases} \text{Add } n_i & if \quad \frac{Sim(q,n_{i+1})}{Sim(q,n_i)} > \text{NPC} \\ 0 & otherwize \end{cases} \quad (3.1)$$

where $|kNN(q,k)| = k$, and $|.|$ represents the cardinality of a set.

$$kNN(q,k) = \{n_1, n_2, ...n_k\}, \quad where\ kNN(q,k) \subseteq R \quad (3.2)$$

3.1.3.2 Graph Construction

Different features, $F = F_1, F_2...F_n$, are extracted from images in the dataset D and the query image Q, where each feature encodes discriminative information of the given image in different aspects. We then compute the distance between the given images based on a distance metric function $d'(I_i, I_j)$, where I_i and I_j denote the given feature vector extracted from image i and j respectively. Following, we compute symmetric affinity matrices $A'_1, A'_2, ... A'_n$ from each distance matrix D_i using a similarity function $S(D_i)$. We then apply minimax normalization on each similarity matrix as: $A_i = \frac{V_\alpha^{ij} - min(V_\alpha)}{max(V_\alpha) - min(V_\alpha)}$, where V_α is a column vector taken from matrix A'_i, which comprises the pairwise similarity score between a given image V_α^i and images in the dataset V^j, which is denoted as V_α^{ij}. Next, we build undirected edge-weighted graphs with no self-loops $G_1, G_2...G_n$ from the affinity matrices $A_1, A_2, ...A_n$, respectively. Each graph G_n is defined as $Gn = (V_n, E_n, w_n)$, where $V_n = 1, ..., n$ is vertex set, $E_n \subseteq V_n \times V_n$ is the edge set, and $w_n : E \longrightarrow \mathbb{R}_+^*$ is the (positive) weight function. Vertices in G correspond to the given images, edges represent neighborhood relationships, and edge-weights reflect similarity between pairs of linked vertices.

3.1.3.3 PIW Using Entropy of CDS

Since the nearest neighbor selection method heavily relies on the initial pairwise similarity, it is possible that the NN set easily includes false matches to the given query. This usually happens due to the lack of technics which consider the underlying structure of the data manifold, especially the inherent asymmetry of a neighborhood is a major shortcoming of such systems.

For instance, although $Sim(n_i, q) = Sim(q, n_i)$, the nearest neighbor relationship between query Q and image n_i may not be symmetric, which implies that $m_i \in kNN(q, k)$ but $m_i \notin kNN(n_i, k)$. As demonstrated in the past retrieval works, the k-reciprocal neighbors [114] and similarity diffusion process [64] have been vastly taken as the optimal options to tackle this issue. However, the existing methods are not computationally efficient. In this work, we remedy the existing limitations using an unsupervised constrained clustering algorithm whereby we exploit the pairwise similarity to find a cohesive cluster which incorporates the specified query.

3.1.3.4 Constrained Clustering for Coherent Neighbor Selection

Towards collecting true matches to the query image, we employ an unsupervised clustering algorithm on the top of the previous steps. Our hypothesis to tackle the asymmetry problem between the given query and its nearest neighbors is that images which are semantically similar to each other tend to be clustered in some feature space. As can be seen in the synthetic example (See Fig. 3.2), retrieved image i_4 and i_6 are outliers or false positives to the query image Q. We can confirm this by observing the common neighbors of Q with i_4 and i_6. But due to the lack of contextual information, the system considers them as a true match (neighbor) to the query. In our proposed model, to attack this issue, we represent the set of kNN as a graph G' accordingly to subsection 3.1.3.2. Then, we treat outliers finding problem as an unsupervised clustering problem. We first convert graph G' into a symmetric affinity matrix A, where the diagonal corresponding to each node is set to 0, and the $ij - th$ entry denotes the edge-weight w_{ij} of the graph so that $A_{ij} \equiv A_{ji}$. Accordingly, given graph G' and query Q, we cast detecting outliers from a given NN set as finding the most compact and coherent cluster from graph G', which is constrained to contain the query image Q. To this end, we adopt constrained dominant sets [172], [99] which is a generalization of a well known graph-theoretic notion of a cluster. We are given a symmetric affinity matrix A and parameter $\mu > 0$, and then we define the following parametrized quadratic program

$$
\begin{aligned}
\text{maximize} \quad & f_Q^\mu(X) = X'(A - \mu\hat{\Gamma}_Q)X \\
& f_Q^\mu(X) = X'\hat{A}X \\
\text{subject to} \quad & X \in \Delta
\end{aligned}
\tag{3.3}
$$

where a prime denotes transposition and

$$\Delta = \left\{ X \in R^n \ : \ \sum_{i=1}^{n} X_i = 1, \text{ and } X_i \geq 0 \text{ for all } i = 1 \ldots n \right\}$$

Δ is the standard simplex of R^n. $\hat{\Gamma}_Q$ represents $n \times n$ diagonal matrix whose diagonal elements are set to zero in correspondence to the query Q and to 1 otherwise. And $\hat{A}$ is defined as,

$$\hat{A} = A - \mu\hat{\Gamma}_Q = \begin{pmatrix} 0 & . & . & . \\ . & -\mu & . & . \\ . & . & -\mu & . \\ . & . & . & -\mu \end{pmatrix}$$

where the dots denote the ij th entry of matrix A. Note that matrix $\hat{A}$ is scaled properly to avoid negative values.

Let $Q \subseteq V$, with $Q \neq \emptyset$ and let $\mu > \lambda_{max}(A_{V \setminus Q})$, where $\lambda_{max}(A_{V \setminus q})$ is the largest eigenvalue of the principal submatrix of A indexed by the element of $V \setminus q$. If X is a local maximizer of $f_Q^\mu(X)$ in Δ, then $\delta(X) \cap Q \neq \emptyset$, where, $\delta(X) = i \in V : X_i > 0$. We refer the reader to [172] for the proof.

The above result provides us with a simple technique to determine a constrained dominant set which contains the query vertex Q. Indeed, if Q is the vertex corresponding the query image, by setting

$$\mu > \lambda(A_{V \setminus Q}) \tag{3.4}$$

we are guaranteed that all local solutions of eq (3.3) will have a support that necessarily contains the query element. The established correspondence between dominant set (coherent cluster) and local extrema of a quadratic form over the standard simplex allow us to find a dominant set using straightforward continuous optimization techniques known as replicator dynamics, a class of dynamical systems arising in evolutionary game theory [21]. The obtained solution provides a principled measure of a cluster cohesiveness as well as a measure of vertex participation. Hence, we show that by fixing an appropriate threshold ζ on the membership score of vertices, to extract the coherent cluster, we could easily be able to detect the outlier nodes from the k-nearest neighbors set. For each X^i, ζ^i is dynamically computed as

$$\zeta^i = \Lambda(1 - max(X^i) + min(X^i))/L \tag{3.5}$$

where $max(X)$ and $min(X)$ denote the maximum and minimum membership score of X^i, respectively. Λ is a scaling parameter and L stands for length of X^i. Moreover, we show an effective technique to quantify the usefulness of the given features based on the dispersive degree of the obtained characteristics vector X.

3.1.3.5 PIW Using Entropy of Constrained Cluster.

Entropy has been successfully utilized in a variety of computer vision applications, including object detection [133], image retrieval [37] and visual tracking [95]. In this chapter, we exploit the entropy of a membership-score of nodes in the constrained dominant set to quantify the usefulness of the given features. To this end, we borrowed the concept of entropy in the sense of information theory (Shannon entropy). We claim that the discriminative power of a given feature is inversely proportional to the entropy of the score distribution, where the score distribution is a stochastic vector. Let us say we are given a random variable C with possible values $c_1, c_2, ...c_n$, according to statistical point of view the information of the event $(C = c_i)$ is inversely proportional to its likelihood, which is denoted by $I(C_i)$ and defined as

$$I(C_i) = log\Big(\frac{1}{P(c_i)}\Big) = -log(p(c_i)). \qquad (3.6)$$

Thus, as stated by [123], the entropy of C is the expected value of I, which is given as

$$H(C) = -\sum_{i=1}^{N} P(c_i)log(P(c_i). \qquad (3.7)$$

For each characteristic vector $X^i, X^{i+1}...X^z$, where $X^i = \{X_\mu^i, X_{\mu+1}^i...X_n^i\}$, we compute the entropy $H(exp(X^i))$. Each X^i corresponds to the membership score of nodes in the CDS, which is obtained from the given feature F^i. Assume that the top NNs obtained from feature x are irrelevant to the query Q, thus the resulting CDS will only contain the constraint element Q. Based on our previous claim, since the entropy of a singleton set is 0, we can infer that the feature is highly discriminative. Although this conclusion is right, assigning a large weight to feature with irrelevant NNs will have a negative impact on the final similarity. To avoid such unintended impact, we consider the extreme case where the entropy is 0. Following, we introduce a new term C_a, which is obtained from the cardinality of a given cluster, K_c, as $Ca^i = \frac{K_c^i}{\sum_{i=1}^{z} K_c^i}$. As a result, we formulate the PIW computa-
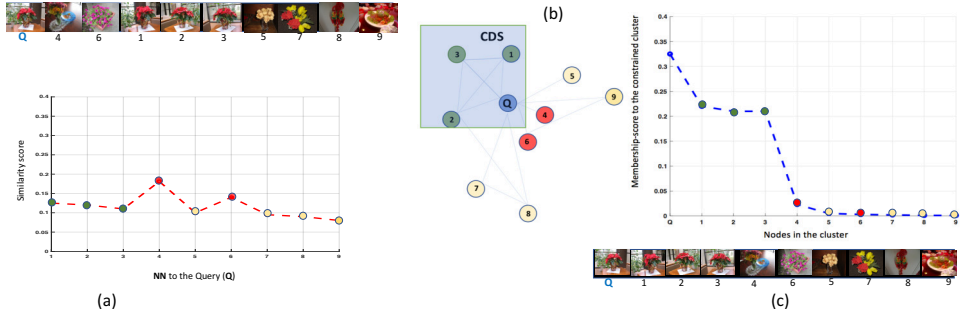
Figure 3.2: *(a) Initial score distribution of the top k nearest neighbors to the query Q, green and red points denote the false-negative and false-posetive NNs. (b) Graph G', built from the initial pairwise similarity of the k-nearest neighbor set. And the blue box contains the CDS nodes which are obtained by running CDS on graph G'. (c) The resulting constrained dominant set membership-score distribution.*

tion from the additive inverse of the entropy $\varepsilon^i = 1 - H(X^i)$, and C_a^i, as:

$$PIW^i = \frac{\vartheta^i}{\sum_{i=1}^z \vartheta^i} \quad Thus, \sum_{i=1}^z PIW^i = 1 \tag{3.8}$$

where $\vartheta^i = \varepsilon^i + C_a^i$, and i represents the corresponding feature.

3.1.3.6 Naive Voting Scheme and Similarity Fusion

In this section, we introduce a simple yet effective voting scheme, which is based on the member nodes of k-nearest neighbor sets and the constrained dominant sets, let $NN_1, NN_2...NN_z$ and $CDS_1, CDS_2...CDS_z$ represent the NN and CDS sets respectively, which are obtained from $G'_1, G'_2...G'_z$. Let us say $\xi = 2(z-1) - 1$, and then we build ξ different combinations of NN sets, $\varphi_1, \varphi_2...\varphi_\xi$. Each φ represents an intersection between $z-1$ unique combinations of NN sets. We then form a super-set ϖ which contains the union of φ sets, with including repeated nodes. Assume that $NNs = \{NN_1, NN_2, NN_3\}, \xi = 3$, thus each φ set contains the intersection of two NN sets as $\varphi_1 = \{NN_1 \cap NN_2\}, \varphi_2 = \{NN_1 \cap NN_3\}$ and $\varphi_3 = \{NN_2 \cap NN_3\}$. Hence the resulting ϖ is defined as $\varpi = (\varphi_1 \ominus \varphi_2 \ominus \varphi_3)$, where $(.\ominus.)$ is an operator which returns the union of given sets, including repeated nodes. We have also collected the union of CDS sets as $\omega = (CDS_1 \ominus CDS_2 \ominus CDS_3)$. Following, we compute κ as $(\kappa = \varphi_1 \cap \varphi_2 \cap ...\varphi_\xi)$. Thereby we find super-sets ϖ, ω and κ. Next, we design three different counters, which are formulated to increment when the NN node appears in the corresponding super-

sets. Based on the value obtained from each counter, we finally compute the vote scores for each NN node to the query as $v_1 = v_1/\eta, v_2 = v_2/\theta$ and $v_3 = v_3/\iota$, where η, θ and ι are parameters which are fixed empirically. Note that the outlier detecting capability of our framework is encoded in the voting process. Thus, if a NN node n_i is contained in more than one cluster, its probability to be given a large weight is higher. This is due to the number of votes it gets from each cluster.

