Multiple Semantic Matching on Augmented *N*-Partite Graph for Object Co-Segmentation

Chuan Wang, Hua Zhang, Liang Yang, Xiaochun Cao, *Senior Member, IEEE*, and Hongkai Xiong, *Senior Member, IEEE*

***Abstract*— Recent methods for object co-segmentation focus on discovering single co-occurring relation of candidate regions representing the foreground of multiple images. However, region extraction based only on low and middle level information often occupies a large area of background without the help of semantic context. In addition, seeking single matching solution very likely leads to discover local parts of common objects. To cope with these deficiencies, we present a new object co-segmentation framework, which takes advantages of semantic information and globally explores multiple co-occurring matching cliques based on an *N*-partite graph structure. To this end, we first propose to incorporate candidate generation with semantic context. Based on the regions extracted from semantic segmentation of each image, we design a merging mechanism to hierarchically generate candidates with high semantic responses. Second, all candidates are taken into consideration to globally formulate multiple maximum weighted matching cliques, which complement the discovery of part of the common objects induced by a single clique. To facilitate the discovery of multiple matching cliques, an *N*-partite graph, which inherently excludes intra-links between candidates from the same image, is constructed to separate multiple cliques without additional constraints. Further, we augment the graph with an additional virtual node in each part to handle irrelevant matches when the similarity between the two candidates is too small. Finally, with the explored multiple cliques, we statistically compute pixel-wise co-occurrence map for each image. Experimental results on two benchmark data sets, i.e., iCoseg and MSRC data sets achieve desirable performance and demonstrate the effectiveness of our proposed framework.**

***Index Terms*— Object co-segmentation, semantic candidate, multiple matches, *N*-partite graph.**

I. Introduction

OBJECT co-segmentation [1]–[3] aims to segment common objects with the same category and appearance among a set of images. Compared with object segmentation from a single image, object co-segmentation can utilize both local context information from a certain image and global common information among multiple images. It is beneficial to a variety of applications in computer vision, e.g., image classification [4] and object/instance detection [5], [6].

Most of existing methods [7]–[9] cast the co-segmentation to a graph-based candidate matching problem. They model the connectivity of candidates based on a complete graph (as in Fig. 1 (a)) with pairwise feature similarities such as color histograms [7] and gradient histograms [10]. To measure the commonality among candidates from different images, they seek to find an optimal matching solution [9], [11] with the maximal summation of similarities of the selected candidates. Segmentation methods [3], [12] are employed to obtain final binary segmentations. In the pipeline, both candidate generation and matching have great importance on common object discovery.

Firstly, due to diverse deformation of foreground and high complexity of background, the extracted candidates may cover background. Context-based candidate generation methods [13]

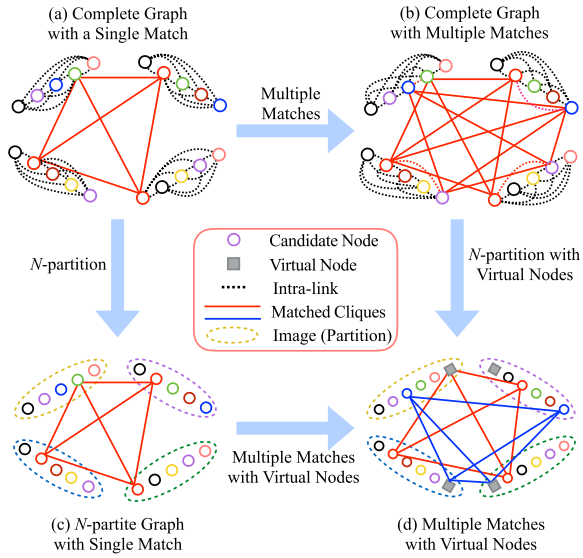

Fig. 1. The differences between our proposed multiple cliques matching on augmented *N*-partite graph (d) and the existing graph-based method (a). On one hand, transforming single clique matching to multiple (from (a) to (b)) can alleviate the problem of discovery of the part of common objects. On the other hand, employing the *N*-partite graph can effectively satisfy the requirement, that candidates in each clique come from different images, without adding exclusive constraints (from (a) to (c)). Furthermore, with introduction of a virtual node to each image (d) the multiple matches can be globally explored with the exclusion of irrelevant matches.

