Simultaneously Discovering and Localizing Common Objects in Wild Images

Zhenzhen Wang and Junsong Yuan, *Senior Member, IEEE*

***Abstract*— Motivated by the recent success of supervised and weakly supervised common object discovery, in this paper, we move forward one step further to tackle common object discovery in a fully unsupervised way. Generally, object co-localization aims at simultaneously localizing objects of the same class across a group of images. Traditional object localization/detection usually trains specific object detectors which require bounding box annotations of object instances, or at least image-level labels to indicate the presence/absence of objects in an image. Given a collection of images without any annotations, our proposed fully unsupervised method is to simultaneously discover images that contain common objects and also localize common objects in corresponding images. Without requiring to know the total number of common objects, we formulate this unsupervised object discovery as a sub-graph mining problem from a weighted graph of object proposals, where nodes correspond to object proposals, and edges represent the similarities between neighbouring proposals. The positive images and common objects are jointly discovered by finding sub-graphs of strongly connected nodes, with each sub-graph capturing one object pattern. The optimization problem can be efficiently solved by our proposed maximal-flow-based algorithm. Instead of assuming that each image contains only one common object, our proposed solution can better address wild images where each image may contain multiple common objects or even no common object. Moreover, our proposed method can be easily tailored to the task of image retrieval in which the nodes correspond to the similarity between query and reference images. Extensive experiments on PASCAL VOC 2007 and Object Discovery data sets demonstrate that even without any supervision, our approach can discover/localize common objects of various classes in the presence of scale, view point, appearance variation, and partial occlusions. We also conduct broad experiments on image retrieval benchmarks, Holidays and Oxford5k data sets, to show that our proposed method, which considers both the similarity between query and reference images and also similarities among reference images, can help to improve the retrieval results significantly.**

***Index Terms*— Common object discovery, image retrieval, unsupervised learning, sub-graph mining.**

Manuscript received July 31, 2017; revised January 19, 2018; accepted May 7, 2018. Date of publication May 25, 2018; date of current version June 11, 2018. This work was supported in part by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE2015-T2-2-114 and in part by start-up grants of the Computer Science and Engineering Department, University at Buffalo. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jie Liang. *(Corresponding author: Zhenzhen Wang.)*

Z. Wang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: zwang033@e.ntu.edu.sg).

J. Yuan is with the Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260-2500 USA (e-mail: jsyuan@buffalo.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2839901

I. Introduction

LOCALIZING and detecting objects in images are among the most widely studied computer vision problems. They are quite challenging due to intra-class variation, inter-class diversity, and noisy annotations, especially in wild images. Thus, a large body of fully/strongly annotated data is crucial to train detectors to achieve satisfactory performance. However, manually labeling the presence of objects and even their locations in images is time-consuming, expensive and laborious. On the other hand, the explosive visual data available at almost no cost on websites such as Flickr and Facebook have not been fully utilized. Therefore, building object localization and detection frameworks with weak supervision or even no supervision has been of great interest in recent years.

Weakly-supervised object localization (WSOL) [1], [2] has drawn much attention recently. It aims at localizing common objects across images using the annotations to indicate the presence/absence of the objects of interest. In this paper, we focus on simultaneously discovering and localizing common objects in real-world images, which shares the same type of output as WSOL, but does not require the annotation of presence/absence of objects. In addition, we tackle this problem in a more challenging scenario where (1) multiple common object classes are contained in the given collection of images, which means this is a totally unsupervised problem; (2) multiple objects or even no object is contained in some of the images (see Fig. 1), so that there may exist some outliers which do not contain any common objects. It shows in our experiments that even when the outlier ratio reaches as high as 60%, our method can still achieve satisfactory performance. The detailed comparisons of the required supervision for weakly-/un-supervised methods and our proposed method are shown in Table I. It can be clearly seen that, our proposed method could save all annotations on images and bounding boxes and even the number of categories in the given image set, while allowing high rate of noisy images.

To localize the objects in images, we propose to approach the common object discovery and localization in images based on off-the-shelf object proposals. Many common patterns discovery methods [8]–[10] usually treat the objects as a composition of visual primitives, *e.g.*, a bicycle can be composed by one triangle primitive and two circle primitives. In this paper, we treat each object as a whole and aim to find clusters of proposals which appear frequently and contain the whole object. To this end, we formulate the common object class discovery process as a constrained sub-graph mining

TABLE I
COMPARISON OF DIFFERENT SUPERVISION REQUIRED FOR THE WEAKLY-/UN-SUPERVISED METHODS AND OURS

	Wk.-supervised [3], [4], [1]	Un.-supervised [5], [6], [7]	Ours
Image annotations	Yes	No	No
Bounding box annotations	No	No	No
# object categories	One	Many	Many
# noisy images	Not included or only a few	Not included or only a few	Included, as high as 60%

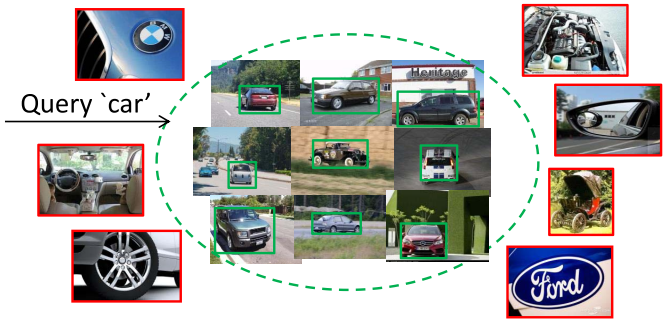


Fig. 1. Our goal is to simultaneously discover images that contain common objects and also localize common objects in corresponding images. (Better viewed in color.)

problem, which characterizes a whole image group as a graph composed of the object proposals as nodes and the similarities between the proposals as edges. Different from traditional sub-graph mining methods which rely on either edges or nodes to perform solution, the proposed method considers both factors thus can bring better performance. Moreover, our proposed method can be easily extended to the task of image/instance retrieval where the nodes of the graph are defined as the similarity between query and reference images, and the edges are the similarities between each pair of the reference images.

Our proposed method is evaluated on PASCAL VOC 2007 (part and all) and Object Discovery datasets in terms of the detected locations of common objects, and on Holidays and Oxford5k datasets in terms of the retrieval results. The first part of experiments show that our approach not only can discover the images that contain the objects while removing the outlier images, but also can localize common objects effectively even without any image annotations. The experiments on image/instance retrieval further demonstrate the effectiveness of our proposed method. To validate the robustness of our method, we also evaluate our method in a more challenging scenario where the number of outlier images (without any common objects in it) accounts for a high proportion (*e.g.*, as high as 60%). To summarize, our main contributions are three folds:

- A new sub-graph selection objective function for unsupervised common object discovery is proposed to formulate the scenario where the number of objects contained in each image is unconstrained.
- A novel solution which is inspired by the maximal flow algorithm is proposed to optimize the sub-graph selection problem efficiently.
- A very first attempt to simultaneously perform common object discovery and localization with high outlier ratios (*e.g.*, as high as 60%) in a fully unsupervised situation.

The remaining of the paper is organized as follows: Section II briefly reviews the development of common object discovery strategies and the sub-graph selection based methods. Section III describes the proposed approach, *i.e.*, the problem formulation and the optimization process, in detail. Experimental results on and implementation details on co-localization and image retrieval are elaborated in Section IV and section V, respectively. Finally, we conclude this paper in Section VI.

II. RELATED WORK

Common object discovery plays an important role in computer vision tasks, such as object detection/localization [2], [11], [12], image co-segmentation/saliency [13], [14]. The main paradigm of these tasks are similar: the inputs are usually real-world images with incomplete labels or sometimes even without any supervision information, then the key step is to discover the most frequently occurring pattern by methods such as local feature matching, sub-graph mining, etc. Based on the results of pattern discovery, the outputs differ according to the targets of tasks, for example, detection/localization draws bounding boxes around objects, co-saliency and co-segmentation predict pixel-wise labels. In the following, we will review some representative studies on the common object discovery according to the different supervision required, and also its extension to image/instance retrieval, then the strategies of the sub-graph mining which inspire us a novel solution to the proposed method.

A. Weakly-Supervised Object Discovery

Weakly-supervised strategies are widely used in the tasks such as object detection/localization [1], [2], [11], image co-segmentation/saliency [14]–[17]. Object discovery from videos [18]–[20] can also be seen as weakly-supervised methods since we only know the frame-level labels. Although many studies claim to be weakly-supervised methods, the supervision required is a little different. Some methods require the image-level labels to indicate the categories, while some only need the binary labels to indicate whether there exists an object or not. No matter which case, obtaining the image-level labels is much easier than annotating object locations or pixel-wise labels. Some studies [21], [22] on mid-level representations also adopt the weakly-supervised skill by requiring only the image labels rather than the patch labels, so that they can leverage the large amount of tagged images on the Internet.