3.1.3.7 Final Similarity.

After obtaining the aforementioned terms, we compute the final similarity as follows: say we are given n features, Q is the query image and D denotes image dataset, then the initial similarity of D to Q, with respect to feature $F_i, i = 1...n,$,is given as $S_{D,Q}^{(i)}$. Let $PIW_Q^{(i)}, i = 1...n$, encode the weight of feature F_i for query Q, and then the final similarity score, $F_{sim(Q,D)}$, between Q and D is given as

$$N_s = \prod_{i=1}^{k}(S_{D,Q}^{(i)})^{PIW_Q^{(i)}} \tag{3.9}$$

$$F_{sim(Q,D)} = \lambda N_s + (1 - \lambda)\sum_{\Omega=1}^{\Psi} v_\Omega \tag{3.10}$$

where $\Psi = 3$, is the total number of voter sets. And $\lambda \in [0, 1]$ defines the penalty factor which penalizes the similarity fusion, when $\lambda = 1$ only F_s is considered, otherwise, if $\lambda = 0$, only v is considered.

3.1.4 Experiments

In this section, we present the details about the features, datasets and evaluation methodology we used along with rigorous experimental analysis.

3.1.4.1 Datasets and Metrics

To provide a thorough evaluation and comparison, we evaluate our approach on INRIA Holiday, Ukbench, Oxford5k and Paris6k datasets.

Ukbench Dataset [105]. Contains 10,200 images which are categorized into 2,550 groups, each group consists of three similar images to the query which undergo severe illumination and pose variations. Every image in this dataset is used as a query image in turn while the remaining images are considered as dataset images, in "leave-one-out" fashion. As customary,

Figure 3.3: *five relevant images to the query where the green and red frame indicate the True and False posetives to the query, respectively. **Top-row (a) and (b):** show the top five relevant images of our proposed method. **Bottom row (a) and (b):** show the top five relevant images obtained from a Naive fusion of several features.*

we used the N-S score to evaluate the performance of our method, which is based on the average recall of the top 4 ranked images.

INRIA Holiday Dataset [67]. Comprises 1491 personal holiday pictures including 500 query images, where most of the queries have one or two relevant images. Mean average precision (MAP) is used as a performance evaluation metric.

Oxford5k Dataset [110]. It is one of the most popular retrieval datasets, which contains 5062 images, collected from flicker-images by searching for landmark buildings in the Oxford campus. 55 queries corresponding to 11 buildings are used.

Paris6k Dataset [111]. Consists of 6392 images of Paris landmark buildings with 55 query images that are manually annotated.

3.1.4.2 Image Features

Object Level Deep Feature Pooling (OLDFP) [101]. OLDFP is a compact image representation, whereby images are represented as a vector of pooled CNN features describing the underlying objects. Principal Component Analysis (PCA) has been employed to reduce the dimensionality of the compact representation. We consider the top 512-dimensional vector in the case of the Holiday dataset while considering the top 1024-dimensional vector to describe images in the Ukbench dataset. As suggested in [101], we have applied power normalization (with exponent 0.5), and l2 normalization on the PCA projected image descriptor.

44

BOW. Following [179], [180], we adopt Hamming Embedding [67]. SIFT descriptor and Hessian-Affine detector are used in feature extraction, and we used 128-bit vector binary signatures of SIFT. The Hamming threshold and weighting parameters are set to 30 and 16 respectively, and three visual words are provided for each key-point. Flickr60k data [67] is used to train a codebook of size 20k. We also adopt root sift as in [11], average IDF as defined in [178] and the burstiness weighting [68].

NetVLAD [10]. NetVLAD is an end-to-end trainable CNN architecture that incorporates the generalized VLAD layer.

HSV Color Histogram. Like [158], [180], for each image, we extract 1000-dimensional HSV color histograms where the number of bins for H, S, V are 20, 10, 5 respectively.

Table 3.1: *The performance of baseline features on Holidays, Ukbench, Oxford5k and Paris6k datasets.*

Datasets	Metrics	NetVLAD [10]	BOW	OLDFP	HSV	R_{res} [64]	G_{res} [64]	R_{vgg} [64]	G_{vgg} [64]
Holidays	MAP	84	80	87	65	-	-	-	-
Ukbench	N-S score	3.75	3.58	3.79	3.19		-	-	-
Oxford5k	MAP	69	-	-	-	95.8	87.7	93	-
Paris6k	MAP	-	-	-	-	96.8	94.1	96.4	95.6

3.1.4.3 Experiment on Holiday and Ukbench Datasets

As it can be seen in Fig.3.3(a), the noticeable similarity between the query image and the irrelevant images, in the Holiday dataset, makes the retrieval process challenging. For instance, (See Fig.3.3(a)), at a glance all images seem similar to the query image while the relevant are only the first two ranked images. Moreover, we can observe that the proposed scheme is invariant to image illumination and rotation change. Table 3.2 shows that our method significantly improves the MAP of the baseline method [101] on Holiday dataset by 7.3 % while improving the state-of-the-art method by 1.1 %. Likewise, it can be seen that our method considerably improves the N-S score of the baseline method [101] on the Ukbench dataset by 0.15 while improving the state-of-the-art method by 0.03.

Furthermore, to show how effective the proposed feature-weighting system is, we have experimented by fusing the given features with and without PIW. Naive fusion (NF) denotes our approach with a constant PIW for all features used, thus the final similarity F_s defined as $F_s = \frac{1}{k}(\prod_{i=1}^{k}(S_{D,Q}^{(i)}))$. In Fig.3.6 we have demonstrated the remarkable impact of the proposed
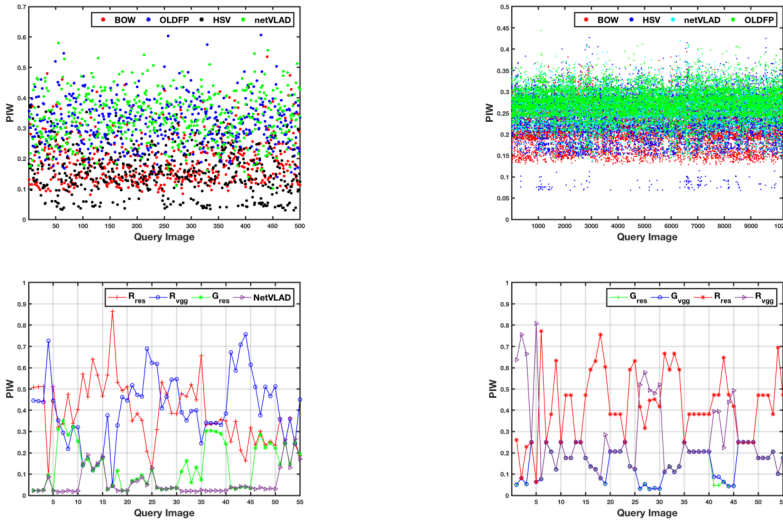
Figure 3.4: *Feature positive-impact weights (PIW's) learned by our algorithm. Top-left, top-right, bottom-left, and bottom-right: on Holiday, Ukbench, Oxford5k and Paris6 datasets, respectively.*

PIW. As can be observed, our scheme effectively uplifts the impact of a discriminative feature while downgrading the inferior one. Note that in the PIW computation we have normalized the minimum entropy (See eq.3.8), thus its values range between 0 and 1. Accordingly, one implies that the feature is highly discriminative, while zero shows that the feature is indiscriminate.

In order to demonstrate that our scheme is robust in handling outliers, we have conducted an experiment by fixing the number of NNs (disabling the incrimental NNs selection) to different numbers. As is evident from Fig.3.6, the performance of our method is consistent regardless of the number of kNN. As elaborated in subsection 3.1.3.4, the robustness of our method to the number of k comes from the proposed outlier detection method. Since the proposed outliers detector is formulated in a way that allows us to handle the outliers, we are easily able to alleviate the false matches which are incorporated in the nearest neighbors set. This results in finding a nearly constant number of nearest neighbors regardless of the choice of k.
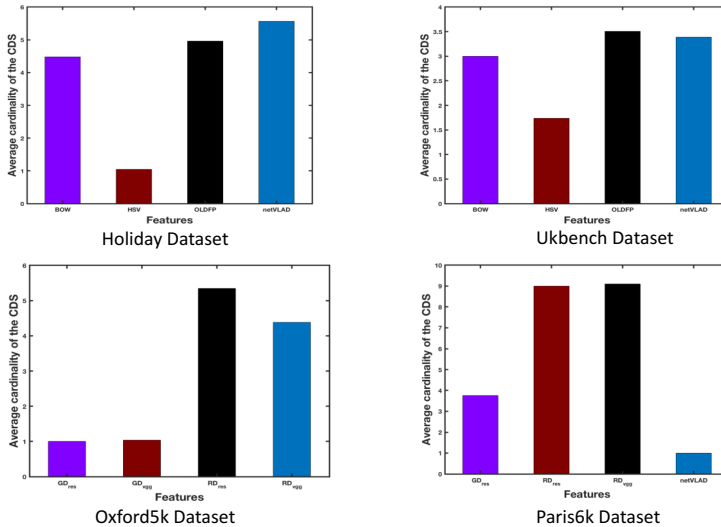
Figure 3.5: *The cardinality of constrained dominant sets for the given features.*

3.1.4.4 Experiment on Oxford5k and Paris6k Datasets

In the same fashion as the previous analysis, we have conducted extensive experiments on the widely used Oxford5k and Paris6k datasets. Unlike the Holiday and Ukbench datasets, we adapt affinity matrices which are obtained through a diffusion process on a regional $Resnet$ and VGG representation [64], and they are denoted as R_{res} and R_{vgg} respectively, as well as affinity matrices G_{res} and G_{vgg} which are also obtained through a diffusion process on a global $Resnet$ and VGG representation, respectively. Table 3.2 shows that the proposed method slightly improves the state-of-the-art result. Even if the performance gain is not significant, our scheme marginally achieves better MAP over the state-of-the-art methods. Furthermore, as shown in Fig 3.4, the proposed model learns the PIW of the given features effectively. Therefore, a smaller average weight is assigned to G_{vgg} and $NetVLAD$ feature comparing to R_{res} and R_{vgg}.

3.1.4.5 Robustness of Proposed PIW

As can be seen in Fig 3.4, for all datasets, our algorithm has efficiently learned the appropriate weights to the corresponding features. Fig. 3.4 shows how our algorithm assigns PIW in a query adaptive manner. In Holiday and Ukbench datasets, the average weight given to HSV feature is much smaller than all the other features used. Conversely, a large PIW is assigned to OLDFP and NetVLAD features. Nevertheless, it is evident that in some

Table 3.2: *Comparison among various retrieval methods with our method on benchmark datasets, where QALF is implemented with the same baseline similarities used in our experiments.*

Datasets	Metrics	Baselines	QALF [180]	[158]	NF	ED [19]	[54]	[115]	[156]	[15]	Ours
Ukbench	N-S score	3.79 [101]	3.84	3.86	3.86	3.93	-	-	-	3.76	**3.94**
Holiday	MAP	87 [101]	88	88	91	93	90	83	89	77	**94**
Oxford5k	MAP	95.8 [64]	-	76.2	94.4	-	89.1	79.7	81.4	67.6	**96.2**
Paris6k	MAP	96.8 [64]	-	83.3	-	-	91.2	83.8	88.9	-	**97.4**
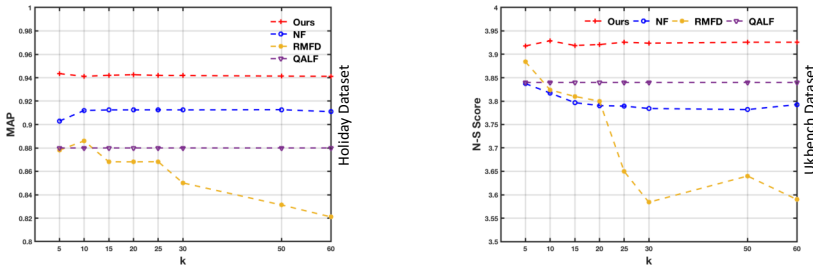
Figure 3.6: *Comparison with state-of-the-art fusion methods with respect to varying k. Naive Fusion (NF), Reranking by Multi-feature Fusion (RMFD) [158], and QALF [180].*

cases a large value of PIW is assigned to HSV and BOW features as well, which is appreciated considering its effectiveness on discriminating good and bad features in a query adaptive manner.