Manuscript received October 22, 2016; revised April 24, 2017 and July 16, 2017; accepted August 25, 2017. Date of publication September 8, 2017; date of current version September 21, 2017. This work was supported by in part by the National Key Research and Development Plan under Grant 2016YFB0800603, in part by the National Natural Science Foundation of China under Grant 61422213 and Grant U1636214, and in part by the Beijing Natural Science Foundation under Grant 4172068, and in part by the Key Program of the Chinese Academy of Sciences under Grant QYZDB-SSW-JSC003. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianfei Cai. *(Corresponding author: Xiaochun Cao.)*

C. Wang, H. Zhang, and X. Cao are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wangchuan@iie.ac.cn; zhanghua@iie.ac.cn; caoxiaochun@iie.ac.cn).

L. Yang is with the School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China (e-mail: yangliang@vip.qq.com).

H. Xiong is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xionghongkai@sjtu.edu.cn).

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.2017.2750410

may not be able to discriminate the background and would mislead the explored common objects. Moreover, the low level representation of candidate parts is sensitive to the deformation of objects.

Secondly, when the similarity between object parts is larger than that between the entire objects, single matching clique that merely selects one candidate from each image may result in the incomplete coverage of objects. Thus the matching clique may run away from common objects when there exists large variances of pose or viewpoint among objects. Although increasing the number of selected candidates from one image [9] would help as shown in Fig. 1 (b), it still has risk of introducing irrelevant information.

To address the aforementioned problems in candidate generation and matching, we propose a multiple semantic candidate matching framework for object co-segmentation. We propose to extract candidates based on semantic segmentation from Fully Convolutional Network (FCN). Since semantic information not only has capability to seek out object-like pixels, but also it excludes backgrounds like sky and grass in images. Besides, to handle the deficiency induced by lacking global view in FCN, we accumulatively produce new candidates via hierarchically selecting and merging two initial candidates, which are spatially close and semantically similar. The new candidate with high semantic response is selected to represent foreground. Subsequently, to deal with the incomplete discovery of common objects induced by single maximum weighted matching clique, we design a novel co-occurrence exploration algorithm to globally discover multiple co-occurrences, e.g., discovering the red and blue links in Fig. 1 (d) together. Specifically, we explore multiple maximum weighted matching cliques to represent co-occurrences of the common objects with consideration of all candidates. To separate the multiple cliques from each other without additional constraints, we adopt an N-partite graph whose nodes in the same part are not connected (Fig. 1 (d)). Further, to cope with irrelevant matches caused by occlusion or variation, we augment the N-partite graph by adding a virtual node to each part. The connection between a candidate node and a virtual node indicates that there does not exist similar candidates in the image.

The main contributions of this paper are summarized as follows. 1) We integrate semantic information and object-level information into candidate generation, and obtain candidates with semantic contents and global view of the objects. 2) We explore multiple maximum weighted matching cliques for object co-segmentation to fully explore co-occurrence of the common objects and eliminate incomplete discovery of the objects induced by single clique. 3) We introduce a new graph structure, i.e., N-partite graph, for co-segmentation to meet the exclusivity among candidates selected from the same image. Thus we can release the constraints on controlling number of selections from the same image. 4) We augment the N-partite graph by adding a virtual node to each part to make the algorithm robust to the nonexistence of similar candidates in some images.

The rest is organized as follows. We illustrate the framework in Section III-A. Section III-B describes the generation of semantic candidates. In Section III-C, we present the exploration of multiple matching cliques with augmented N-partite complete graph. We show the effectiveness of our co-segmentation approach on four widely used co-segmentation datasets in Section IV and conclude the work in Section V.