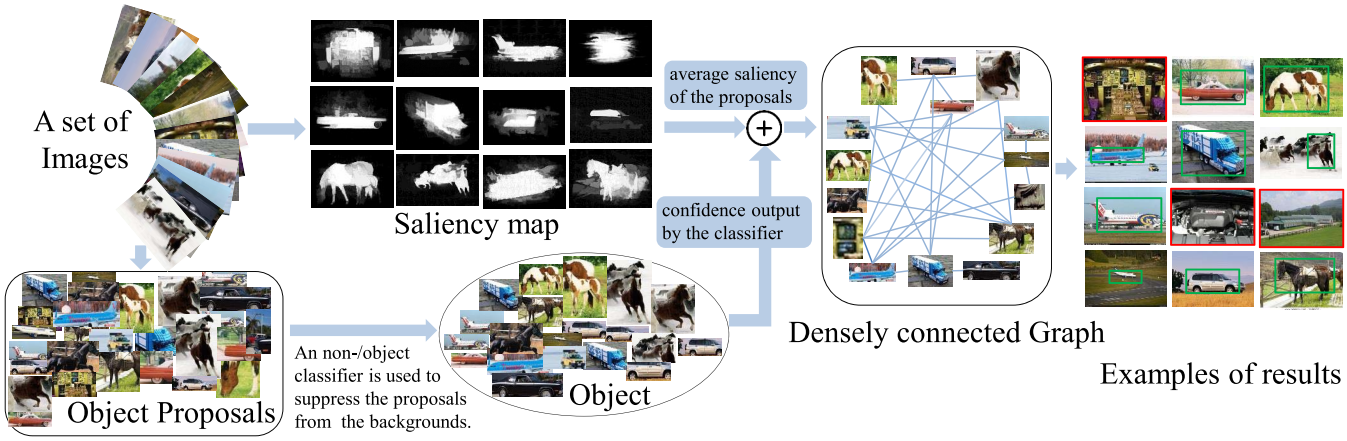


Fig. 2. Schematic process of the key working mechanism. Given a set of images, we first extract the object proposals and then these proposals are filtered by a pretrained two-class classifier [23] which better distinguishes the objects from backgrounds. The weighted combination of the proposal's confidence output from the classifier together with the average saliency within the proposal is the final probability of the proposals containing an object. Then we build a densely connected graph which takes the scored proposals as nodes, and the similarities between these object proposals as the edges. Two models are proposed to handle different scenarios where 1) a given image group contains multiple common object classes, but each image contains at most one target object; 2) an individual image and the image group are both likely to contain multiple common objects. We only show the results under the first assumption in this figure; the common objects are highlighted in green and the outlier images are in red. (Better viewed in color.)

B. Unsupervised Object Discovery

More recently, researchers tend to study object discovery tasks in wild images without any annotations, including noisy images without target objects and images containing multiple objects. The most representative studies on common object co-localization are [5]–[7], all the methods together with our proposed method all take advantage of the object proposals. However, these methods are based on the assumption that there is only one target object in each positive image, and the image groups are usually without any outliers or with a rather low outlier ratio. Since these restrictions are not always satisfied in real-world image sets, our proposed method is extended to tackle scenarios where individual images could contain multiple target objects, and also scenarios with high outlier ratios (*e.g.*, as high as 60%). Clustering-based methods [24], [25] are also well studied for classification or co-segmentation, they can also be used as a post-processing of the co-localization to group the patches in the same category.

C. Object Instance/Image Search

The goal of the image/instance retrieval is to select the images, which are similar to or contain the same instance as the given query, from a set of reference images. The difference of the image retrieval and instance retrieval is that the former prefers to select images that are globally similar to the query [26], [27], while the latter cares more about the foreground objects contained in the reference images [28], [29]. The resulting ranks of the two tasks both depend on the similarity to the query, and usually there will be a re-ranking stage to refine the search results, such as AML [29] and QE [30]. However, it is widely acknowledged that if a reference $\mathcal{I}$ is associated with a query $\mathcal{Q}$, then it is possible that the reference images which enjoy high similarity with $\mathcal{I}$ may also be associated with $\mathcal{Q}$, and vice versa. Thus, our proposed framework, which employs both the similarity between query and reference images and the similarities among reference images, can help to improve the resulting ranks by adding or removing some outlier images from the top-ranked list.

D. Sub-Graph Selection

Frequent feature selection with graph [31], [32] is one of the arms of data mining, it has long been used in computer vision for common object/pattern. For example, the methods of sub-graph mining have been introduced to solve the task of thematic pattern discovery in videos [33], [34], where each node of the graph is represented by visual words and the affinity graph is built based on the relationships of these visual words. The sub-set of visual words having the maximum overall mutual information scores is selected. Although the algorithms work well in their scenarios, to find a sub-graph only depends on graph edges but neglects the weights of nodes is obviously inadequate for our settings. For instance, given a group of images containing airplane, the patches containing sky can also be the frequently occurring sub-sets since the airplane usually occurs together with sky. In order to solve such confusion, we propose to apply prior knowledge like saliency, ratio of the width to the height to discriminate the foreground objects out of backgrounds if they have similar frequency.

III. Methods

In this section, we will present our proposed approach as shown in Fig. II. Given a collection of real world images which may contain outlier/negative images, our goal is to simultaneously discover images that contain common objects and also localize common objects in corresponding images. Moreover, to make the proposed method applicable to wild images, we do not assume to know the total number of object classes in the given dataset, as well as the number of objects that occur in one single image.

A. Simultaneous Object Discovery and Localization

To localize the common objects in wild images, we use off-the-shelf object proposals at the first step, then the key problems are 1) to discover the images that contain at least one common object while eliminating the outlier images; and 2) to locate the proposals corresponding to the common objects. To achieve these two goals simultaneously, we formulate the discovery process as a sub-graph mining problem, where the whole graph is built based on all the object proposals from all images. Once the subset of proposals containing common objects is found, we can easily identify the images that contain these proposals, and the locations of objects in the images are where the corresponding proposals extracted. In the following, we will introduce how to build graphs in detail based on two different assumptions: 1) there is at most one common object in a single image, which is the most popular setting in previous works and 2) there may be multiple common objects in each image, which is a more challenging scenario.

Given a set of images $\{\mathcal{I}_i\}_{i=1}^N$, and a set of proposals $\{\mathcal{P}_j\}_{j=1}^M$ extracted from all images, we denote an image as positive if it contains the common object; similarly, a proposal is called positive if it corresponds to a common object. Two vectors $\boldsymbol{g} \in \{0,1\}^N$, $\boldsymbol{f} \in \{0,1\}^M$ are used to infer the labels (positive or negative) of the images and proposals, respectively. As our task is to simultaneously discover and locate common objects, we need to infer $\boldsymbol{g}$ and $\boldsymbol{f}$ from the whole image dataset together. It is clear that an image should be labeled positive if one of the proposals extracted from it is positive. The relation between $\boldsymbol{f}$ and $\boldsymbol{g}$ can thus be captured by a binary matrix $A \in \mathbb{R}^{M\times N}$, where $A_{j,i} = 1$ if the jth proposal is extracted from the ith image. The correlation of all proposals can be represented by a weighted affinity graph, where each edge represents the similarity between two proposals, and the weight of each node represents the probability of being positive.

Without loss of generality, we first assume that each positive image contains one common object: this scenario is termed as *'Single Instance version of Unsupervised Object Localization (S-UOL).'* Then the following constraint must hold: $A^\top \boldsymbol{f} = \boldsymbol{g}$. This constraint guarantees that there must be one positive proposal ($f_j = 1$) being selected in each positive image ($g_i = 1$), while no positive proposal ($f_j = 0$) will be selected in each negative image ($g_i = 0$). Then finding a sub-graph of positive proposals is to simultaneously infer variables $\boldsymbol{f}, \boldsymbol{g}$ which maximize the following objective:

$$\begin{aligned} &\max_{\boldsymbol{f},\boldsymbol{g}} \boldsymbol{c}^\top \boldsymbol{f} - \lambda \boldsymbol{f}^\top L \boldsymbol{f} - \eta \|\boldsymbol{g}\|_0 \\ &\text{s.t. } A^\top \boldsymbol{f} = \boldsymbol{g} \end{aligned} \tag{1}$$

where each element c_j in $\boldsymbol{c} \in \mathbb{R}^M$ is confidence of the jth proposal being positive. The first term $\boldsymbol{c}^\top \boldsymbol{f}$ aims at maximizing the cumulative score of the selected sub-graph. Laplacian matrix L is defined as $L = D - W$, where $W \in \mathbb{R}^{M\times M}$ is the similarity matrix based on the proposal features, and D is the diagonal matrix in which $D_{j,j} = \sum_{k=1}^M W_{j,k}$. The second term $\boldsymbol{f}^\top L \boldsymbol{f}$ is to minimize the effect of negative proposals by emphasizing more on the edge connections with higher weights, and the parameter λ controls the influence of this connectivity. The last term $\|\boldsymbol{g}\|_0$ attempts to eliminate negative images. The reason for combining the outlier discovery and object localization into a joint formulation is that the two processes can help each other during learning, *e.g.*, the more frequently occurring images are more likely to contain common objects, and vice versa. It is worth noting that most previous studies [5], [6] only consider image sets without outliers or with a small proportion of outliers, however, our method can handle data with a large portion of outlier images.