3.1.4.6 Impact of Parameters

To evaluate the robustness of our method we have performed different experiments by changing one parameter at a time. Thereby, we have observed that setting Λ to a large value results in assigning insignificant PIW to indiscriminate features. The reason is that after the application of CDS, the cluster membership-score of the dissimilar images to the query will become smaller. Thus, since the threshold fixed to choose the true neighbors is tighter, the resulting constrained dominant set will be forced to yield a singleton cluster. As a result, we obtained a very small PIW due to the cardinality of the constrained-cluster. In addition to that, we observe that the MAP start to decline when λ is set to a very large value (See. Fig 3.7, right).
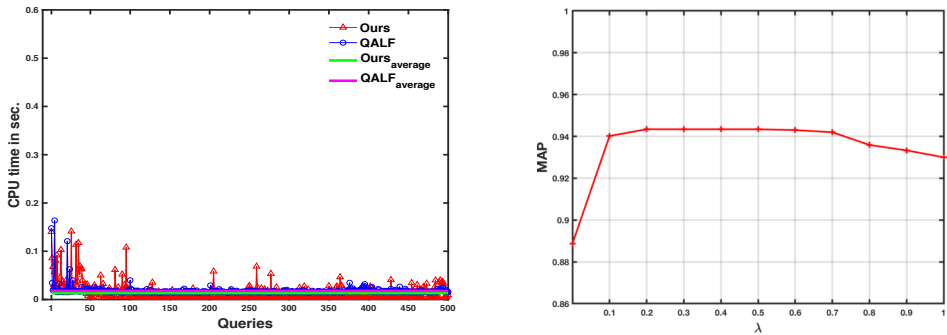
Figure 3.7: *Left: Time complexity of our algorithm (red) and QALF [180] (blue) on Holiday dataset. Right: The impact of λ on the retrieval performance, on Holiday dataset.*

3.1.4.7 Impact of Cluster Cardinality

On the Ukbench dataset, as can be observed in Fig. 3.5, the average cardinality of the constrained clusters which is obtained from HSV and BOW feature is 3 and 1.7, respectively. In contrast, for NetVLAD and OLDFP, the average cluster cardinality is 3.4 and 3.5, respectively . Similarly, in the case of the Holiday dataset, the cluster cardinality obtained from HSV feature is one while for BOW, NetVLAD and OLDFP is 4.5, 5 and 5.6, respectively. Thus, from this, we can draw our conclusion that the cardinality of a constrained dominant set, in a certain condition, has a direct relationship with the effectiveness of the given feature.

3.1.4.8 Computational Time

In Fig. 3.7 we depict the query time taken to search for each query image, red and blue lines represent our method and QALF, respectively. The vertical axis denotes the CPU time taken in seconds, and the horizontal axis shows the query images. As can be seen from the plot, the proposed framework is faster than the fastest state-of-the-art feature-fusion method [180]. As for time complexity, in our experiment we used a replicator dynamics to solve problem (3.3), hence, for a graph with N nodes, the time complexity per step is $O(N^2)$, and the algorithm usually takes a few steps to converge, while that of [18] is $O(N^3)$. However, we note that by using the Infection-immunization algorithm [118] we can achieve even faster convergence as its per-step complexity would be linear in the number of nodes.

3.1.5   Summary

In this chapter, we addressed a multi-feature fusion problem in CBIR. We developed a novel and computationally efficient CBIR method based on a constrained-clustering concept. In particular, we showed an efficient way of estimating a positive impact weight of features in a query-specific manner. Thus it can be readily used for feature combination. Furthermore, the proposed scheme is fully unsupervised, and can easily be able to detect false-positive NNs to the query, through the diffused similarity of the NNs. To demonstrate the validity of our method, we performed extensive experiments on benchmark datasets. Besides the improvements achieved on the state-of-the-art results, our method shows its effectiveness in quantifying the discriminative power of given features. Moreover, its effectiveness on feature-weighting can also be exploited in other computer vision problems, such as person re-identification, object detection, and image segmentation.

3.2   Constrained Dominant Sets for Image Retrieval

Learning new global relations based on an initial affinity of the database objects has shown significant improvements in similarity retrievals. Locally constrained diffusion process is one of the recent effective tools in learning the intrinsic manifold structure of a given data. Existing methods, which constrained the diffusion process locally, have problems - manual choice of optimal local neighborhood size, do not allow for intrinsic relation among the neighbors, fix initialization vector to extract dense neighbor - which negatively affect the affinity propagation. We propose a new approach, which alleviate these issues, based on some properties of a family of quadratic optimization problems related to dominant sets, a well-known graph-theoretic notion of a cluster which generalizes the concept of a maximal clique to edge-weighted graphs. In particular, we show that by properly controlling a regularization parameter which determines the structure and the scale of the underlying problem, we are in a position to extract dominant set cluster which is constrained to contain user-provided query. Experimental results on standard benchmark datasets show the effectiveness of the proposed approach.

3.2.1   Similarity Diffusion for Image Retrieval

Retrieval has recently attracted considerable attention within the computer vision community, especially because of its potential applications such as database retrieval, web and mobile image search. Given user provided

query, the goal is to provide as output a ranked list of objects that best reflect the user's intent. Classical approaches perform the task based on the (dis)similarity between the query and the database objects. The main limitation of such classical retrieval approaches is that they does not allow for the intrinsic relation among the database objects.

Recently, various techniques, instead of simply using the pairwise similarity, they try to learn a better similarities that consider manifold structures of the underlying data. Qin *et al.* in [114] try to alleviate the asymmetry problem of the k-nearest neighbor (k-nn) using the notion of k-reciprocal nearest neighbor. In [63] the notion of shared nearest neighbor is used to build secondary similarity measure, which stabilize the performance of the search, based on the primary distance measure. In [44] shape meta-similarity measure, which is computed as the **L1** distance between new vector representation which considers only the k-nn set of similarities fixing all others to 0, was proposed. Choosing the right size of the neighbor is important. In [76], the notion of shortest path was used to built a new affinity for retrieval.

Diffusion process is one of the recent effective tools in learning the intrinsic manifold structure of a given data [40, 161, 162]. Given data, a weighted graph is built where the nodes are the objects and the edge weight is a function of the affinity between the objects. The pairwise affinities are then propagated following structure of the weighted edge links in the graph. The result of the affinity propagation highly depends on the quality of the pairwise similarity [79, 134]. Inaccurate Pairwise similarity results in a graph with much noise which negatively affects the diffusion process. Constraining the diffusion process locally alleviate this issue [40, 134, 162]. Dominant neighbor (DN) and k-nn are two notions used by the recent existing methods to constrain the diffusion process locally [40, 161, 162]. In [40], it has been shown that affinity learning constraining relation of an object to its neighbors effectively improves the retrieval performance and was able to achieve 100 % bull's eye score in the well known MPEG datset. The author of [40] put automatically selecting local neighborhood size (K) as the main limitation of the approach and is still an open problem. The influence of selecting different K values was also studied which proved that the parameter is a serious problem of the approach. For MPEG7 dataset, the choice is insignificant while for the other two datasets YALE and ORL choosing the reasonable K is difficult which resulted in a decrease in performance for the right value of K. Moreover, it is obvious that the selection of k-nn is prone to errors in the pairwise similarities [162]. Since any k-nn decision procedure relies only on affinities of an object to all other objects,

k-nn approach is handicapped in resisting errors in pairwise affinities and in capturing the structure of the underlying data manifold.

Yang *et al.* in [162], to avoid the above issues, proposed the notion of dominant neighbors (DN). Instead of the k-nn, here a compact set from the k-nn which best explains the intrinsic relation among the neighbors is considered to constrain the diffusion process. However, the approach follows heuristic based k-nn initialization scheme. To capture dominant neighbors, the approach first choose a fixed value of K, collect the K nearest neighbors and then initialize the dynamics, the dynamics which extracts dense neighbors, to the barycenter of the face of the simplex which contains the neighbors. It is obvious to see that the approach is still dependent on K. Moreover, as fixing K limits the dynamics to a specified face of the simplex, objects out of k-nn(q) which form a dominant neighbor with q will be loosed. The chosen k-nn may also be fully noisy which might not have a compact structure.

In this section, we propose a new approach to retrieval which can deal naturally with the above problems. Our approach is based on some properties of a family of quadratic optimization problems related to dominant sets, a well-known graph-theoretic notion of a cluster which generalizes the concept of a maximal clique to edge-weighted graphs. In particular, we show that by properly controlling a regularization parameter which determines the structure and the scale of the underlying problem, we are in a position to extract dominant-set cluster which is constrained to contain user-specified query.

The resulting algorithm has a number of interesting features which distinguishes it from existing approaches. Specifically: 1) it is able to constrain the diffusion process locally extracting dense neighbors whose local neighborhood size (K) is fixed automatically; different neighbors can have different value of K. 2)it has no any initialization step; the dynamics, to extract the dense neighbors, can start at any point in the standard simplex 3) it turns out to be *robust* to noisy affinity matrices.

The rest of this chapter is organized as follows. In the next section we will discuss the most related works to our approach. The experimental results are given in section 3.2.3.

3.2.2 Diffusion Process

Given a set of, say n, objects, the relation among them can be represented as an undirected edge-weighted graph $G = (V, E, w)$, where $V = \{1, ..., n\}$ 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, the graph G is represented with the corresponding weighted adjacency (or similarity) matrix, which is the $n \times n$ nonnegative, symmetric matrix $A = (a_{ij})$ defined as $a_{ij} = w(i, j)$, if $(i, j) \in E$, and $a_{ij} = 0$ otherwise. A diffusion process then starts from a predefined initialization, say $\mathcal{V}$ and propagate the affinity value through the underlying manifold based on a predefined transition matrix, say $\mathcal{T}$, and diffusion scheme ($\mathcal{S}$).

Off-the-shelf diffusion processes, which basically differ based on the choice of $\mathcal{V}$, $\mathcal{T}$ and $\mathcal{S}$, the most related ones to this work are [162] and [160]. In both cases, the diffusion process is locally constrained. While in [160] the notion of k-nn is used to constrain the diffusion process locally, dominant neighbor notion ($\mathcal{DN}$) is used by [162].

3.2.2.1 Nearest Neighbors

In the first case, the edge-weights of the k-nn are kept i.e define locally constrained affinity $\mathcal{L} = (l_{ij})$ defined as $l_{ij} = w(i, j)$, if $(i, j) \in$ k-nn(q), and $l_{ij} = 0$ otherwise. Then the diffusion process, setting $\mathcal{V}$ as the affinity A, is performed by the following update rule.

$$\mathcal{V}_{t+1} = \mathcal{L}\mathcal{V}\mathcal{L} \tag{3.11}$$

Nearest neighbors constrained diffusion process, alleviating the issue of noisy pairwise similarity, significantly increases the retrieval performance. However, the approach has two serious limitations: First, automatically selecting local neighborhood size (K) is very difficult and is still an open problem [40]. In [40] the influence of selecting different K values was studied which proved that the parameter is a serious problem of the approach. For MPEG7 dataset, the choice was insignificant while for the other two datasets, YALE and ORL, choosing the reasonable K was difficult which even resulted in a decrease in performance, for ORL from 77.30% to 73.40% and for YALE 77.08% to 73.39%, for the right value of K. Moreover, it is obvious that the selection of k-nn is prone to errors in the pairwise similarities [162].

3.2.2.2 Dominant Neighbors

Yang *et al.* in [162], to avoid the above issues, proposed the notion of dominant neighbors ($\mathcal{DN}$). Instead of the k-nn, here a compact set from the k-nn which best explains the intrinsic relation among the neighbors is

MPEG7	B1	B2	B3	B4	B5	B6(Ours)
A1	99.91	99.93	**100**	**100**	99.88	**100**
A2	99.92	99.93	**100**	**100**	99.88	**100**
A3	99.93	99.94	**100**	**100**	99.88	**100**
A4	99.92	99.94	**100**	**100**	99.88	**100**

Table 3.3: *Results on MPEG7 dataset. Bull's eye score for the first 40 elements*

considered to constrain the diffusion process. To do so, the author used the dominant set framework by Pavan and Pelillo [103].

A dominant neighbor ($\mathcal{DN}$) is set as a dominant set, say $\mathcal{DS}$, from the k-nn which contains the user provided query q, lets call it $\mathcal{DS}(q)$.

3.2.2.3   Proposed Method

Given a query q, we scale the affinity and run the replicator 1.10, say the dynamics converges to $\mathbf{x}^*$. The support of $\mathbf{x}^*$, $\sigma(\mathbf{x}^*)$, is the constrained dominant set which contains the query q, let us call it $\mathcal{CDS}(q)$. The edge-weights of the $\mathcal{CDS}(q)$ are then kept i.e define locally constrained affinity $\mathcal{L} = (l_{ij})$ defined as $l_{ij} = w(i,j)$, if $(i,j) \in \mathcal{CDS}(q)$, and $l_{ij} = 0$ otherwise. The diffusion process is then performed by the same update rule as in 3.11. For the proof of convergence of the update rule we refer the reader to [161].