II. Related Work

Object co-segmentation [1], focuses on discovering and segmenting the common objects with similar appearance. Joulin *et al.* [14] propose a weakly-supervised framework based on discriminative clustering for co-segmentation with the assumption that foreground must be an object. Then co-segmentation between two images is extended to multiple images which contain foreground with large variances and cluttered background [2], [8], [15]–[18]. Without surprise, co-segmentation among multiple images is full of complexity and difficulty, and the greatest challenge in co-segmentation is how to effectively estimate global relationships of the co-occurring objects with diverse appearances and deformation. The attention of recent works on co-segmentation among multiple images can be roughly classified into two aspects, i.e., foreground candidate refinement [9], [15], [19] and co-occurrence formulation [8], [16], [20], [21].

Refining foreground candidates focuses on improving the discriminative ability of separating foreground from background. The direct inspiration is to enforce the confidence of foreground by incorporating additional information like saliency and so on. Rubinstein *et al.* [15] introduce saliency information for each image and treat the salient regions as candidate foregrounds. Meng *et al.* [19] propose to integrate multi-modal information, i.e., superpixel segmentations [22], object detection results [23], [24] and saliency maps [25], together to improve the accuracy of located candidates. Except saliency information, Fu *et al.* [9] employ depth information [26] into the identification of the candidates. However, due to the variation of scales, illumination and depth of objects, the estimated saliency and depth information can be easily misled and heavily affect the performance of co-segmentation. In addition to enforcing the confidence of candidates, there also exist many works focusing on reducing the probability of being candidates. Compared with the difficulty and uncertainty of estimating foreground, estimating information of background is easier to achieve. Zhang *et al.* [27] propose to explore negative prior knowledge from the most similar images of different categories. Quan *et al.* [28] propose to explore context priors of background. Both schemes [27], [28] rely on superpixels, which lack global semantic view of objects and would fail when the common object in one image occupies closely to the edge of the image. We propose to employ semantic information on candidate generation. The semantic information can exclude background from candidates by removing background labels like grass and bench. Besides, it can also integrate pixels with similar appearance and form candidates with the same object category.

The other aspect of co-segmentation aims to explore better formulation of co-occurring relationships. Rubio *et al.* [16]