The above formulation can be easily extended to address a more general scenario where for a group of images, some of them may contain multiple common objects. We call this setting as *"Multiple Instance version of Unsupervised Object Localization (M-UOL)"*. Then the objective can be written as:

$$\max_{\boldsymbol{f}} \boldsymbol{c}^\top \boldsymbol{f} - \lambda \boldsymbol{f}^\top L \boldsymbol{f} - \eta \|\boldsymbol{f}\|_0 \tag{2}$$

Definitions of variables are the same as Eq. (1), but we give up the constraint as it is unknown how many positive proposals each positive image may contain. Note that we reward sparse solutions by the last term due to there are dozens of proposals extracted from each image and the majority of them tend to be negative ones, thus a sparsity regularizer on $\boldsymbol{f}$ is still necessary for co-localization problems to avoid selecting large numbers of proposals. The parameter η controls the influence of sparsity. The *M-UOL* model can also be applied to the problem of image and instance retrieval, where the node confidences represent how relevant the reference images and the query are, and the similarity matrix W is composed of the similarity between reference images. Different from traditional retrieval methods which only consider the relation between query and reference images, our method can also take advantage of the relations among the reference images.

Compared with most of the previous graph-based methods, we make a significant improvement by leveraging both node and edge weights to build graphs. The roles of the nodes and edges on the sub-graph mining results are shown in Fig. 3. Based on our formulation, the isolated nodes even with high possibilities of containing an object are less likely selected, such as the node corresponding to a bike (not a common object) in Fig. 3. A set of cohesively connected nodes may also not be the ideal sub-graph if the weight of each node is low (*i.e.*, less likely to be objects), such as the cluster of sky highlighted in red in Fig. 3. Only nodes enjoying high confidence of being the objects and correlating cohesively with each other are selected (*e.g.*, patterns of airplanes and cats), which are exactly the first and second terms in Eq. (1) and Eq. (2) designed for.

B. Optimization

To solve the sub-graph mining problems defined in Eq. (1) and Eq. (2), we resort to the maximal flow algorithm proposed by Boykov and Kolmogorov [35]. It has been demonstrated in [36] and [37] that a minimum cut can be efficiently computed with the maximum flow algorithm, meanwhile, it can be proved that our objective functions are combinations of

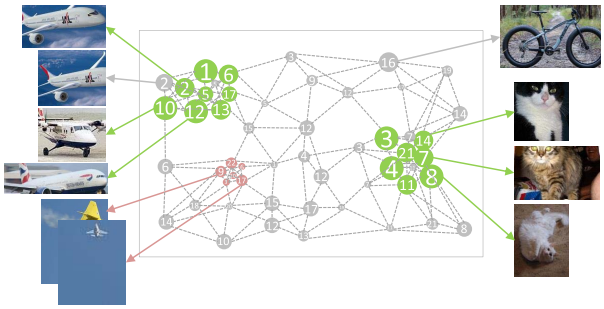


Fig. 3. Illustration of sub-graph mining from a densely connected graph. The nodes of the graph are represented by circles in different sizes to indicate the confidence of proposals containing common objects. The different numbers in circles denote the proposals are extracted from different images. The edges of the graph are represented by the distances between circles, the more similar two proposals are, the closer they are in the figure. Each node is connected to the nearest 5 neighbors. The resulting sub-set of nodes are highlighted by green which enjoying high confidences and are cohesive with each other. Although close to each other, the red nodes are unlikely to contain object due to low weights of these nodes, and some of the grey ones in spite of enjoy high confidences but are unlikely to be common objects since they are unique to other nodes. (Better viewed in color with magnification.)

cut-functions, so that to optimize our objective functions is equivalent to find a min-cut on a graph.

Proof: Given a graph $\mathcal{G}$ over nodes $V = \{1, \ldots, M\}$, the goal of minimum cut is to find a sub-set $S \subset V$ that minimizes the cut-function $\sum_{p\in S}\sum_{q\notin S} E_{p,q} = \sum_{p=1}^{M}\sum_{q=1}^{M} f_p(1-f_q)E_{p,q}$, where E denotes the adjacency matrix of the graph, f_p is 1 if $p \in S$ and 0 otherwise. The graph-regularized subset selection problem formalized in Eq. (1) can be represented as following by substituting the constraint into the objective function:

$$\min_{\boldsymbol{f}\in\{0,1\}^M} \left(\eta\|A^\top \boldsymbol{f}\|_0 - \boldsymbol{c}^\top \boldsymbol{f}\right) + \lambda \boldsymbol{f}^\top L \boldsymbol{f} \tag{3}$$

By introducing artificial nodes s and t to the graph, the first term can be encoded as a *cut-function*:

$$\begin{aligned}
&\eta\|A^\top \boldsymbol{f}\|_0 - \boldsymbol{c}^\top \boldsymbol{f} \\
&= \eta \sum_{i=1}^{N}\sum_{p=1}^{M} A_{p,i} f_p - \sum_{p=1}^{M} c_p f_p \\
&= \eta \sum_{p=1}^{M} f_p \sum_{i=1}^{N} A_{p,i} - \sum_{p=1}^{M} c_p f_p \\
&= \eta \sum_{p=1}^{M} f_p - \sum_{p=1}^{M} c_p f_p \\
&= \sum_{p=1}^{M} (\eta - c_p) f_p \\
&= \sum_{p\in S, c_p<\eta} (\eta - c_p) + \sum_{p\in V, c_p\geq\eta} (\eta - c_p) - \sum_{p\notin S, c_p\geq\eta} (\eta - c_p) \\
&= \sum_{p=1}^{M} E_{s,p} f_s (1-f_p) + \sum_{p=1}^{M} E_{p,t} f_p (1-f_t) + \mathcal{C}
\end{aligned} \tag{4}$$

where $\mathcal{C} = \sum_{p\in V, c_p\geq\eta} (\eta - c_p)$ is a constant, $f_s = 1$, $f_t = 0$ and the edges of the two augmented nodes s and t are given by $E_{p,q} = \lambda W_{p,q}$ for $1 \leq p, q \leq n$ and $E_{s,p} = \max\{c_p - \eta, 0\}$ and $E_{t,p} = \eta - c_p + E_{s,p}$, W is the adjacency matrix of the original graph without augmentation. As $f_s = 1$ and $f_t = 0$ enforce that $s \in S$ and $t \notin S$, it follows that Eq. (1) is an s/t min-cut problem on the transformed graph defined by the adjacency matrix E over the nodes of $\mathcal{G}$ augmented by s and t. The aforementioned situation still holds if W is a weighted adjacency matrix, so the min-cut reformulation can also be applied to a weighted network.

The second term of Eq. (3) can be easily represented as a cut-function over graph $\mathcal{G}$:

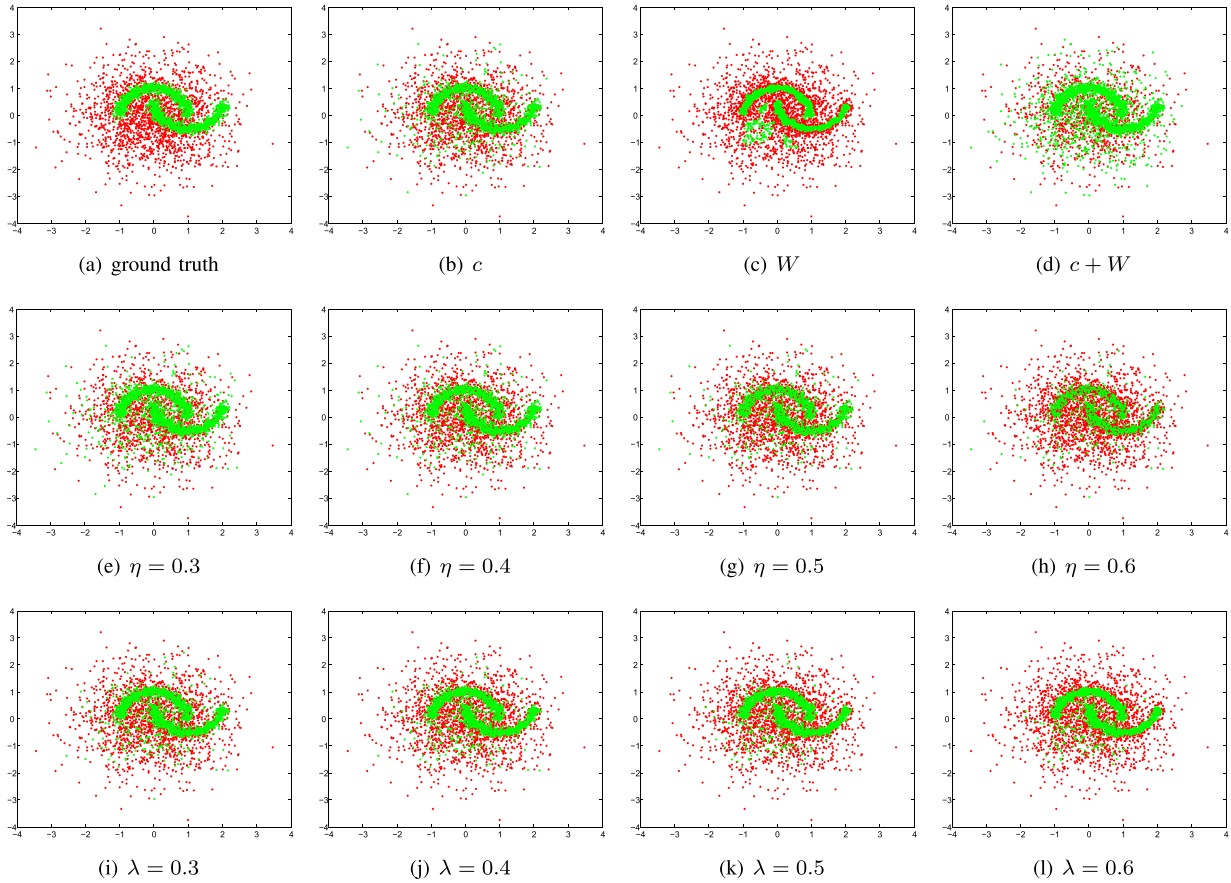
$$\begin{aligned}
\boldsymbol{f}^\top L \boldsymbol{f} &= \sum_{p=1}^{M} f_p \left(D_{p,p} - \sum_{q=1}^{M} W_{p,q} f_q\right) \\
&= \sum_{p=1}^{M}\sum_{q=1}^{M} W_{p,q} f_p \left(1 - f_q\right)
\end{aligned} \tag{5}$$