3.2.3   Experiments

The performance of the approach is presented in this section. The approach was tested against three well known data sets in the field of retrieval: MPEG7(shape), YALE(faces) and ORL(faces). For all test data sets the number of iterations for the update rule is set to 200. A given pairwise distance $\mathcal{D}$ is transformed to similarity (edge-weight) using a standard Gaussian kernel

$$\mathtt{A}_{ij}^{\sigma} = \mathbb{K}_{i \neq j} exp(-\mathcal{D}/2\sigma^2)$$

where σ is the free scale parameter, and $\mathbb{K}_P = 1$ if P is true, 0 otherwise. $\mathcal{L}$ is then built, from $\mathtt{A}$, using the constrained dominant set framework. The diffusion process is then computed using the update rule 3.11 which resulted in the final learned affinity for ranking.

A similar experimental analysis as of [40] has been conducted. In [40], a generic framework with 72 different variant of diffusion processes was defined which are resulted from three steps: initialization, definition of transition matrix and diffusion process. In our experiment, the update scheme is fixed to 3.11 which has proven to be effective. The four different types

of initialization schemes are Affinity Matrix **A** (A1) [142], Identity Matrix **I** (A2), Transition Matrix **P** which is the standard random walk transition matrix (A3) [100] and Transition Matrix $\mathbf{P}_{kNN}$ which is the random walk transition matrix constrained to the k-nearest neighbors (A4) [100]. Including our transition matrix (B6), we have in total 6 different types of transition matrices: **P** (B1), Personalized PageRank Transition Matrix $\mathbf{P}_{PPR}$ (B2) [100], $\mathbf{P}_{kNN}$ (B3), Dominant Set Neighbors $\mathbf{P}_{DS}$ [162] (B4), and Affinity Matrix **A** (B5)

$\mathcal{R}$	20	25	30	35	40
B3	94.321	97.871	**98.614**	99.357	**100**
B4	94.296	97.846	**98.614**	99.357	**100**
Ours	**94.354**	**97.896**	**98.614**	**99.360**	**100**

Table 3.4: *Results on MPEG7 dataset varying the first $\mathcal{R}$ returned objects*

Metric: The Bull's eye score is used as a measure of retrieval accuracy. It measures the percentage of objects sharing the same class with a query q in the top $\mathcal{R}$ retrieved shapes. Let us say $\mathcal{C}$ is the set of objects in the same class of the query q and $\mathcal{O}$ is the set of top $\mathcal{R}$ retrieved shapes. The Bull's eye score ($\mathcal{B}$) is then computed as $\mathcal{B} = \frac{|\mathcal{O} \cap \mathcal{C}|}{|\mathcal{C}|}$

MPEG7: a well known data set for testing performance of retrieval and shape matching algorithms. It comprises 1400 silhouette shape images of 70 different categories with 20 images in each categories. In all reported results, Articulated Invariant Representation (AIR) [53], best performing shape matching algorithm, is used as the input pairwise distance measure. The retrieval performance is measured fixing $\mathcal{R}$ to 40.

The retrieval performance has also been tested by varying the first $\mathcal{R}$ returned objects, the set in which instances of the same category are checked in. For the purpose of this experiment we use the best diffusion variants (B3 and B4 initialized with A2). The performance of the algorithms is shown in Table 3.4.

MPEG7 has been used, most frequently, for testing retrieval algorithms. Table 3.5 shows the comparison against different state-of-the-art approaches.

YALE: [22] a popular benchmark for face clustering which consists of 15 unique people with 11 pictures for each under different conditions: normal, sad, sleepy, center light, right light, ... etc that include variations of pose, illumination and expression. Similar procedure of [71] and [40] were followed to build the distance matrix. Down sample the image, normalize to 0-mean and 1-variance, and compute the Euclidean distance between the

[88]	[20]	[53]	[89]	[162]	[40]	Ours
85.40	91.61	93.67	95.96	99.99	**100**	**100**

Table 3.5: *Retrieval performance comparison on MPEG7 dataset. **Up:** methods, **Down:** Bull's eye score for the first 40 elements*

vectorized representation. The retrieval performance is measured fixing $\mathcal{R}$ to 15.

YALE	B1	B2	B3	B4	B5	B6(Ours)
A1	71.74	71.24	**75.59**	75.31	70.25	75.15
A2	71.96	70.69	77.30	76.20	69.92	**77.41**
A3	72.07	70.57	74.93	76.14	70.30	**75.37**
A4	72.23	70.74	77.08	76.10	70.25	**77.36**

Table 3.6: *Results on YALE dataset. Bull's eye score for the first 15 elements*

Results of the algorithm on YALE data set varying $\mathcal{R}$ is shown in Table 3.7.

$\mathcal{R}$	20	25	30	35	40
B3	71.240	**74.105**	77.303	79.559	80.826
B4	70.854	72.176	76.198	77.741	79.063
Ours	**71.350**	74.050	**77.411**	**80.000**	**81.653**

Table 3.7: *Results on YALE dataset varying the first $\mathcal{R}$ returned objects*

ORL: face data set of 40 different persons with 10 grayscale images per person with slight variations of pose, illumination, and expression. Similar procedure as of YALE data set was followed and The retrieval performance is measured fixing $\mathcal{R}$ to 15.

Results of the algorithm on ORL data set varying $\mathcal{R}$ is shown in Table 3.9.

3.2.4 Summary

In this work, we have developed a locally constrained diffusion process which, as of existing methods, has no problems such as choosing optimal local neighbor size and initializing the dynamics to extract dense neighbor

ORL	B1	B2	B3	B4	B5	B6(Ours)
A1	72.75	73.48	**74.25**	73.90	70.58	**74.25**
A2	72.75	73.75	**77.42**	74.82	70.15	**77.42**
A3	73.12	73.75	**75.52**	75.35	71.05	**75.52**
A4	73.12	73.75	**77.32**	75.50	71.40	**77.32**

Table 3.8: *Results on ORL dataset. Bull's eye score for the first 15 elements*

$\mathcal{R}$	20	25	30	35	40
B3	**70.950**	**75.250**	**77.425**	**79.275**	**80.550**
B4	68.850	72.900	74.825	76.775	77.700
Ours	**70.950**	**75.250**	**77.425**	**79.275**	**80.550**

Table 3.9: *Results on ORL dataset varying the first $\mathcal{R}$ returned objects*

which constrain the diffusion process. The framework alleviates the issues with an up-tick in the results.

Deep Constrained Dominant Sets for Person Re-identification

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 reformulate 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
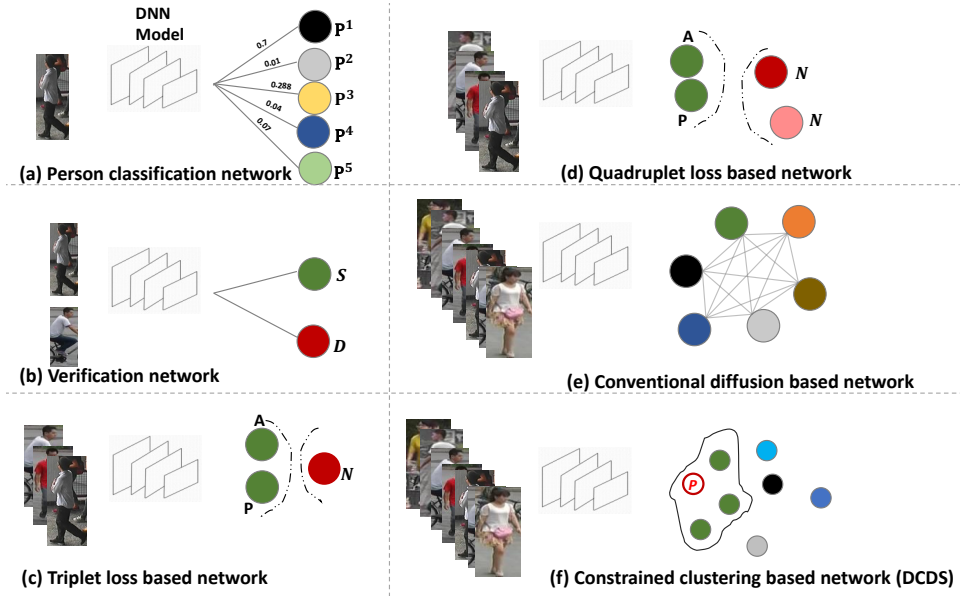
Figure 4.1: *Shows a variety of existing classification and similarity-based deep person re-id models. (a) Depicts a classification-based deep person re-id model, where P^i refers to the ith person. (b) Illustrates a verification network whereby the similarity S and dissimilarity D for a pair of images 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.*

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.

4.1 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, view and illumination variation can be mentioned.

Person re-id can be taken as a person retrieval problem based on the ranked similarity score, which is obtained from the pairwise affinities be-

tween 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 that considers the intrinsic manifold structure of the given affinity matrix [17], [92], [41], [19]. 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. Aiming to improve the discriminative power of a DNN model, there have been recent works which incorporate a similarity diffusion process in an end-to-end manner [124], [125], [28]. Following [23], which applies a random walk in an end-to-end fashion for solving semantic segmentation problem, authors in [124] proposed a group-shuffling random walk network for fully utilizing the affinity information between gallery images in both the training and testing phase. Also, the authors of [125] proposed 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 chapter, 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:

- For the very first time, the dominant sets clustering method is integrated in a DNN and optimized in end-to-end fashion.

- A one-to-one correspondence between person re-identification and constrained clustering problem is established.

- State-of-the-art results are significantly improved.

The chapter 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.

4.2 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 [30], [176], [141], [3], [46]. For instance, verification network (V-Net) [85], Figure 4.1(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 [29]. 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 4.1(c), triplet loss is exploited to enforce the correct order of relative distances among image triplets [30], [38], [176] . In Figure 4.1(d), Quadruplet loss [29] 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, [28], [45], [124], [98], [125] proposed graph based end-to-end diffusion methods shown in Figure 4.1(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.* [125] 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.* [157] 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.* [23] 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, [124] integrates random walk method in end-to-end learning to tackle person re-identification problem. In [124], 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 [125], [75], [124], [28], Shen *et al.* [125] learn the edge weights by exploiting the training label supervision, thus they are able to learn more accurate feature fusion weights for updating nodes feature.

Recent applications of dominant sets. Dominant sets (DS) clustering [103] and its constraint variant constrained dominant sets (CDS) [172] have been employed in several recent computer vision applications ranging from person tracking [138], [139], geo-localization [173], image retrieval [170], [8], 3D object recognition [143], to Image segmentation and co-segmentation [99], [13]. Zemene *et al.* [172] presented CDS with its applications to interactive Image segmentation. Following, [99] uses CDS to tackle both image segmentation and co-segmentation in interactive and unsupervised setup. Wang *et al.* [143] 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 model, 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 4.1(f), as can be seen, the objective here is to find a coherent constrained cluster which incorporates the given probe image P.
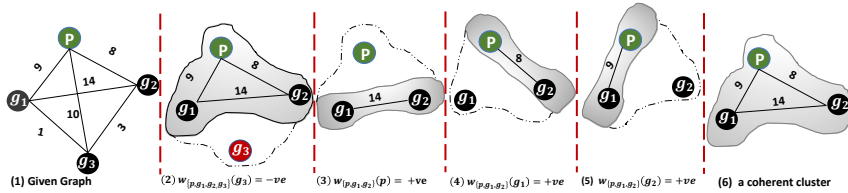
Figure 4.2: *Let $S = \{P, g_1, g_2, g_3\}$ comprises probe, P, and gallery images g_i. As can be observed from the above toy example, the proposed method assess the contribution of each participant node $i \in S$ with respect to the subset $S \backslash i$. (1) shows graph G, showing the pairwise similarities of query-gallery images. (2-5) show the relative weight, $W_{\Gamma}(i)$ (Equ. 4.1), 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).*

4.3 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).