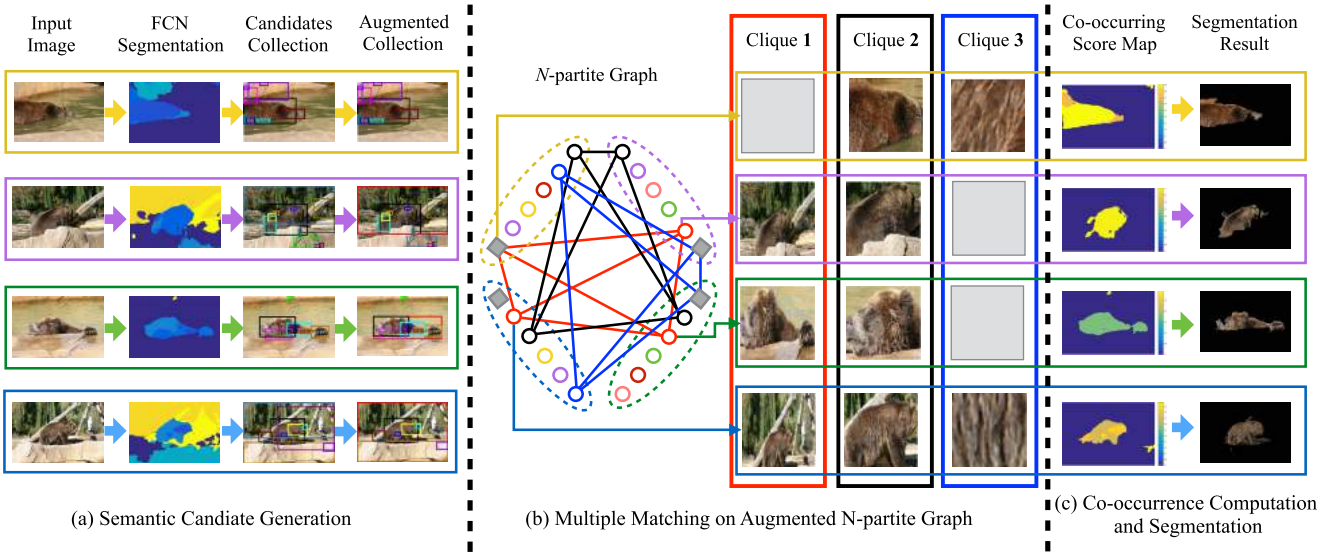

Fig. 2. Overview of our method. Given a set of images with their segmentation results from FCN, firstly we propose to extract candidates which may hold more object-level information (as shown in 4^{th} column in part (a)). We hierarchically merge two regions, which are spatially close in the image and semantically similar, and preserve new candidates with high "objectness" response. Secondly, based on the generated candidates, we build an N-partite graph each part and node of which represent one image and one candidate, respectively. We additionally introduce virtual nodes in the graph to prevent irrelevant selections, shown as gray block in (b). Then we explore multiple maximum weighted matching cliques to represent co-occurrence relations among candidates. Thirdly, with the discovered cliques we compute pixel-level co-occurring maps via counting the existence of pixels. Then the final segmentation results are obtained with Grab-cut method.

model the appearance distribution both for foreground and background of each image, and provide combination of appearance distributions for each image. However, due to huge dependence on candidate initialization [29], the co-segmentation results may be easily misguided to a local minimum, especially when large variations exist among the common objects. Faktor and Irani [8] define a new concept of "good" co-segmentation that common objects should be composed easily by candidates from inter-images but difficultly by the rest ones from the same image. By calculating the matching scores from collected pool of segments, which consists of hierarchical segmentation results for all images [13], each candidate can be evaluated by similar ones from other images in the set. Zhou *et al.* [30] propose a coarse-to-fine clustering method based on a combination of global feature and local feature to cluster near-duplicate images. Chen *et al.* [31] handle matching under the problems of textureless regions, overlap and partial loss via novel range computation and confidence estimation method. But when it comes to images with distracting background, which contain instances re-occurred in other images, the method degenerates. Faktor and Irani [8] and Rubio *et al.* [16] build the co-segmentation based on results from non-semantic priors. The co-segmentation results can be influenced if the priors provide misleading information like similar patterns from background. By incorporating semantic regions, we provide high-level estimation of foreground and background. We also propose a multiple matching mechanism to explore multiple co-occurring object-object, part-object and part-part, which take all candidates into consideration and provide abundant estimations of co-occurring relation.

III. Proposed Framework

A. Overview and Notations

To jointly label foreground pixels in a set of images containing similar objects, we propose a co-segmentation framework by exploring multiple semantic candidate matching cliques. This framework consists of three main components, i.e., semantic candidate generation, candidate matching with global consistency and segmentation according to the matching results. The flowchart is shown in Fig. 2.

Given images $\mathbf{I} = \{\mathbf{I}^1, \mathbf{I}^2, \ldots, \mathbf{I}^N\}$, our goal is to segment the common objects, e.g., bears existed in the images in Fig. 2 (a), without prior knowledges like object category or the number of common objects. Firstly, we obtain semantic segmentation map $\mathbf{L}^i$ (2^{nd} column of Fig. 2 (a)) from FCN for each image $\mathbf{I}^i$. According to the semantic map, we extract bounding boxes of non-adjacent semantic regions and build initial candidate collection $\mathbf{Q}^i = \{q_1^i, q_2^i, \ldots, q_{|\mathbf{Q}^i|}^i\}$ (3^{rd} column of Fig. 2 (a)). Secondly, to aggregate object-level information held by candidates, we hierarchically merge two candidates, which are spatially close and semantically similar, to formulate a new candidate with larger area and richer object information. The new candidate is added into the collection $\mathbf{Q}^i$ if it has higher semantic response compared with those corresponding to the two child nodes. The process is described in detail in Section III-B.