Similarly, it can be proved that the objective Eq. (2) can also be represented as the combination of two cut-functions. □

It is clearly shown from the above proof that both of the proposed objectives could be represented by cut-functions, thus could be optimized by maximal flow algorithms. In our implementation, we use Boykov-Kolmogorov algorithm [35] to solve our node selection problems. Although the worst case complexity is in $\mathcal{O}(M^2en)$, where e represents the number of edges in the graph and n is the size of the minimum cut, it performs efficiently in practice since the graph is rather sparse. Comparing our *S-UOL* Model with the joint model proposed in [5], one can find that our method can directly output the binary solution while [5] needs to set an extra threshold for the binarization of their real-value solution.

C. Ablation Simulation

To visualize the effectiveness of our proposed framework and its components, we run a test experiment on some simulated 2D data points (see Fig. 4) based on the general assumption that there may exist multiple common objects in a single positive image. In this simulation experiment, the node confidence is set as its density calculated following [38], and the similarity matrix is calculated by $W_{k,l} = \exp(-\|\boldsymbol{x}_{i,k} - \boldsymbol{x}_{j,l}\|^2)$. The affinity graph is constructed in the same way as described in Section III-A. The proposed sub-graph mining technique is tested on these data, and the results are shown in Fig. 4. Each point represents a node, and the green ones are positive, while red ones are negative. There are 1600 positive points and 2000 negative points in total. In order to show the relative importance of each part in the affinity graph formulated in Eq. (2), we display the results, which are optimized based on the nodes only (Unary term in Eq. (2)), on the edges only, and on both the nodes and edges (Unary and Laplacian term in Eq. (2)), in the first row. Fig. 4(b) pictures the case where we treat the top 2000 points with higher node confidences as positive. It can be seen that some false positives will be chosen if we only rely on the node scores. Fig. 4(c) shows results of the 1600-nearest-neighbor degree of the nodes, although the majority of the positive



(a) ground truth (b) c (c) W (d) $c + W$

(e) $\eta = 0.3$ (f) $\eta = 0.4$ (g) $\eta = 0.5$ (h) $\eta = 0.6$

(i) $\lambda = 0.3$ (j) $\lambda = 0.4$ (k) $\lambda = 0.5$ (l) $\lambda = 0.6$

Fig. 4. The simulation results using *M-UOL* Model. **Top:** Groundtruth, and results using only node confidence, edge similarity and both. **Middle**: Results vary in terms of different values of η with the optimal λ (grid-search on ranges of values). **Bottom**: Results vary in terms of different values of λ with η being fixed as 0.3. (Better viewed in color.) (a) ground truth. (b) c. (c) W. (c) $c + W$. (d) $\eta = 0.3$. (e) $\eta = 0.4$. (f) $\eta = 0.5$. (g) $\eta = 0.6$. (h) $\lambda = 0.3$. (i) $\lambda = 0.4$. (j) $\lambda = 0.5$. (k) $\lambda = 0.6$.

points are choose, the rate of false negative is rather high. From fig. 4(d), which shows the case of depends on both nodes and edges but without the regularization term, we can find that almost all the positive points are correctly recalled, but the false positive rate is so high, thus a regularization term is necessary for removing those false positive points.

The results on the joint objective (Eq. (2)) are visualized in the second and third row, with various η and λ showing the influence of the two parameters on the results. Compared with fig. 4(d), it can be found that with the regularization term, most of the false positive points can be eliminated, and the coefficient η controls how sparsity of the results $\boldsymbol{f}$ could be. The parameter λ influences the results by tuning the relative importance of the node confidences and the similarities between nodes. We should note that there is no standard metric for common object or common pattern discovery since the definition of "*common*" depends on the dataset and the task, so theoretically there are no optimal parameters of η and λ. In the following experiments on the real-world images, the parameters are fixed to fulfill the evaluation metric of the specified task.

D. Applications to Real-World Images

To utilize the proposed framework for common object discovery and localization in real-world images, we first extract object proposals for each image using off-the-shelf method [39]. After generating a pool of object proposals, we then resort to a pre-trained two-class classifier, called DeepBox [23], to distinguish the objects from backgrounds. The proposals with lower probabilities of containing objects are removed, so that the size of proposal pool is much smaller.

Due to CNN [40] has been demonstrated to have the capability to capture high-level image representations, we thus employ the popular AlexNet [41] to extract features for region proposals. The AlexNet is finely trained on the ImageNet dataset with 1000 outputs at the last layer, we use the outputs of the second fc-layer, that is 4096-dimensional feature for each proposal.

1) Proposal Confidence: We introduce a prior for each proposal to represent the confidence of being positive. The average saliency [44] within the box, which provides useful common foreground prior and has been extensively used in foreground co-segmentation and image co-localization [5], [7], can be used as an instructor of positive proposals. And also, the scores of DeepBox [23], which is used at previous step to filter out proposals based on a binary classifier, could also be used to indicate the objectness of proposals. Finally, our proposal score $\boldsymbol{c}$ in the first term of our objective functions can be expressed as the weighted combination of the above two pieces of prior information, the weight is fixed as 0.7/0.3 in

Algorithm 1 Cohesive Sub-Graph Mining

Input : A set of unlabeled images $\{\mathcal{I}_i\}_{i=1}^N$.
(**S1**) Extract object proposals using any off-the-self methods, denoted by $\{\mathcal{P}_j\}_{j=1}^M$ and the corresponding index matrix $A \in \mathcal{R}^M \times N$;
(**S2**) Compute the confidence $\{c_j\}_{j=1}^M$ of being positive for each proposal based on prior knowledge;
(**S3**) Extract features from each proposal, and compute the similarity matrix and Laplacian matrix, denoted respectively by $W, L \in \mathcal{R}^{M\times M}$;
(**S4**) Formulate the task as *S-UOL* Model or *M-UOL* Model according to the assumption whether the positive images only contain a single common object;
(**S5**) Solve the *S/M-UOL* Model using our proposed optimization method to get the proposal labels $f \in \mathcal{R}^M$ (and image labels $g \in \mathcal{R}^N$ if necessary);
Output: The proposal labels f and image labels g if necessary.

our experiments. It is noted that the choice of proposal scores can be so flexible and it depends on the specific task and the prior knowledge.

2) Similarity Matrix: Given the proposal representations, we compute the similarity matrix $W_{i,j}$ with all the proposals. Based on the observation that the number of positive proposals accounts for only a small proportion of the total proposals and the similarity matrix should be sparse, we set the similarities between proposals from the same image to be 0 for the proposed *"S-UOL"* Model. On the other hand, for the *"M-UOL"* Model where positive images containing multiple common objects, we set the similarities to be 0 only when the IoU of two proposals is larger than a certain value (*e.g.*, $\frac{area(\mathcal{P}_i \cap \mathcal{P}_j)}{area(\mathcal{P}_i \cup \mathcal{P}_j)} > 0.5$). In addition, the similarity W_{ij} is also set to be 0 if the ith and jth proposals are not the K-reciprocal neighbors to each other, K is fixed as the number of images in our experiments. This makes sure the quality of the edge connections. We summarize the working flow of our proposed unsupervised common object discovery framework in Algorithm 1.

IV. Unsupervised Object Discover Experiments

To evaluate the performance of our proposed method on co-localization and discovery, we conduct experiments on two benchmarks: the Object Discovery dataset [14] and the PASCAL VOC 2007 dataset [45]. The proposed method is compared with some representative weakly-supervised methods [1]–[4], [46]–[49] and also fully unsupervised methods [5]–[7], [14], [25], [43], [50], [51] in terms of the detected locations of common objects. To evaluate the robustness of our method in the presence of outlier images, we also conduct experiments on the Object Discovery dataset with outlier ratios ranging from 0% to 60%.