4.3.1 Modeling person re-id as a Dominant Set

Recent methods [28], [23] have proposed different models, which leverage local and group similarity of images in an end-to-end manner. Authors in [28] 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,
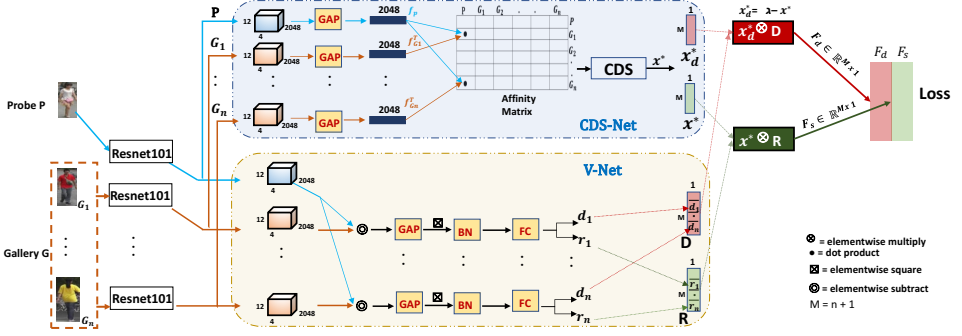
Figure 4.3: *Workflow of the proposed DCDS. Given n number of gallery images, G, and probe image P, we first extract their Resent101 features right before the global average pooling (GAP) layer, which are then fed to CDS-Net (upper stream) and V-Net (lower stream) branches. In the CDS-branch, after applying 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. 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 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, $D \in \mathbb{R}^{M \times 1}$. Afterward, 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 $G_t \in \mathbb{R}^{M \times 1}$.*

and their similarity as edge-weight w_{ij}. Given vertices V, and $S \subseteq V$ 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:

$$w_S(i) = \begin{cases} 1, & \text{if} \quad |S| = 1, \\ \sum_{j \in S \setminus \{i\}} \phi_{S \setminus \{i\}}(j, i) w_{S \setminus \{i\}}(j), & \text{otherwise} \end{cases} \quad (4.1)$$

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 $S \setminus \{i\}$. 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** 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 characters of DS are inherited.

4.3.1.1   A Set of a person images as a constrained cluster

We cast person re-identification as finding a constrained cluster, where, elements of the cluster correspond to a set of same person images and the constraint refers to the probe image used to extract the corresponding cluster. 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, the characteristics vector, which contains the membership scores of all participants is always a stochastic vector, meaning that $\sum_{i=1}^M z_i = 1$, where z_i denotes the membership score of each image in the cluster.

As can be seen from the toy example in Figure 4.2, the initial pairwise similarities between the query and gallery images hold valuable information, which define the relation 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 with defining a weight of each image p, g_1, g_2, g_3 with regard to subset $S \backslash i$ as depicted in Figure 4.2, $(2-5)$, respectively. As can be observed from Figure 4.2, 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 $S = \{p, g_1, g_2\}$ has a positive impact on the coherency of the cluster. It is evident that the illustration in Figure 4.2

verifies that the proposed DCDS (Deep Constrained Dominant Set) 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.

4.3.2 CDS Based End-to-end Learning

In this section, we discuss the integration of CDS in end-to-end learning. We adopt a siamese based Resent101, with a novel verification loss to find probe-gallery similarity, R, and dissimilarity, D scores. As can be seen from Figure 4.3, we have two main branches: CDS network branch (CDS-Net) and verification network branch (V-Net). In the CDS-Net, the elements of pairwise affinity matrix are computed first as a dot product of the global pooling feature of a pair of images. Afterward, the replicator dynamics [148] is applied, which is a discrete time solver of the parametrized quadratic program, Equ. 4.2, whose solution corresponds to the CDS. Thus, assuming that there are M images in the given mini-batch, the replicator dynamics, Equ. 1.10, is recursively applied M times taking each image in the mini-batch as a constraint. Given 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 parameter $-\alpha$ to the diagonal corresponding to the subset $V \backslash P$ and zero to the diagonal corresponding to the constraint image P. Next, the modified adjacency matrix, B, is feed 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:

$$\text{maximize} \quad f_P^\alpha(\mathbf{x})^i = \mathbf{x}'B\mathbf{x} \quad where, B = A - \alpha \hat{I}_p.$$
$$\text{subject to} \quad \mathbf{x} \in \Delta \tag{4.2}$$

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 4.3) 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

$$
\begin{aligned}
F_s &= \beta(Y) \otimes (1 - \beta)(S'), \\
F_d &= \beta(Y_d) \otimes (1 - \beta)(D'), \quad where, \quad Y_d = \delta - Y
\end{aligned}
\tag{4.3}
$$

δ is empirically set to 0.3. 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 4.3 illustrates the case of a single probe-gallery, whereas Equ. 4.3 shows the solution of M probe images in a given mini-batch.

4.3.3 Auxiliary Net

In this work, we integrate an auxiliary net to further improve the performance of our model. The auxiliary net is trained based on the multi-scale prediction of Resnet50 [57]. It is a simple yet effective architecture, whereby we can easily compute both triplet and cross entropy loss of different layers of Resnet50 [57], 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 4.4, we employ three features at different layers from Resnet50 $conv5\_x$ Layer, 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 each feature which comes from the Relu and FC layers, respectively. During testing phase we concatenate the features that come from the DCDS and Auxiliary Net to find 4096 dimensional feature. We then apply CDS to find the final ranking_score, (See Figure 4.5).
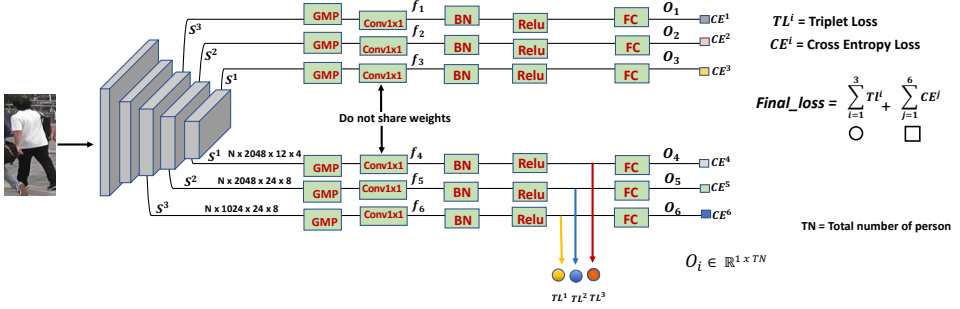
Figure 4.4: *Illustrates the auxiliary net, which consists of two branches which are jointly trained. We first use features at different layers, S_1, S_2, S_3, and then feed these to Global Maxpooling (GMP), Conv, BN, Relu and FC layers for further encoding. We then compute triplet losses employing the 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 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 finally compute the final loss as the sum of the average of the triplet and cross entropy losses.*

4.3.4 Constraint Expansion During Testing

We propose a new scheme (illustrated in Figure 4.6) 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 4.5), we first collect k-NNs 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.

4.4 Experiments

To validate the performance of our method we have conducted several experiments on three publicly available benchmark datasets, namely CUHK03 [85], Market1501 [177], and DukeMTMC-reID [182].

4.4.1 Datasets and evaluation metrics

Datasets: CUHK03 [85] dataset comprises 14,097 manually and automatically cropped images of 1,467 identities, which are captured by two cam-
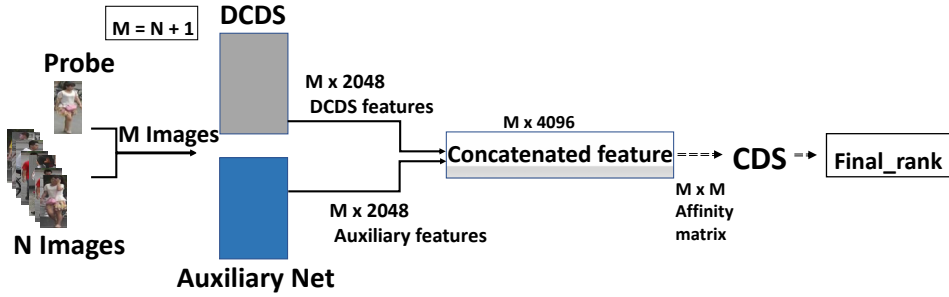
Figure 4.5: *During testing, given a probe and gallery images, we extract DCDS and auxiliary features and concatenate them to find a single vector. Afterward, we build M x M affinity matrix and run CDS with constraint expansion mechanism to find the final probe-gallery similarity rank.*

eras on campus; in our experiments, we have used manually annotated images. Market1501 dataset [177] 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) [49] is utilized. DukeMTMC-reID is generated from a tracking dataset called DukeMTMC. DukeMTMC 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.

In multiple dataset (MD) setup, we first train our model on eight datasets: CUHK03 [85], CUHK01 [84], Market1501 [177], DukeMTMC-reID [182], Viper [34], MSMT17 [147], GRID [93], and ILIDS [181]. Next, we finetune and evaluate on each of CUHK03 [85], Market1501 [177], and DukeMTMC-reID [182] datasets.

Evaluation Metrics: Following the recent person re-id methods, we use mean average precision (mAP) as suggested in [177], 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 [177].

Methods	mAP	rank-1	rank-5
SGGNN [125] ECCV18	82.8	92.3	96.1
DKPM [126] CVPR18	75.3	90.1	96.7
DGSRW [124] CVPR18	82.5	92.7	96.9
GCSL [28] CVPR18	81.6	93.5	-
CPC [146] CVPR18	69.48	83.7	-
MLFN [26] CVPR18	74.3	90.0	-
HA-CNN [86] CVPR18	75.7	91.2	-
PA [130] ECCV18	74.5	88.8	95.6
HSP [73] CVPR18	83.3	93.6	97.5
Ours	**85.8**	**94.81**	**98.1**
$RA_{w/RR}$ [145] CVPR18	86.7	90.9	-
$PA_{w/RR}$ [130] ECCV18	89.9	93.4	96.4
$HSP_{w/RR}$ [73] CVPR18	90.9	94.6	96.8
Ours$_{w/RR}$	**93.3**	**95.4**	**98.3**

Table 4.1: *A comparison of the proposed method with state-of-the-art methods on Market1501 dataset. Upper block, without re-ranking methods. Lower block, with re-ranking method, w/RR, [183].*

4.4.2 Implementation Details

We implement DCDS based on Resnet101 [57] architecture, which is pretrained on imagenet dataset. We adopt the training strategy of Kalayeh *et al.* [73], and aggregate eight different person re-id benchmark dataset to train our model. In total, the merged dataset contains 89,091 images, which comprises 4937 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 size is set to 64, each mini-batch has 16 person-ID 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 [184]. For optimization we use Adam, we initially set the learning rate to 0.0001, and drop it by 0.1 in every 40 epochs. The fusing parameter in Equ. 4.3, β, is set to 0.9.

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 4.2: *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.*

Methods	rank-1	rank-5
SGGNN [125] ECCV18	95.3	99.1
DKPM [126] CVPR18	91.1	98.3
DGSRW [124] CVPR18	94.9	98.7
GCSL [28] CVPR18	90.2	98.5
MLFN [26] CVPR18	89.2	-
CPC [146] CVPR18	88.1	-
PA [130] ECCV18	88.0	97.6
HSP [73] CVPR18	94.28	99.04
Ours	**95.8**	**99.1**

Table 4.3: *A comparison of the proposed method with state-of-the-art methods on CUHK03 dataset.*

4.4.3   Results on Market1501 Datasets

As can be seen from Table 4.1, on Market dataset, our proposed method improves state-of-the-art method [73] by $2.5\%, 1.21\%$, and 0.6% in mAP, rank-1 and rank-5 scores, respectively. Moreover, comparing to state-of-the-art graph-based DNN method, SGGNN [125], 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 [183], 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 4.1, our method has gain mAP of 2.2% over HSP [73], and 10.5 % over SGGNN [125], 10.8 % over DGSRW.

Methods	mAP	rank-1	rank-5
SGGNN [125] ECCV18	68.2	81.1	88.4
DKPM [126] CVPR18	63.2	80.3	89.5
DGSRW [124] CVPR18	66.4	80.7	88.5
GCSL [28] CVPR18	69.5	84.9	-
CPC [146] CVPR18	59.49	76.44	-
MLFN [26] CVPR18	62.8	81.0	-
RAPR [145] CVPR18	80.0	84.4	-
PA [130] ECCV18	64.2	82.1	90.2
HSP [73] CVPR18	73.3	85.9	92.9
Ours	**75.5**	**87.5**	-
$PA_{w/RR}$ [130] ECCV18	83.9	88.3	93.1
$HSP_{w/RR}$ [73] CVPR18	84.99	88.9	94.27
Ours $_{w/RR}$	**86.1**	88.5	-

Table 4.4: *A comparison of the proposed method with state-of-the-art methods on DukeMTMC-reID dataset.Upper block, without re-ranking methods. Lower block, with re-ranking method,w/RR, [183].*

	Train on Duke, CUHK03 → Test on Market1501	
Methods	mAP	rank-1
PUL [48]	20.5	45.5
Ours	**24.5**	**51.3**

Table 4.5: *A comparison of the proposed method with PUL [48] on Market1501 dataset.*

4.4.4 Results on CUHK03 Datasets

Table 4.5 shows the performance of our method on CUHK03 dataset. Since most of the Graph-based DNN models report their result on the standard protocol [85], we have experimented on the standard evaluation protocol, to make fair comparison. As can be observed from Table 4.5, our method gain a marginal improvement in the mAP. Using a reranking method [183], we have reported a competitive result in all evaluation metrics.