With the candidate collections, $\mathbf{Q} = \{\mathbf{Q}^1, \mathbf{Q}^2, \ldots, \mathbf{Q}^N\}$, for all images, we explore the co-occurrence among candidates via discovering multiple matching cliques, each of which is a fully connected subgraph with N nodes, in N-partite complete graph. We illustrate this process in Fig. 2 (b). Firstly,

TABLE I
NOTATIONS

$\mathbf{I}^i$	The i^{th} image in the set
$\mathbf{L}^i$	Semantic segmentation map of the i^{th} image
$\mathbf{Q}^i$	The candidate set of the i^{th} image
τ^i	The merging tree for image i
q_m^i	The m^{th} candidate of the i^{th} image and $m \in \{1, 2, \ldots, \lvert\mathbf{Q}^i\rvert\}$
$\mathbf{s}_m^i$	Semantic feature vector of candidate q_m^i
s_m^i	Maximal semantic response of candidate q_m^i
$\mathbf{V}^i$	The i^{th} part in the N-partite graph
v_m^i	The m^{th} node of the i^{th} part in the N-partite graph
v_v^i	The virtual node of the i^{th} part in the N-partite graph
e_{jn}^{im}	The edge between nodes v_m^i and v_j^n in the N-partite graph
$\mathbf{C}_c$	The c^{th} matching clique
c_c^i	The i^{th} candidate in the c^{th} matching clique
N	The number of images
K	The number of cliques

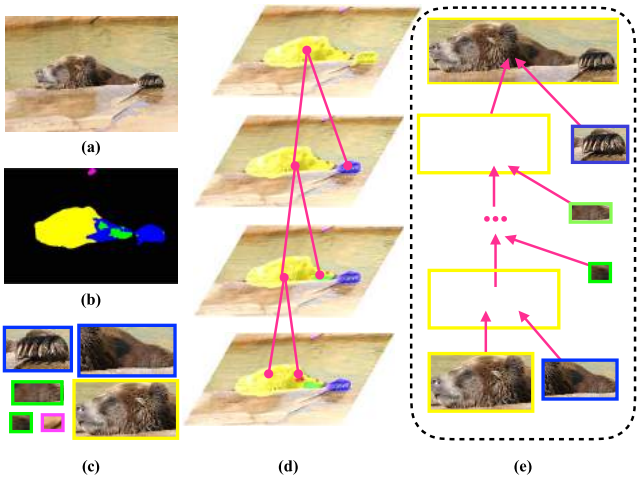

Fig. 3. Illustration of semantic candidate generation. Given an image (a) with its FCN segmentation (b), we extract segmented regions as initial candidates as shown in (c). The color of a rectangle in (c) corresponds to that of a candidate in (b). We construct a binary merging tree to hierarchically generate new candidates based on spatial distances in the image and feature similarity of two candidates, as shown in (d). Neighboring candidates with the most similar feature distributions are primarily selected. Then generated candidates with higher semantic response are added into candidate collection.

we construct the N-partite graph, each part of which represents a candidate collection $\mathbf{Q}^i$ of image $\mathbf{I}^i$. Nodes in each part represent candidates $q_1^i, q_2^i, \ldots, q_{|\mathbf{Q}^i|}^i$ in the collection. Here, we assign an additional virtual node to each part to handle irrelevant match within a clique, which is caused by absence of common components.

Secondly, we formulate the proposed multiple maximum cliques discovery problem as a global binary assignment problem of candidates. Each candidate must be selected once and only once among multiple cliques. We form the clique collection $\mathbf{C}$, i.e., $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_K\}$. Each clique $\mathbf{C}_k$ contains N components (nodes) $\{c_k^1, c_k^2, \ldots, c_k^N\}$ that come from N different images (parts). In a clique, if one part does not contain candidates with consistent appearance to others, the corresponding component will be replaced by the virtual node of that part. The details of constructing N-partite graph and solving multiple maximum cliques are described in Section III-C.1 and Section III-C.2, respectively.