Following [6], we implement two types of settings: 1) The *separate-class* experiment, which is conducted across images with the same class labels (for positive images), so that there is only one common object class in such a setting. It should be noted that the supervision of this setting is even weaker than traditional weakly-supervised setting where all images are labeled, however, there can exist anonymous noisy images which do not contain any common objects in our setting, and the number of noisy images is unknown either. 2) The *mixed-class* experiment, which is in a fully unsupervised way without any prior knowledge and there may be multiple common object classes contained in the given image sets. This scenario is common for image/object search, object discovery in videos, etc. In such a case, the positive images may contain different types of objects, while the negative images are unknown. This problem hasn't been well studied before. To save space, we abbreviate *separate-class* to "sep." and *mixed-class* to "mix.", and we use "sep.-c" to denote the setting where only the nodes are considered (removing the Laplacian term in the objective) during mining the subgraphs.

A. Implementation Details

We first extract 300 proposals from each image using [39], then remove the proposals predicted as backgrounds by [23], the number of remained object proposals varies from 32 to 180. Grid-search experiments over ranges of values are conducted to find the parameters in the two proposed models, *i.e.*, *S-UOL* and *M-UOL*. For the *S-UOL* Model, the correct localization (CorLoc) metric is adopted for a fair comparison with previous methods, which is defined as the percentage of images correctly localized according to the PASCAL criterion: $\frac{area(\mathcal{P}_i \cap \mathcal{P}_{gt})}{area(\mathcal{P}_i \cup \mathcal{P}_{gt})} > 0.5$, where $\mathcal{P}_i$ is the predicted box and $\mathcal{P}_{gt}$ is the ground-truth box. For these images which contain multiple objects, we follow the commonly used criterion that regarding the image as correctly co-localized if any of its objects are localized correctly.

For the *M-UOL* Model, since the last term in the objective function (*i.e.*, Eq. (2)) only controls the number of selected object proposals in total and there is no supervision or assumption to indicate the exact number of common objects in each image, and it is likely that the resulting number of proposals in each image may be more or less than the number of ground truth objects. Thus in our experiments, the *M-UOL* Model is evaluated by F1-score which is a measurement in instance-level rather than the image-level metric CorLoc adopted in the *S-UOL* Model. The F1-score is defined as:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precison} + \text{Recall}},$$

where precision is the proportion of correctly localized positives (IoU>0.5) relative to the number of identified positives, and recall is the proportion of correctly localized objects relative to the total number of ground truth objects.

1) The Object Discovery Dataset: The Object Discovery dataset [14] contains 15k images in three categories: airplane, car, and horse. As it is automatically collected by the Bing API using image queries of these three categories, the dataset contains outlier/negative images without the three common object classes. To compare with previous co-segmentation and co-localization method fairly, we use a subset of the dataset containing 100 images for each category. There are 18, 11, 7 outliers for three categories, respectively. The dataset is

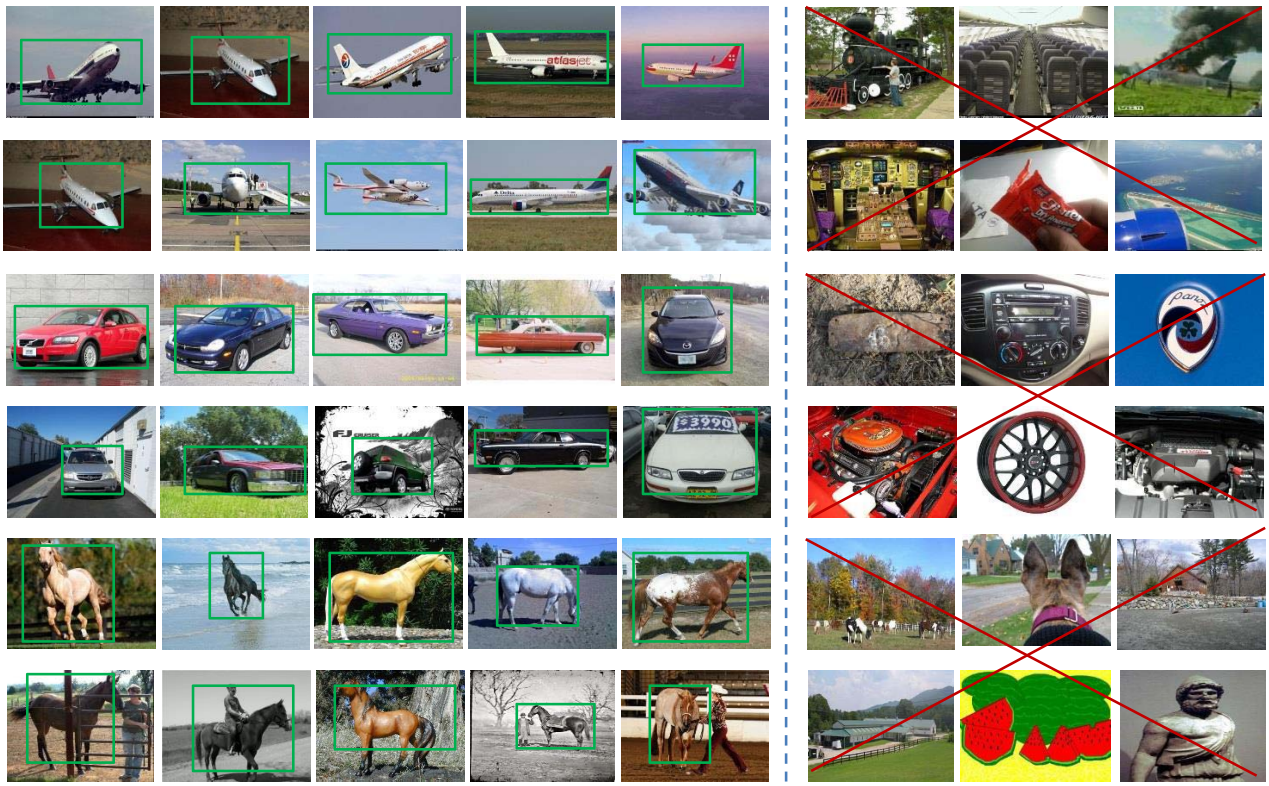
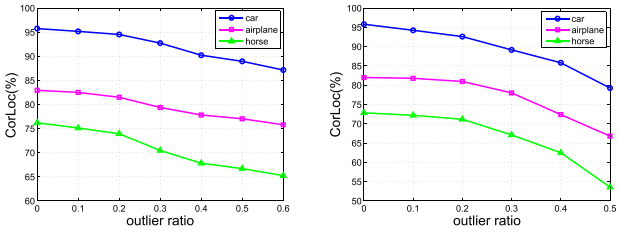


Fig. 5. Examples of successful co-localization results for the Object Discovery subset in mixed-class setting. **Left:** Results using our *S-UOL* Model; **Right:** Outlier/Negative images.

TABLE II
CORLOC(%) FOR THE OBJECT DISCOVERY SUBSET USING *S-UOL* MODEL

Method	Airplane	Car	Horse	Av.
Kim *et al.* [42]	21.95	0.00	16.13	12.69
Joulin *et al.* [25]	32.93	66.29	54.84	51.35
Joulin *et al.* [43]	57.32	64.04	52.69	58.02
Rubinstein *et al.* [14]	74.39	87.64	63.44	75.16
Tang *et al.* [5]	71.95	93.26	64.52	76.58
Cho *et al.* [6]-sep.	82.93	94.38	75.27	84.19
Cho *et al.* [6]-mix.	81.71	94.38	70.97	82.35
Ours-sep.-c	69.73	89.0	68.19	75.64
Ours-sep.	82.1	95.26	76.0	84.45
Ours-mix.	81.13	94.17	72.52	82.61



(a) (b)

Fig. 6. CorLoc(%) for the Object Discovery dataset with outliers using our *S-UOL* Model. (a) separate-class. (b) mixed-class.

originally designed for co-segmentation with pixel-level labels, we convert the ground-truth segmentations to object locations by drawing tight bounding boxes around positive pixels.

Table II shows the comparison of state-of-the-art methods and our *S-UOL* Model, which assumes that each image contains at most one common object, in the separate-class and mixed-class settings. From the table we can see that based on the separate-class setting (Ours-sep.), our method is comparable to the state-of-the-art [6]. The gap between "sep.-c" and "sep." demonstrates that both nodes and edges will contribute to the final performance. Although the results on the mixed-class setting are about 2% lower in average compared with the separated-setting, our method still outperforms the majority of previous methods. The reason for the gap between results of these two settings is that the separate-class experiments are conducted in each category with 100 images so that there is only one common object class in this case and the mixed-class experiment are based on all the three categories with 36 negative images in total.

In Figure 6, we show the performance of CorLoc over outlier ratios using the proposed *S-UOL* Model. The outlier images are randomly sampled from PASCAL VOC 2007 (except the images labeled airplane, car and horse). In the separate-class experiment where there is only one common object class among the image sets, the performance is still satisfactory even when the outlier ratio reaches to as high as 60%. In the mixed-class experiment where there are three common object classes, the performance is steady in lower outlier ratios, while drops heavily in high outlier ratios. This is because, with high outlier ratio, such as 50%, each of the common object class only accounts for about 16% which is already not so 'dominant'.