4.4.5 Results on DukeMTMC-reID Dataset

Likewise, in DukeMTMC-reID dataset, the improvements of our proposed method is noticeable. Our method has surpassed state-of-the-art method [73] by $1.7\%/1.6\%$ in mAP/rank-1 scores. Moreover, comparing to state-of-the-art graph-based DNN, our method outperforms DGSRW [124], SG-GNN [125] and GCSL [28] by $9.1\%, 7.3\%$, and 6% in mAP, respectively.

4.4.6 Ablation Study

To investigate the impact of each component in our architecture, we have performed an ablation study. Thus, we have reported the contributions of each module in Table 4.2. To make a fair comparison with the baseline and graph-based DNN models, the ablations study is conducted in a single-dataset (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 [126], [124], [125]. On Market1501 dataset, 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 [124], [125]. As a result, on DukeMTMC-reID dataset our method surpass [124] by $9.1\%/6.8\%$, and [125] by 17.9 % / 7.4 % 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 4.7. Figure 4.7(a) shows the effect of fusing parameter, β, Equ. (4.3) 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 4.7(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.

In the supplementary material, we provide additional experiments on cross-dataset person-re-identification (re-id) using the proposed deep constrained dominant sets (DCDS) on Market1501 dataset. In section one, we summarize the datasets we used in our experiments. In section two, we

present the experiments we have performed on cross-dataset person re-id. And, in section three, we provide hyper parameter analysis on DukeMTMC-reID and CUHK03 datasets. Figure 4.8 illustrates an example of our method training-output (left) and learning objective, target matrix, (right). Figure 4.9 demonstrates the similarity fusing process, between the V-Net and CDS-Net, alongside sample qualitative results.

Experiments on cross-datasets evaluation. Due to the lack of abundant labeled data, cross-dataset person re-id has attracted great interest. Recently, Fan *et al.* [48] have developed a progressive clustering-based method to attack cross-dataset person re-id problem. To further validate our proposed DCDS, we apply our method on cross-dataset person re-id problem and compare it with progressive unsupervised learning (PUL) [48]. To this end, we train our model on DukeMTMC-reID and CUHK03 datasets and test it on Market1501 dataset. We then compare it with PUL [48], which has also been trained on CUHK03 and DukeMTMC-reID datasets. As can be observed from Table 4.5, even though our proposed method is not intended for cross-dataset re-id, it has gained a substantial improvements over PUL [48], that was mainly designed to attack person re-id problem in a cross-dataset setup.

4.4.7 Parameter Analysis

Similar to the parameter analysis reported in the main manuscript, we report hyper parameter analysis on DukeMTMC-reID and CUHK03 dataset. The performance of our method with respect to the fusing parameters on DukeMTMC-reID and CUHK03 are shown in Figure 4.10 (a) and Figure 4.10 (b), respectively. Thereby, as can be observed, the results show similar phenomena as in Market1501, where the mAP increases with a larger β value. Figure 4.11 shows the similarity distribution given by the baseline and the proposed DCDS using three different probe-images, with a batch size of 64, and setting Ω to 4, 8 and 16.

4.5 Summary

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.

CDS with single constraint on the top k-NN

I^i P^j

CDS with multiple constraints over the entire gallery images
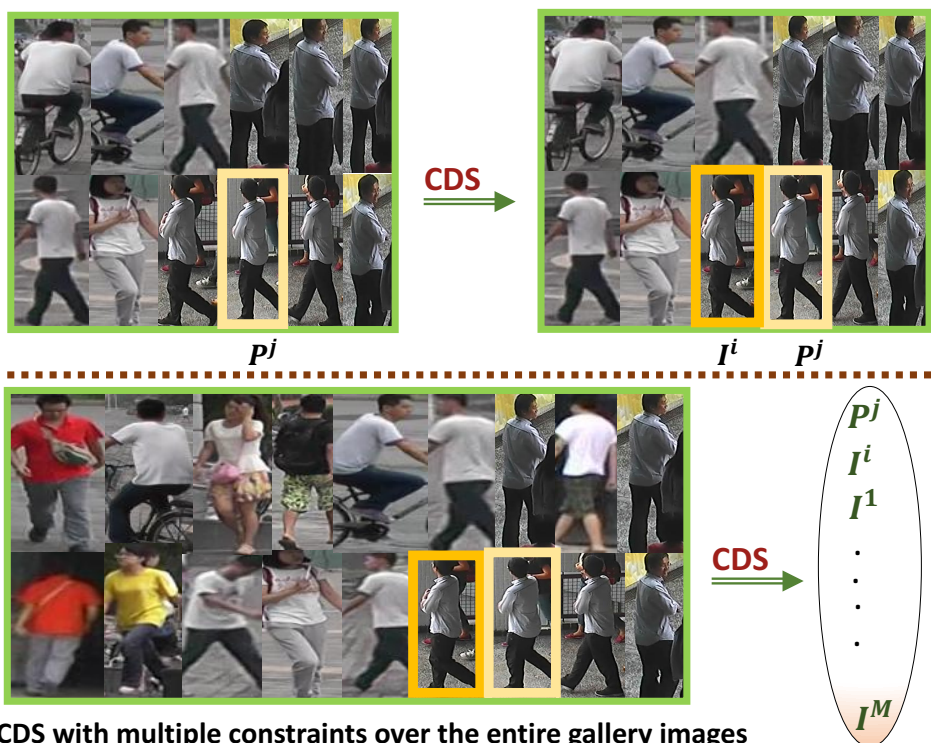
Figure 4.6: *Given a constraint (probe-image) P^j, we first collect k-NNs to the probe-image, based on the pairwise similarities. Subsequently, we run CDS on the graph of the k-NN. Then, based on the cluster membership score obtained, we choose image I^i, with the highest membership score and re-run CDS, considering P^j and I^i as constraints, over the graph of the all set of images, I^M, in the minibatch. Afterward, we consider the solution as our final rank.*
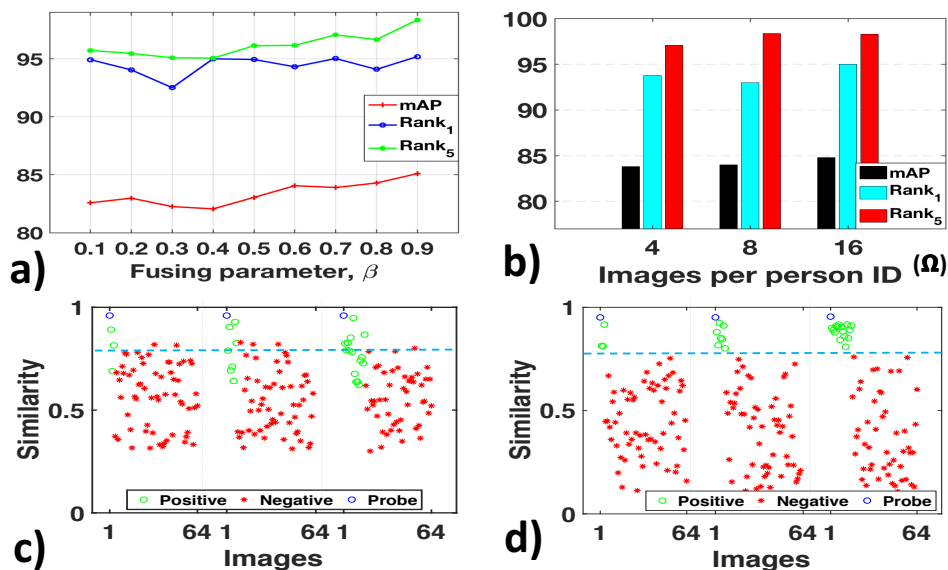
Figure 4.7: *Illustrates different experimental analysis performed on Market1501 dataset. a) shows the impact of fusing parameter β in Equ. 4.3. b) shows the performance of our model with varying the number of images per person in a given batch. c) and d) illustrate the similarity between the probe and 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.*
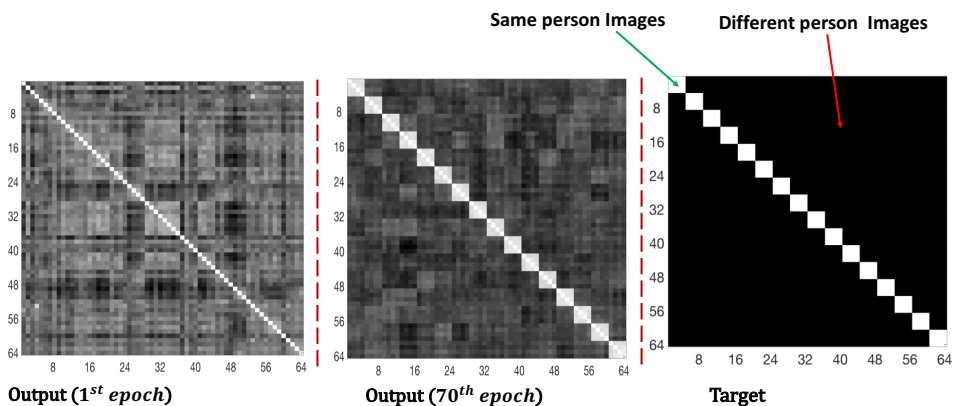
Figure 4.8: *On the right hand side, the target matrix is shown. There are total 16 persons in the mini-batch and 4 images per ID ($\Omega = 4$), batch size = 64. In the target matrix, the white-blocks represent the similarity between the same person-images in the mini-batch, whereas the black-blocks of the matrix define the dissimilarities between different person images. In the similarity matrix shown left (after one epoch) and middle (after 70th epochs) each row of the output matrix denotes the fused similarity obtained from the CDS-Net and V-Net, per Equ. (6) in the main manuscript. Thus, we optimize our model until we obtain an output with a similar distribution of the target matrix. As can be seen, our model has effectively learned and gives a similarity matrix (shown in the middle) which is closer to the target matrix.*
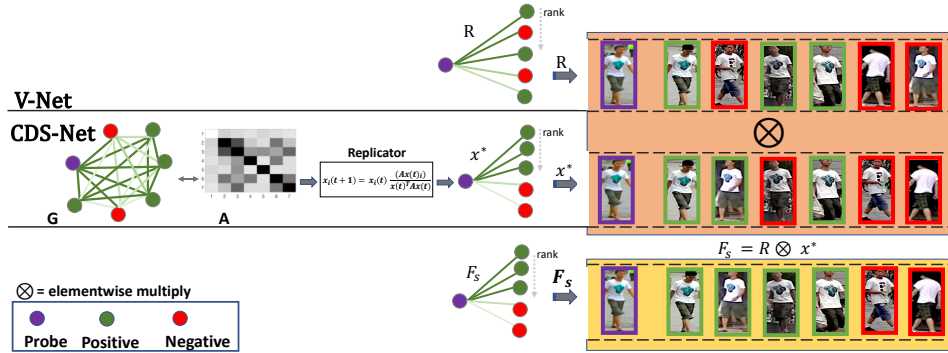
Figure 4.9: *Exemplar results obtained as a result of the similarity fusion between the V-Net and CDS-Net. The Upper-row shows the probe and gallery similarity (R) obtained from the V-Net, where the green circles show persons similar to the probe (shown by purple-circle), while the red circles denote persons different from the probe image. Middle-row shows the workflow in CDS-Net. First, graph G is formed using the similarity obtained from the dot products. We then construct the modified affinity matrix B, followed by application of replicator dynamics on B to obtain the probe gallery similarity (X^*). Finally, We elementwise multiply X^* and R to find the final probe-gallery similarity (F_s), shown in the third row. The intensity of the edges in, G, R, x^*, and F_s define the similarity value, where the bold ones denote larger similarity values, whereas the pale-edges depict smaller similarity values.*
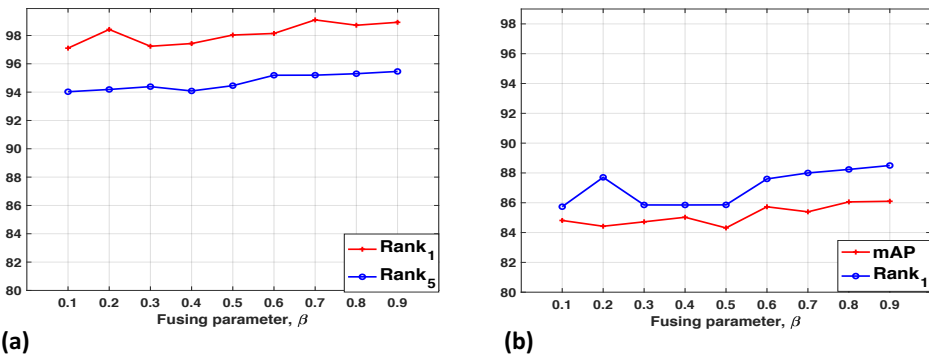
Figure 4.10: *Performance of our model with respect to fusing parameter β, on (a) CUHK03, and (b) DukeMTMC-reID, datasets.*
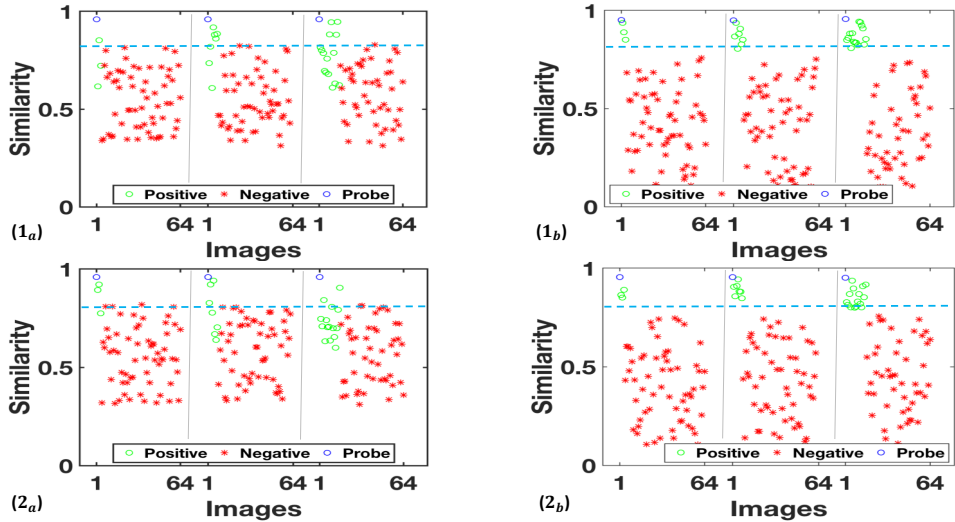
Figure 4.11: *Shows experimental analysis performed on CUHK03 $(1_{a,b})$, and DukeMTMC-reID $(2_{a,b})$ datasets. $1_a, 2_a$ and $1_b, 2_b$ illustrate the similarity between the probe-gallery images obtained from the baseline and DCDS method, respectively. It can be observed that the baseline method has assigned 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. Note that, for better visibility, we have randomly assigned a large (close to 1) self-similarity value to the probe (blue-circle).*

Conclusion

In this thesis, we have proposed several schemes which exploit constrained clustering mechanism to tackle different computer vision problems such as, Image Segmentation and Co-segmentation, Image Retrieval, and Person Re-identification. Thereby, we validate the indispensability of the proposed graph-based algorithms. Moreover, the usage of constrained dominant sets (CDS) in an end-to-end manner demonstrates the advantage of integrating graph-based classical methods into a deep neural network (DNN) model.

In Chapter 2, we have demonstrated the applicability of CDS to problems such as interactive image segmentation and co-segmentation (in both the unsupervised and the interactive flavor). In our perspective, these can be thought of as "constrained" segmentation problems involving an external source of information (being it, for example, a user annotation or a collection of related images to segment jointly) which somehow drives the whole segmentation process. The approach is based on some properties of a family of quadratic optimization problems related to dominant sets which show that, by properly selecting a regularization parameter that controls the structure of the underlying function, we are able to "force" all solutions to contain the constraint elements. The proposed method is flexible and is capable of dealing with various forms of constraints and input modalities,

such as scribbles and bounding boxes, in the case of interactive segmentation. Extensive experiments on benchmark datasets have shown that our approach considerably improves the state-of-the-art results on the problems addressed.

In Chapter 3, we addressed the content-based image retrieval problem. We developed a novel and computationally efficient CBIR method based on a constrained-clustering concept. In particular, we showed an efficient way of estimating a positive impact weight of features in a query-specific manner. Thus it can be readily used for feature combination. Furthermore, the proposed scheme is fully unsupervised, and can easily be able to detect false-positive NNs to the query, through the diffused similarity of the NNs. To demonstrate the validity of our method, we performed extensive experiments on benchmark datasets. Besides the improvements achieved on the state-of-the-art results, our method shows its effectiveness in quantifying the discriminative power of given features. Moreover, its effectiveness on feature-weighting can also be exploited in other computer vision problems, such as person re-identification, object detection, and image segmentation.

On the other hand, in CDSIR, we have developed a locally constrained diffusion process which, as of existing methods, has no problems such as choosing optimal local neighbor size and initializing the dynamics to extract dense neighbor which constrain the diffusion process. The framework alleviates the issues while improving the performance. Experimental results on three well known datasets in the field of retrieval demonstrate that the approach compares favorably with state-of-the-art algorithms.

In Chapter 4, 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 similarity based applications.

Bibliography

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012.

[2] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(6):641–647, 1994.

[3] 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.

[4] A.Joulin, F.R.Bach, and J.Ponce. Discriminative clustering for image co-segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 1943–1950, 2010.

[5] F.R.Bach A.Joulin and J.Ponce. Multi-class cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 542–549, 2012.

[6] Selim Aksoy and Robert M. Haralick. Graph-theoretic clustering for image grouping and retrieval. In *1999 Conference on Computer Vision and Pattern Recognition (CVPR'99), 23-25 June 1999, Ft. Collins, CO, USA*, page 1063, 1999.

[7] Andrea Albarelli, Samuel Rota Bulò, Andrea Torsello, and Marcello Pelillo. Matching as a non-cooperative game. In *Proc. IEEE 12th International Conference on Computer Vision (ICCV)*, pages 1319–1326, 2009.

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

[9] Leulseged Tesfaye Alemu, Mubarak Shah, and Marcello Pelillo. Deep constrained dominant sets for person re-identification. *CoRR*, abs/1904.11397, 2019.

[10] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: CNN architecture for weakly supervised place recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA, June 27-30, 2016*, pages 5297–5307, 2016.

[11] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2911–2918, 2012.

[12] Pablo Arbelaez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33:898–916, 2011.

[13] Sinem Aslan and Marcello Pelillo. Weakly supervised semantic segmentation using constrained dominant sets. In *International Conference on Image Analysis and Processing*, pages 425–436. Springer, 2019.

[14] Yannis S. Avrithis and Yannis Kalantidis. Approximate gaussian mixtures for large scale vocabularies. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III*, pages 15–28, 2012.

[15] Artem Babenko, Anton Slesarev, Alexander Chigorin, and Victor S. Lempitsky. Neural codes for image retrieval. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 584–599, 2014.

[16] Junjie Bai and Xiaodong Wu. Error-tolerant scribbles based interactive image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 392–399, 2014.

[17] 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.

[18] 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.

[19] 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.

[20] Xiang Bai, Xingwei Yang, Longin Jan Latecki, Wenyu Liu, and Zhuowen Tu. Learning context-sensitive shape similarity by graph transduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5):861–874, 2010.

[21] Xue Bai and Guillermo Sapiro. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *Int. J. Computer Vision*, 82(2):113–132, 2009.

[22] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, 1997.

[23] 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.

[24] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 105–112, 2001.

[25] Bryan C. Catanzaro, Bor-Yiing Su, Narayanan Sundaram, Yunsup Lee, Mark Murphy, and Kurt Keutzer. Efficient, high-quality image contour detection. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 2381–2388, 2009.

[26] 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.

[27] Morteza Haghir Chehreghani. Adaptive trajectory analysis of replicator dynamics for data clustering. *Machine Learning*, 104(2-3):271–289, 2016.

[28] 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.

[29] 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.

[30] 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.

[31] Selene E. Chew and Nathan D. Cahill. Semi-supervised normalized cuts for image segmentation. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 1716–1723, 2015.

[32] Ondrej Chum, Andrej Mikulík, Michal Perdoch, and Jiri Matas. Total recall II: query expansion revisited. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Colorado Springs, CO, USA, 20-25 June 2011*, pages 889–896, 2011.

[33] C.Rother, T.P. Minka, A.Blake, and V.Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *IEEE Conference on Computer Vision and Pattern Recognition,(CVPR*, pages 993–1000, 2006.

[34] S. Brennan D. Gray and H. Tao. Evaluating appearance models for recognition, reacquisition. In *IEEE International workshop on performance evaluation of track- ing and surveillance,*, pages 31–44, 2007.

[35] D.Batra, A.Kowdle, D.Parikh, J.Luo, and T.Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 3169–3176, 2010.

[36] Cheng Deng, Rongrong Ji, Wei Liu, Dacheng Tao, and Xinbo Gao. Visual reranking through weakly supervised multi-graph learning. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 2600–2607, 2013.

[37] Thomas Deselaers, Tobias Weyand, and Hermann Ney. Image retrieval and annotation using maximum entropy. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, pages 725–734, 2006.

[38] 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.

[39] Thanh-Toan Do, Quang D. Tran, and Ngai-Man Cheung. Faemb: A function approximation-based embedding method for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR , Boston, MA, USA, June 7-12, 2015*, pages 3556–3564, 2015.

[40] Michael Donoser and Horst Bischof. Diffusion processes for retrieval revisited. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 1320–1327, 2013.

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

[42] Olivier Duchenne, Jean-Yves Audibert, Renaud Keriven, Jean Ponce, and Florent Ségonne. Segmentation by transduction. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, 2008.

[43] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification, 2nd Edition*. Wiley, 2001.

[44] Amir Egozi, Yosi Keller, and Hugo Guterman. Improving shape retrieval by spectral matching and meta similarity. *IEEE Trans. Image Processing*, 19(5):1319–1327, 2010.

[45] Ismail Elezi, Alessandro Torcinovich, Sebastiano Vascon, and Marcello Pelillo. Transductive label augmentation for improved deep network learning. In *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018*, pages 1432–1437, 2018.

[46] Ismail Elezi, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixé. The group loss for deep metric embedding. *CoRR*, 2019.

[47] Anders P. Eriksson, Carl Olsson, and Fredrik Kahl. Normalized cuts revisited: A reformulation for segmentation with linear grouping constraints. *Journal of Mathematical Imaging and Vision*, 39(1):45–61, 2011.

[48] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *TOMCCAP*, 14(4):83:1–83:18, 2018.

[49] 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.

[50] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Pearson, 2011.

[51] Gerald Friedland, Kristian Jantz, and Raúl Rojas. SIOX: Simple interactive object extraction in still images. In *(ISM*, pages 253–260, 2005.

[52] Yoram Gdalyahu, Daphna Weinshall, and Michael Werman. Self-organization in vision: Stochastic clustering for image segmentation, perceptual grouping, and image database organization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(10):1053–1074, 2001.

[53] Raghuraman Gopalan, Pavan K. Turaga, and Rama Chellappa. Articulation-invariant representation of non-planar shapes. pages 286–299, 2010.

[54] Albert Gordo, Jon Almazán, Jérôme Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 241–257, 2016.

[55] Leo Grady. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(11):1768–1783, 2006.

[56] Avik Hati, Subhasis Chaudhuri, and Rajbabu Velmurugan. Image co-segmentation using maximum common subgraph matching and region co-growing. In *Computer Vision - ECCV 2016 - 14th European Conference*, pages 736–752, 2016.

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

[58] H.Fu, X.Cao, and Z.Tu. Cluster-based co-saliency detection. *IEEE Trans. Image Processing*, 22(10):3766–3778, 2013.

[59] H.Li and K.N.Ngan. A co-saliency model of image pairs. *IEEE Trans. Image Processing*, 20(12):3365–3375, 2011.

[60] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[61] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Geometric context from a single image. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 654–661, 2005.

[62] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

[63] Michael E. Houle, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *SSDBM*, pages 482–500, 2010.

[64] Ahmet Iscen, Giorgos Tolias, Yannis S. Avrithis, Teddy Furon, and Ondrej Chum. Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, July 21-26, 2017*, pages 926–935, 2017.

[65] Mihir Jain, Rachid Benmokhtar, Hervé Jégou, and Patrick Gros. embedding similarity-based image classification. In *International Conference on Multimedia Retrieval, ICMR '12, Hong Kong, China, June 5-8, 2012*, page 19, 2012.

[66] Suyog Dutt Jain and Kristen Grauman. Predicting sufficient annotation strength for interactive foreground segmentation. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 1313–1320, 2013.

[67] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, pages 304–317, 2008.

[68] Herve Jegou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, 20-25 June 2009, Miami, Florida, USA*, pages 1169–1176, 2009.

[69] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1704–1716, 2012.

[70] Hervé Jégou and Andrew Zisserman. Triangulation embedding and democratic aggregation for image search. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Columbus, OH, USA, June 23-28*, pages 3310–3317, 2014.