TABLE III
F1-SCORES FOR THE OBJECT DISCOVERY SUBSET USING THE *M-UOL* MODEL

Method	Airplane	Car	Horse	Av.
EdgeBox [39]	9.99	11.79	8.48	10.09
DeepBox [23]	48.59	60.9	39.2	49.56
Ours-sep.	83.07	82.03	76.81	80.64
Ours-mix.	71.14	72.44	64.07	69.22

TABLE IV
AVERAGE CORLOC(%) FOR PASCA07-6×2 USING THE *S-UOL* MODEL

Method	Average Corloc
Deselaers *et al.* [52]	50
Siva and Xiang *et al.* [53]	49
Tang *et al.* [5]	39
Cho *et al.* [6]-sep.	68
Cho *et al.* [6]-mix.	54
Ours-sep.	67
Ours-mix.	56

We compare the proposed *M-UOL* Model with EdgeBox [39] and DeepBox [23] in terms of the F1-score. To calculate the measurement, we use the top 300 object proposals for the EdgeBox and the positives identified by DeepBox from the 300 proposals are used to evaluate itself. The results are listed in Table III.

Figure 5 shows example results using the proposed *S-UOL* Model and *M-UOL* Model in the mixed-class setting. In spite of accounting for small size, the *S-UOL* is able to localize the target objects from cluttered backgrounds. It is noted that for the *M-UOL* Model, we have no constraint on the exact number of objects in each image. Thus, the output proposals are usually more or less than the ground truth objects. We also show some outlier images that can be successfully eliminated by both models.

2) PASCAL VOC 2007 Dataset: The PASCAL VOC 2007 dataset [45] is one of the most popular datasets for computer vision tasks such as classification, detection, and segmentation, and it contains 5011 images in 20 object categories. This dataset is quite challenging for the task of co-localization because: 1) most of the objects in this set of images are with considerable clutter, occlusion, and different viewpoints; 2) many images in this dataset contain multiple common objects. Following the experimental setup defined in [6], we evaluate our method on two sizes of sets: PASCA07-6×2 and PASCA07-all. The PASCA07-6×2 consists of all images from 6 classes (aeroplane, bicycle, boat, bus, horse, and motorbike) of train+val dataset from the left and right aspects each. Each of the 12 class/viewpoints combinations contains between 21∼50 images for a total of 463 images. And for the large-scale dataset PASCA07-all, we use all images from train+val datasets to evaluate our method.

a) PASCA07-6×2: In Table VI, we compare our *S-UOL* model with previous methods in terms of CorLoc. In both separate-/mixed-class settings, our method can achieve satisfactory performance. The results of the separate-class without the Laplacian term (sep.-c) is also among the top performances. Though being lower in certain categories, the average performance of our method is comparable to that of [6] in the separate-class setting (sep.) and much higher in the mixed-class setting. Fig. 7 presents some examples generated by the *S-UOL* Model. And the results of the *M-UOL* Model for this dataset are showed in Table V with the average F1-scores.

TABLE V
AVERAGE F1-SCORES FOR THE PASCAL07 USING THE *M-UOL* MODEL

Method	PASCAL07-6×2	PASCAL07-all
EdgeBox [39]	9.20	8.15
DeepBox [23]	34.61	32.21
Ours-sep.	76.91	75.01
Ours-mix.	69.42	63.69

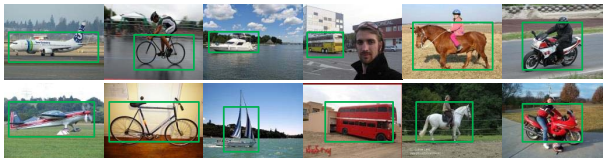


Fig. 7. Examples of co-localization for the PASCAL07-6×2 using the *S-UOL* Model under mixed-class setting.

b) PASCA07-All: Here we tackle a much more challenging and large-scale discovery task, using all images from the PASCAL07 train+val dataset. The comparison of our *S-UOL* model to the state-of-the-arts is shown in Table VII, where the top 10 rows are weakly supervised localization methods, [6] and [7] are fully unsupervised methods, and [6] also provides results in weakly-supervised setting. From Table VII we can see that our *S-UOL* model outperforms state-of-the-art unsupervised method [6] in separate-class setting, and also slightly outperforms the fully-unsupervised setting of [7] and [6]-mix. As weakly-supervised setting, the results achieved by "our-sep" are inferior to some previous weakly-supervised methods, the reasons might rely on: 1) as we have stated above, the supervision of "Our-sep" is even weaker than traditional weakly-supervised setting; 2) some of the weakly-supervised methods use additional supervision or prior knowledge [47], [55]. We show some examples of results in Fig. 8. Table V shows the comparisons of our *M-UOL* Model with EdgeBox [39] and DeepBox [23] in terms of the average F1-scores.

3) Discussion: In this section, we conduct more ablation studies on different strategies of proposal selection and on multiple kinds of features, especially hand-crafted features. Table VIII(a) shows the average CorLoc(%) on object discovery benchmarks. "All" denotes using all 300 proposals extracted by [39], and "Top-x%" denotes using top x% proposals ranked by objectness score [39]. The average number of selected proposals by DeepBox [23] is 67 for each image. From the table we can see that, using all proposals is slightly better than Deepbox. Actually, the main advantage of using DeepBox is to reduce computational complexity. As we stated in Sect. III-B, the worst case complexity of optimizing objective function is $\mathcal{O}(M^2en)$, where M is the number of proposals from all images. With DeepBox, the average number of selected proposals can be reduced to 67 per image, which is significantly efficient compared to 300 per image.

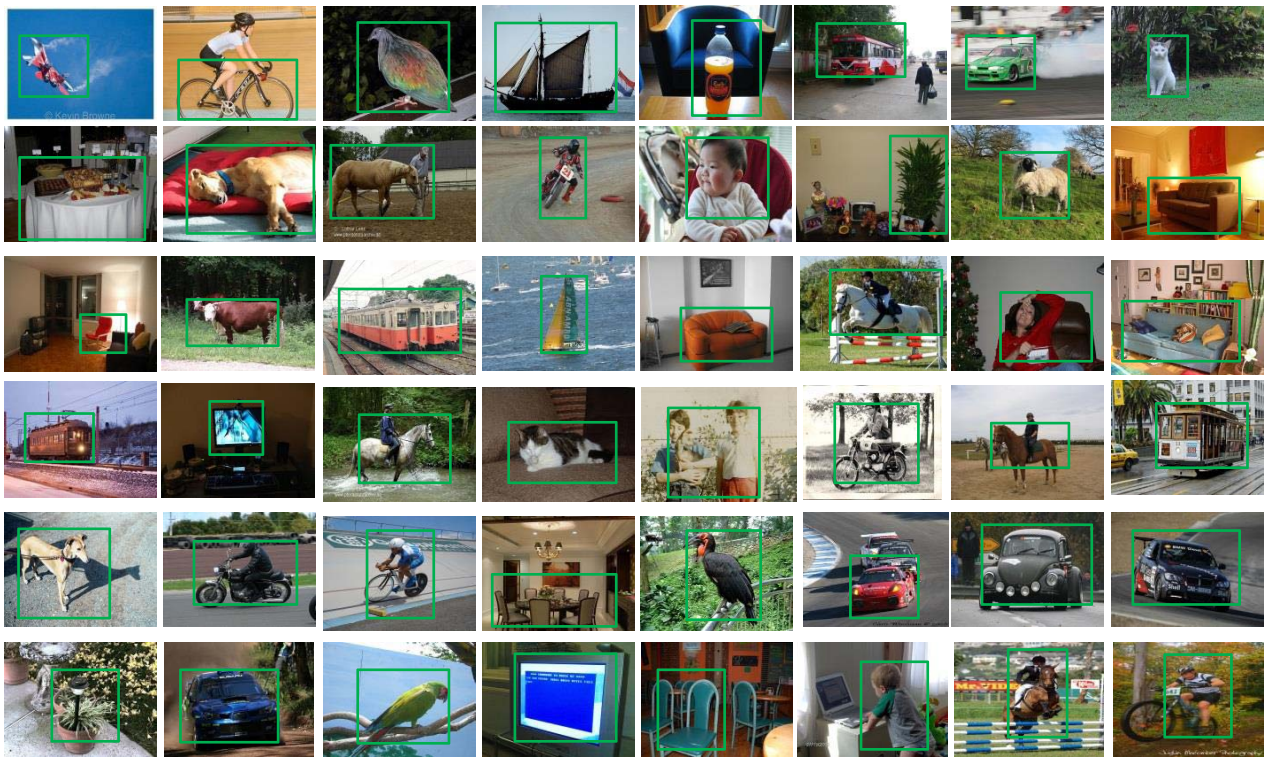


Fig. 8. Examples of co-localization for the PASCAL07-all using the *S-UOL* Model in mixed-class setting, we resize the ratio of width and height in order to show elegantly.