[71] Jiayan Jiang, Bo Wang, and Zhuowen Tu. Unsupervised metric learning by self-smoothing operator. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 794–801, 2011.

[72] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I*, pages 685–701, 2016.

[73] 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.

[74] Sepandar D. Kamvar, Dan Klein, and Christopher D. Manning. Spectral learning. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 561–566, 2003.

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

[76] Peter Kontschieder, Michael Donoser, and Horst Bischof. Beyond pairwise shape similarity analysis. In *ACCV*, pages 655–666, 2009.

[77] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPSAdvances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012.

[78] Brian Kulis, Sugato Basu, Inderjit S. Dhillon, and Raymond J. Mooney. Semi-supervised graph clustering: A kernel approach. *Machine Learning*, 74(1):1–22, 2009.

[79] Stéphane Lafon and Ann B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1393–1403, 2006.

[80] C. Lee, W.D. Jang, J.Y. Sim, and Chang-Su Kim. Multiple random walkers and their application to image cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3837–3845, 2015.

[81] Victor S. Lempitsky, Pushmeet Kohli, Carsten Rother, and Toby Sharp. Image segmentation with a bounding box prior. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 277–284, 2009.

[82] Thomas K. Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Computer Vision*, 43(1):29–44, 2001.

[83] Hongliang Li, Fanman Meng, and King Ngi Ngan. Co-salient object detection from multiple images. *IEEE Trans. Multimedia*, 15(8):1896–1909, 2013.

[84] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Computer Vision - ACCV 2012 - 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I*, pages 31–44, 2012.

[85] 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.

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

[87] Yin Li, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Lazy snapping. *ACM Trans. Graph.*, 23(3), 2004.

[88] Haibin Ling and David W. Jacobs. Shape classification using the inner-distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):286–299.

[89] Haibin Ling, Xingwei Yang, and Longin Jan Latecki. Balancing deformability and discriminability for shape matching. In *Computer Vision - ECCV 2010 - 8th European Conference*, pages 411–424, 2010.

[90] Jiangyu Liu, Jian Sun, and Heung-Yeung Shum. Paint selection. *ACM Trans. Graph.*, 28(3), 2009.

[91] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[92] 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.

[93] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *CVPR*, pages 1988–1995. IEEE Computer Society, 2009.

[94] David G Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*. Springer, New York, 2008.

[95] Lin Ma, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Multiple feature fusion via weighted entropy for visual tracking. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3128–3136, 2015.

[96] Subhransu Maji, Nisheeth K. Vishnoi, and Jitendra Malik. Biased normalized cuts. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 2057–2064, 2011.

[97] Kevin McGuinness and Noel E. O'Connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010.

[98] Benjamin Bruno Meier, Ismail Elezi, Mohammadreza Amirian, Oliver Dürr, and Thilo Stadelmann. Learning neural models for end-to-end clustering. In *Artificial Neural Networks in Pattern Recognition - 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19-21, 2018, Proceedings*, pages 126–138, 2018.

[99] Eyasu Zemene Mequanint, Leulseged Tesfaye Alemu, and Marcello Pelillo. Dominant sets for "constrained" image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(10):2438–2451, 2019.

[100] Cho Minsu and MuLee Kyoung. Authority-shift clustering: Hierarchical clustering by authority seeking on graphs. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 3193–3200, 2010.

[101] Konda Reddy Mopuri and R. Venkatesh Babu. Object level deep feature pooling for compact image representation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Boston, MA, USA, June 7-12, 2015*, pages 62–70, 2015.

[102] Eric N. Mortensen and William A. Barrett. Interactive segmentation with intelligent scissors. *Graphical Models and Image Processing*, 60(5):349–384, 1998.

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

[104] M.Rubinstein, A.Joulin, J.Kopf, and C.Liu. Unsupervised joint object discovery and segmentation in internet images. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 1939–1946, 2013.

[105] David Nistér and Henrik Stewénius. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, 17-22 June 2006, New York, NY, USA*, pages 2161–2168, 2006.

[106] Massimiliano Pavan and Marcello Pelillo. Dominant sets and hierarchical clustering. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 362–369, 2003.

[107] Massimiliano Pavan and Marcello Pelillo. A new graph-theoretic approach to clustering and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 145–152, 2003.

[108] Marcello Pelillo and Andrea Torsello. Payoff-monotonic game dynamics and the maximum clique problem. *Neural Computation*, 18(5):1215–1258, 2006.

Bibliography

[109] Florent Perronnin and Christopher R. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007.

[110] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007.

[111] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, 24-26 June 2008, Anchorage, Alaska, USA*, 2008.

[112] Brian L. Price, Bryan S. Morse, and Scott Cohen. Geodesic graph cut for interactive image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 3161–3168, 2010.

[113] Alexis Protiere and Guillermo Sapiro. Interactive image segmentation via adaptive weighted distances. *IEEE Trans. Image Processing*, 16(4):1046–1057, 2007.

[114] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc J. Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 777–784, 2011.

[115] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 3–20, 2016.

[116] S. Rota Bulò and M. Pelillo. Dominant-set clustering: A review. *European Journal of Operational Research*, 262(1):1–13, 2017.

[117] S. Rota Bulò, M. Pelillo, and I.M. Bomze. Graph-based quadratic optimization: A fast evolutionary approach. *Computer Vision and Image Understanding*, 115(7):984–995, 2011.

[118] S. Rota Bulò, Marcello Pelillo, and Immanuel M. Bomze. Graph-based quadratic optimization: A fast evolutionary approach. *Computer Vision and Image Understanding*, 115(7):984–995, 2011.

[119] Samuel Rota Bulò and Marcello Pelillo. A game-theoretic approach to hypergraph clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(6):1312–1327, 2013.

[120] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "Grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[121] Philippe Salembier and Luis Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Trans. Image Processing*, 9(4):561–576, 2000.

[122] Ozan Sener, Kemal Ugur, and A. Aydin Alatan. Error-tolerant interactive image segmentation using dynamic and iterated graph-cuts. In *IMMPD@ACM Multimedia*, pages 9–16, 2012.

[123] Shannon. *A Mathematical Theory of Communication*. 1948.

[124] 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.

[125] 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.

[126] 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.

[127] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[128] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*, pages 1470–1477, 2003.

[129] Kartic Subr, Sylvain Paris, Cyril Soler, and Jan Kautz. Accurate binary image selection from inaccurate user input. *Comput. Graph. Forum*, 32(2):41–50, 2013.

[130] 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.

[131] S.Vicente, C.Rother, and V.Kolmogorov. Object cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 2217–2224, 2011.

[132] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, 2011.

[133] Raphael Sznitman, Carlos J. Becker, François Fleuret, and Pascal Fua. Fast object detection with entropy-driven evaluation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland CVPR, OR, USA, June 23-28, 2013*, pages 3270–3277, 2013.

[134] Martin Szummer and Tommi S. Jaakkola. Partially labeled classification with markov random walks. In *NIPS*, pages 945–952, 2001.

[135] Meng Tang, Ismail Ben Ayed, and Yuri Boykov. Pseudo-bound optimization for binary energies. In *Computer Vision - ECCV 2014- 12th European Conference*, pages 691–707, 2014.

[136] Meng Tang, Lena Gorelick, Olga Veksler, and Yuri Boykov. Grabcut in one cut. In *IEEE International Conference on Computer Vision, ICCV*, pages 1769–1776, 2013.

[137] Alemu Leulseged Tesfaye. Constrained dominant sets for image retrieval. *MSc Thesis, Universita' Ca' Foscari*, 2016.

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

[139] Y.T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv*, abs/1706.06196, 2017.

[140] Giorgos Tolias and Hervé Jégou. Visual query expansion with or without geometry: Refining local descriptors by feature aggregation. *Pattern Recognition*, 47(10):3466–3476, 2014.

[141] 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.

[142] Bo Wang and Zhuowen Tu. Affinity learning via self-diffusion for image segmentation and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 2312–2319, 2012.

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

[144] Jingdong Wang, Yangqing Jia, Xian-Sheng Hua, Changshui Zhang, and Long Quan. Normalized tree partitioning for image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, 2008.

[145] 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.

[146] 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.

[147] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, pages 79–88. IEEE Computer Society, 2018.

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

[149] Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, and Zhuowen Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 256–263, 2014.

[150] Zhenyu Wu and Richard M. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(11):1101–1113, 1993.

[151] W.Zhu, S.Liang, Y.Wei, and J.Sun. Saliency optimization from robust background detection. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 2814–2821, 2014.

[152] X.Cao, Z.Tao, B.Zhang, H.Fu, and W.Feng. Self-adaptively weighted co-saliency detection via rank constraint. *IEEE Trans. Image Processing*, 23(9):4175–4186, 2014.

[153] X.Dong, J.Shen, L.Shao, and M.H.Yang. Interactive cosegmentation using global and local energy optimization. *IEEE Trans. Image Processing*, 24(11):3966–3977, 2015.

[154] Min Xian, Yingtao Zhang, H. D. Cheng, Fei Xu, and Jianrui Ding. Neutro-connectedness cut. *CoRR*, abs/1512.06285.

[155] Jianxiong Xiao and Long Quan. Multiple view semantic segmentation for street view images. In *Proc. IEEE International Conference on Computer Vision, ICCV*, pages 686–693, 2009.

[156] Jian Xu, Cunzhao Shi, Cheng-Zuo Qi, Chunheng Wang, and Baihua Xiao. Part-based weighting aggregation of deep convolutional features for image retrieval. *CoRR*, abs/1705.01247, 2017.

[157] 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.

[158] Fan Yang, Bogdan Matei, and Larry S. Davis. Re-ranking by multi-feature fusion with diffusion for image retrieval. In *2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, Waikoloa, HI, USA, January 5-9, 2015*, pages 572–579, 2015.

[159] Wenxian Yang, Jianfei Cai, Jianmin Zheng, and Jiebo Luo. User-friendly interactive image segmentation through unified combinatorial user inputs. *IEEE Trans. Image Processing*, 19(9):2470–2479, 2010.

[160] Xingwei Yang, Suzan Köknar-Tezel, and Longin Jan Latecki. Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 357–364, 2009.

[161] Xingwei Yang and Longin Jan Latecki. Affinity learning on a tensor product graph with applications to shape and image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*.

[162] Xingwei Yang, Lakshman Prasad, and Longin Jan Latecki. Affinity learning with diffusion on tensor product graph. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):28–38, 2013.

[163] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):723–742, 2012.

[164] Hongkai Yu, Youjie Zhou, Hui Qian, Min Xian, Yuewei Lin, Dazhou Guo, Kang Zheng, Kareem Abdelfatah, and Song Wang. Loosecut: Interactive image segmentation with loosely bounded boxes. *CoRR*, abs/1507.03060, 2015.

[165] Stella X. Yu and Jianbo Shi. Segmentation given partial grouping constraints. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):173–183, 2004.

[166] Charles T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Computers*, 20(1):68–86, 1971.

[167] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2004.

[168] E. Zemene, Y. Tariku, H. Idrees, A. Prati, M. Pelillo, and M. Shah. Large-scale image geo-localization using dominant sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[169] Eyasu Zemene, Leulseged Tesfaye Alemu, and Marcello Pelillo. Constrained dominant sets for retrieval. In *ICPR 2016*, pages 76–81, 2014.

[170] 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.

[171] Eyasu Zemene and Marcello Pelillo. Path-based dominant-set clustering. In *Proc. 18th International Conference on Image Analysis and Processing (ICIAP), Part I*, pages 150–160, 2015.

[172] 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.

[173] 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.

[174] Shaoting Zhang, Ming Yang, Timothée Cour, Kai Yu, and Dimitris N. Metaxas. Query specific rank fusion for image retrieval. *IEEE, Trans. Pattern Anal. Mach. Intell.*, 37(4):803–815, 2015.

[175] Shiliang Zhang, Ming Yang, Xiaoyu Wang, Yuanqing Lin, and Qi Tian. Semantic-aware co-indexing for image retrieval. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1673–1680, 2013.

[176] 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.

Bibliography

[177] 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.

[178] Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian. Lp-norm IDF for large scale image search. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland CVPR, OR, USA, June 23-28, 2013*, pages 1626–1633, 2013.

[179] Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian. Packing and padding: Coupled multi-index for accurate image retrieval. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Columbus, OH, USA, June 23-28, 2014*, pages 1947–1954, 2014.

[180] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 1741–1750, 2015.

[181] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, pages 1–11. British Machine Vision Association, 2009.

[182] 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.

[183] 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.

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

[185] Youjie Zhou, Lili Ju, and Song Wang. Multiscale superpixels and supervoxels based on hierarchical edge-weighted centroidal voronoi tessellation. In *WACV*, pages 1076–1083, 2015.