TABLE VI

CORLOC(%) FOR PASCAL07-6×2 USING THE *S-UOL* MODEL. THE METHODS LABELED BY '∗' ARE IN WEAKLY SUPERVISED WAY

Method	aeroplane		bicycle		boat		bus		horse		motorbike		
	L	R	L	R	L	R	L	R	L	R	L	R	Av.
∗Siva and Xiang *et al.* [53]	-	-	-	-	-	-	-	-	-	-	-	-	49
∗Pandey *et al.* [48]	-	-	-	-	-	-	-	-	-	-	-	-	61.05
Tang *et al.* [5]	41.86	51.28	25.00	24.00	11.36	11.63	38.10	56.52	43.75	52.17	51.28	64.71	39.31
∗Cho *et al.* [6]-sep.	62.79	71.79	77.08	62.00	25.00	32.56	66.67	91.30	83.33	86.96	82.96	70.59	67.68
Cho *et al.* [6]-mix.	62.79	66.67	54.17	56.00	18.18	18.60	42.86	69.57	70.83	71.74	69.23	44.12	53.73
Niu *et al.* [51]	64.28	87.17	35.41	39.13	16.66	30.00	42.86	60.86	46.80	53.33	59.45	75.75	50.97
∗Ours-sep.-c	49.23	62.46	39.0	32.07	37.64	34.58	42.57	68.42	54.83	62.46	58.0	70.17	50.94
∗Ours-sep.	67.44	76.92	52.08	40.0	45.45	48.84	57.14	95.65	79.17	84.78	74.36	85.29	67.26
Ours-mix.	62.79	71.79	41.67	34.0	24.91	39.53	47.62	81.96	64.58	71.74	69.23	64.41	56.19

To validate our method in fully-unsupervised way, we replace the CNN features with HOG and SIFT features to calculate the similarity matrix while keeping other implementation details unchanged. The results (Cor-Loc(%)) based on the *S-UOL* Model in the mixed-class setting are shown in Table VIII(b). Even with HOG or SIFT features, our method can still obtain good results.

V. IMAGE RETRIEVAL EXPERIMENTS

Our proposed framework can also be easily applied in the problem of image/instance retrieval. Different from traditional image/instance retrieval methods, which only depend on the similarity between query and reference images to rank the references, we also use the similarities among references. Specifically, the initial retrieval score which is obtained by similarities between query and references is used as node confidence and the similarities among references are used to construct the Laplacian matrix in our objective function. The idea to apply our method to improve ranking is based on the following observation. If the neighbor images of an image I match query Q, then it is likely that image I also matches Q, even with a mild similarity between I and Q. On the contrary, if the neighbor images of an image I are distinct to the query Q, then it is likely that I do not match Q either, even though the individual similarity between I and Q is high, so that noisy images can be removed. The output of our proposed algorithm is a binary vector with 1 denoting an image belongs to a specified query, then the selected images are re-ranked by the initial retrieval score. We evaluate our proposed framework on two datasets, Oxford5k building [57] and Holidays dataset [58]. The first is selected for instance/object retrieval, and the latter for scene/image retrieval. For both datasets, we report mAP as the measurement metric. To save space, we use '-c' to

TABLE VII

CORLOC(%) FOR PASCAL07-ALL USING THE *S-UOL* MODEL. THE METHODS LABELED BY '$*$' ARE IN WEAKLY SUPERVISED WAY

Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	pint	she	sofa	trai	tv	Av.
*Pandey *et al.* [48]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	30.31
*Shi *et al.* [49]	54.7	22.6	33.7	24.5	4.6	33.9	42.5	57.0	7.3	39.1	24.1	43.3	41.3	51.5	25.3	13.3	28.0	29.5	54.6	11.8	32.1
*Shi *et al.* [1]	67.3	54.4	34.3	17.8	1.3	46.6	60.7	68.9	2.5	32.4	16.2	58.9	51.5	64.6	18.2	3.1	20.9	34.7	63.4	5.9	36.2
*Cinbis *et al.* [2]	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	10.2	29.0	58.0	64.9	36.7	18.7	56.5	13.2	54.9	59.4	38.8
*Li *et al.* [3]	73.1	45.0	43.4	27.7	6.8	53.3	58.3	45.0	6.2	48.0	14.3	47.3	69.4	66.8	24.3	12.8	51.5	25.5	65.2	16.8	40.0
*Li *et al.* [4]	64.3	54.3	42.7	22.7	34.4	58.1	74.3	36.2	24.3	50.4	11.0	29.2	50.5	66.1	11.3	42.9	39.6	18.3	54.0	39.8	41.2
*Bilen *et al.* [54]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
*Kantorov *et al.* [55]	78.8	66.7	52.9	25.0	26.3	68.0	73.6	44.8	14.9	62.3	45.2	46.3	61.6	82.3	35.3	39.6	69.1	30.9	62.0	69.5	52.8
*Wang *et al.* [47]	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
*Jie *et al.* [56]	70.2	60.0	53.9	26.1	28.3	58.9	75.4	58.9	14.8	63.4	17.9	52.6	51.7	67.0	19.7	46.3	63.9	42.4	67.0	65.1	50.2
Joulin *et al.* [7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.6
*Cho *et al.* [6]-sep.	50.3	42.8	30.0	18.5	4.0	62.3	64.5	42.5	8.6	49.0	12.2	44.0	64.1	57.2	15.3	9.4	30.9	34.0	61.6	31.5	36.6
Cho *et al.* [6]-mix.	40.4	32.8	28.8	22.7	2.8	48.4	58.7	41.0	9.8	32.0	10.2	41.9	51.9	43.3	13.0	10.6	32.4	30.2	52.7	21.8	31.3
*Ours-sep.-c	31.8	25.2	33.5	22.0	16.2	38.3	40.3	38.6	13.6	24.6	10.9	26.0	31.1	38.3	28.5	19.7	50.0	20.7	40.1	11.4	29.4
*Ours-sep.	55.2	35.7	45.9	30.0	20.2	49.2	50.2	53.2	18.1	44.1	13.6	49.1	39.7	43.3	35.4	22.3	67.7	29.7	50.5	25.3	38.9
Ours-mix.	42.2	31.7	36.7	20.0	16.2	36.2	40.2	51.9	15.1	40.0	12.6	44.6	31.5	38.9	31.3	19.7	54.8	28.7	43.9	8.9	32.2

TABLE VIII

ABLATION STUDIES IN TERMS OF DIFFERENT PRIOR KNOWLEDGE. (a) COMPARISON OF DIFFERENT WAYS OF SELECTING PROPOSALS. (b) COMPARISON OF CNN FEATURE WITH HAND-CRAFTED FEATURES

(a)

Selected pps.	Object Discovery	VOC07-6×2	VOC07-all
All	82.97	56.48	32.19
Top-50%	81.29	55.74	31.08
Top-30%	80.95	54.58	30.42
DeepBox [23]	82.61	56.19	32.21

(b)

Feature	Object Discovery	VOC07-6×2	VOC07-all
CNN	82.61	56.19	32.21
SIFT	81.23	55.92	31.47
HOG	82.05	54.87	30.73

represent the case that only considering the similarity between query and reference images.

A. Holidays

This dataset contains 1491 images of which 500 are queries. It contains images of different scenes, items and monuments. For each image, the maximum activation of convolutions (MAC) feature and regional maximum activation of convolutions (R-MAC) feature [29] are extracted. The final dimensions of the two kinds of features are both 512. Let $\boldsymbol{q} \in \mathbb{R}^d$ denote the feature vector of a query, and $I \in \mathbb{R}^{d \times N}$ be the feature matrix of the dataset, where N is the number of reference images. Then the node confidence $\boldsymbol{c} \in \mathbb{R}^N$, which represents the similarity between the query and the reference images, can be computed as:

$$\boldsymbol{c} = \boldsymbol{q}^\top I, \tag{6}$$

and the similarity matrix $W \in \mathbb{R}^{N \times N}$ between the reference images are computed as:

$$W = I^\top I. \tag{7}$$

Table IX lists our performance and some recent studies on this dataset. We can find that with similar dimension of feature vector, our proposed *M-UOL* model can improve $> 2\%$ compared with the results using only the similarity between query and reference images. Note that any state-of-the-art methods which learn the similarity between the query and reference images can be used as the node confidences in our framework, then the retrieval accuracy could be further improved by our proposed *M-UOL* model.

TABLE IX

THE RESULTS OF IMAGE RETRIEVAL ON HOLIDAYS DATASET

Method	Dimension	mAP
Sharif *et al.* [26]	500	82.2
Perronnin *et al.* [59]	512	82.7
	4k	84.7
Arandjelovic *et al.* [27]	256	79.9
Radenović *et al.* [60]	512	82.5
Rezende *et al.* [61]	512	82.3
R-MAC: Ours-c	512	84.5
R-MAC: Ours	512	87.1
MAC: Ours-c	512	83.2
MAC: Ours	512	85.6

B. Oxford5k

This dataset is composed of 5063 reference images and 55 queries. The object bounding boxes in the queries have been given, so it means the query objects are known. The challenging of this dataset is that the scale of the objects in the reference images varies a lot, and all the buildings are very similar. Since this dataset targets at instance retrieval, the objects contained in images should be focused, thus we extract object proposals from each image. Then for each object proposal, the maximum activation of convolutions (MAC) feature and regional maximum activation of convolutions (R-MAC) feature [29] are extracted. The final dimensions of the two kinds of features are both 512. We denote the feature matrix of all the object proposals as $P \in \mathbb{R}^{d \times M}$. In such a case, the node confidence is $\boldsymbol{c'} \in \mathbb{R}^M$, which represents the similarity between the query object and the object proposals:

$$\boldsymbol{c'} = q^\top P, \tag{8}$$

and the similarity matrix $W' \in \mathbb{R}^{M \times M}$ between the object proposals is computed as:

$$W' = P^\top P. \tag{9}$$

Then, the node confidence of each reference image, *i.e.*, the distance between query object to the reference image is determined by the nearest object proposals in the image.

TABLE X
THE RESULTS OF INSTANCE RETRIEVAL ON OXFORD5K

Method	Dimension	mAP
Babenko *et al.* [62]	256	65.7
Toliase*t al.* [29]	512	66.9
	-	77.3
Arandjelovic *et al.* [27]	256	63.5
Radenović *et al.* [60]	512	77.0/80.1
Rezende *et al.* [61]	512	64.1
MAC *et al.* [28]	512	73.9
R-MAC: Ours-c	512	77.6
R-MAC: Ours	512	80.2
MAC: Ours-c	512	73.9
MAC: Ours	512	76.3

The similarity W_{ij} between any two images i and j is determined by averaging the similarities between all the objects proposals in them. From Table X, we can be found that with similar feature dimension, our *M-UOL* model using R-MAC can achieve state-of-the-art performance. Compared with the retrieval results of using only the similarity between the query and reference images ('Ours-c'), the *M-UOL* model can increase about 3% with both MAC and R-MAC, which demonstrates the effectiveness of the proposed *M-UOL* model.

VI. CONCLUSION

In this paper, we propose a framework for common object discovery and localization in wild images. Like most previous methods which are based on the assumption that there is only one object contained in each positive image, we introduce the *S-UOL* Model, which is also demonstrated to be robust even with a significant proportion of outlier images. Then, the *M-UOL* Model is further introduced for a more general scenario where there could be multiple common objects contained in each image. The proposed *M-UOL* Model can also be easily extended to the task of image/instance retrieval. Inspired by min-cut/max-flow algorithms, we then present a constrained sub-graph mining algorithm to optimize the two models. To evaluate the proposed method, we conduct extensive experiments under multiple settings and compare with many representative studies. Empirical results demonstrate that the proposed method performs well despite in a fully unsupervised way, and even when variations in scale, view-point, appearance or partial occlusions frequently occur in images.

ACKNOWLEDGMENT

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, and the Infocomm Media Development Authority, Singapore.

REFERENCES

[1] Z. Shi, T. M. Hospedales, and T. Xiang, "Bayesian joint modelling for object localisation in weakly labelled images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1959–1972, Oct. 2015.

[2] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2409–2416.

[3] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Image co-localization by mimicking a good detector's confidence score distribution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 19–34.

[4] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3512–3520.

[5] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1464–1471.

[6] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1201–1210.

[7] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with frank-Wolfe algorithm," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zürich, Switzerland: Springer, 2014, pp. 253–268.

[8] J. Yuan, G. Zhao, Y. Fu, Z. Li, A. K. Katsaggelos, and Y. Wu, "Discovering thematic objects in image collections and videos," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2207–2219, Apr. 2012.

[9] J. Yuan and Y. Wu, "Spatial random partition for common visual pattern discovery," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[10] M. Wolff, R. T. Collins, and Y. Liu, "Regularity-driven building facade matching between aerial and street views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1591–1600.

[11] X. Wang, Z. Zhu, C. Yao, and X. Bai, "Relaxed multiple-instance SVM with application to object discovery," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1224–1232.

[12] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 275–293, Dec. 2012.

[13] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 594–602.

[14] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in Internet images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1939–1946.

[15] L. Wang, G. Hua, J. Xue, Z. Gao, and N. Zheng, "Joint segmentation and recognition of categorized objects from noisy Web image collection," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4070–4086, Sep. 2014.

[16] L. Wang, G. Hua, R. Sukthankar, J. Xue, Z. Niu, and N. Zheng, "Video object discovery and co-segmentation with extremely weak supervision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2074–2088, Oct. 2017.

[17] A. Faktor and M. Irani, "Co-segmentation by composition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1297–1304.

[18] G. Zhao, J. Yuan, and G. Hua, "Topical video object discovery from key frames by modeling word co-occurrence prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1602–1609.

[19] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2178–2190, Dec. 2010.

[20] C. Lu, R. Liao, and J. Jia, "Personal object discovery in first-person videos," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5789–5799, Dec. 2015.

[21] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Firenze, Italy: Springer, 2012, pp. 73–86.

[22] Y. J. Lee, A. A. Efros, and M. Hebert, "Style-aware mid-level representation for discovering visual connections in space and time," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1857–1864.

[23] W. Kuo, B. Hariharan, and J. Malik, "DeepBox: Learning objectness with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ECCV)*, Dec. 2015, pp. 2479–2487.

[24] A. Faktor and M. Irani, "'Clustering by composition'—Unsupervised discovery of image categories," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Firenze, Italy: Springer, 2012, pp. 474–487.

[25] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1943–1950.

[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.
[27] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5297–5307.
[28] T. Yu, Y. Wu, S. Bhattacharjee, and J. Yuan, "Efficient object instance search using fuzzy objects matching," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4320–4326.
[29] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–12.
[30] A. Kankaria, "Query expansion techniques," *J. Comput.*, vol. 1, no. 2, 2012.
[31] N. Tatti and A. Gionis, "Density-friendly graph decomposition," in *Proc. Int. Conf. World Wide Web*, 2015, pp. 1089–1099.
[32] A. Gionis, F. P. Junqueira, V. Leroy, M. Serafini, and I. Weber, "Piggybacking on social networks," *Proc. Very Large Data Base Endowment*, vol. 6, no. 6, pp. 409–420, 2013.
[33] G. Zhao and J. Yuan, "Discovering thematic patterns in videos via cohesive sub-graph mining," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 1260–1265.
[34] G. Zhao, J. Yuan, J. Xu, and Y. Wu, "Discovering the thematic object in commercial videos," *IEEE Multimedia Mag.*, vol. 18, no. 3, pp. 56–65, Jul./Sep. 2011.
[35] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.
[36] A. V. Goldberg and R. E. Tarjan, "A new approach to the maximum-flow problem," *J. ACM*, vol. 35, no. 4, pp. 921–940, 1988.
[37] C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt, "Efficient network-guided multi-locus association mapping with graph cuts," *Bioinformatics*, vol. 29, no. 13, pp. i171–i179, 2013.
[38] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[39] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zürich, Switzerland: Springer, 2014, pp. 391–405.
[40] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229
[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[42] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 169–176.
[43] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 542–549.
[44] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 FPS," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1404–1412.
[45] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2007). *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
[46] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Firenze, Italy: Springer, 2012, pp. 594–608.
[47] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zürich, Switzerland: Springer, 2014, pp. 431–445.
[48] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1307–1314.
[49] Z. Shi, P. Siva, and T. Xiang, "Transfer learning by ranking for weakly supervised object annotation," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 78.1–78.11.
[50] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1605–1614.
[51] Z. Niu, G. Hua, L. Wang, and X. Gao, "Knowledge-based topic model for unsupervised object discovery and localization," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 50–63, Jan. 2018.
[52] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 452–466.
[53] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 343–350.
[54] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2846–2854.
[55] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "ContextLocNet: Context-aware deep network models for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 350–365.
[56] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4294–4302.
[57] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[58] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.
[59] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3743–3752.
[60] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 3–20.
[61] R. S. Rezende, J. Zepeda, J. Ponce, F. Bach, and P. Pérez, "Kernel square-loss exemplar machines for image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.
[62] A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1269–1277.

Zhenzhen Wang received the B.E. and M.S. degrees from the School of Information and Control, Nanjing University of Information Science and Technology, China, in 2012 and 2015, respectively. She is currently pursuing the Ph.D. degree with the School of Electrical and Electronics Engineering, Nanyang Technological University. Her current research interests include common pattern discovery, image retrieval, and network compression.

Junsong Yuan (M'08–SM'14) received the degree from the Special Class for the Gifted Young, Huazhong University of Science and Technology, China, in 2002, the M.Eng. degree from the National University of Singapore in 2005, and the Ph.D. degree from Northwestern University in 2009.

He was an Associate Professor with the School of Electrical and Electronics Engineering, Nanyang Technological University (NTU), Singapore. He is currently an Associate Professor with the Computer Science and Engineering Department, University at Buffalo, The State University of New York, USA. He received the Best Paper Award from the IEEE TRANSACTIONS ON MULTIMEDIA, the Doctoral Spotlight Award from the IEEE Conference on Computer Vision and Pattern Recognition in 2009, the Nanyang Assistant Professorship from NTU, and the Outstanding EECS Ph.D. Thesis Award from Northwestern University.

Dr. Yuan is a fellow of the International Association of Pattern Recognition. He is a Program Co-Chair of ICME'18 and VCIP'15, and the Area Chair of ACM MM'18, ACCV'18'14, ICPR'18'16, CVPR'17, and ICIP'18'17. He is currently a Senior Area Editor of the *Journal of Visual Communication and Image Representation* and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He served as a Guest Editor for the *International Journal of Computer Vision*.