Finally, with the assistance of explored multiple maximum cliques, we obtain the final segmentation results as shown in Fig. 2 (c). We calculate a pixel-wise weight map of co-occurrence for each image, which is presented in the first column of Fig. 2 (c). We weight each candidate with global co-occurrence, i.e., the number of candidates existed in one clique, and spatial co-occurrence, i.e., co-occurring frequency of neighboring candidates. Then we count the frequency of each pixel by all candidates containing it and obtain the co-occurring score by averaging the frequency. The final binary segmentation of each image $\mathbf{I}^i$ is obtained by employing Grab-cut segmentation method. This process is provided in Section III-D.

We summarize all notations in Table I which also includes others appeared in the following sections.

B. Object Candidate Generation

In this section, we introduce our semantic candidate generation process. Semantic information has been employed in various applications. For example, Li *et al.* [32] employ semantically independent patches as affine transformation priors for key point extraction and obtain superior key point matching performance. The proposed semantic candidate generation process consists of two components, i.e. semantic candidate extraction and object-level candidate generation. Compared with traditional candidate generation algorithms, we propose to generate object candidates based on semantic segmentations from FCN. We fine-tune the model with refined subset of VOC Context dataset, which contains 29 categories and 3480 images, with exclusion of background labels like grass and bench. We find that the FCN outputs for co-segmentation datasets can mostly get the contrasts from objects from unknown categories. Therefore, we neglect the category differences between datasets of FCN and co-segmentation. The fine-tuned FCN model is employed on all datasets.

However, directly using the output of FCN induces two problems. One is the misguided segmentations on object parts. Since FCN is a pixel-level classification model and has no consideration on global view of objects, pixels with similar appearance will be assigned with the same object category irrespective to their relative relation in the context. For example, when pixels of a dog's tail have similar appearance with a cat, these pixels can be incorrectly assigned as cat category [33]. Another problem is that FCN may segment unseen objects into several objects. It would recognize unseen objects as composition of multiple known categories and splits them into several individual semantic regions. To handle this problem, we propose to hierarchically merge semantic regions to form object candidates containing more object regions.

For each image $\mathbf{I}^i$, according to its segmentation map L_i from FCN shown in Fig. 3 (b), we first cluster neighboring pixels with the same category label as individual regions. We remove regions with small size. We construct the initial candidate set $\mathbf{Q}^i = \{q_m^i\}$ by extracting the bounding boxes q_m^i for each region, as shown in Fig. 3 (c).

To improve the representative ability of the collection $\mathbf{Q}^i$, we propose to hierarchically generate candidates with aggregating object-level information. We seek to improve the probability that regions from the same object may be presented in the same candidate. By considering that close regions with high semantic similarity should belong to the same object with high probability, we focus on merging regions with close spatial distances and similar semantic features. For example, the whole bear is contained by the uppermost candidate, and separated in the candidates below the uppermost one, as demonstrated in Fig. 3 (e).

Based on the initial candidate set, we build a binary merging tree for each image to hierarchically merge two candidates with neighboring spatial distance and highest semantic feature similarity, as shown in Fig. 3 (d). We employ the "objectness" score to estimate object-level information contained by candidates. "Objectness" score of a candidate is assigned with the maximal response from its semantic feature. The semantic feature vector $\mathbf{s}_m^i$ of candidate q_m^i is extracted via Region-based Convolutional Neural Network (R-CNN) [34]. We employ the ILSVRC13 model in R-CNN. For each candidate, we obtain a semantic feature vector with 200 dimensions, which is the number of categories in ILSVRC13 dataset.

Specifically, given the collection $\mathbf{Q}^i = \{q_m^i\}$ with their semantic features, the merging mechanism is designed based on pair-wise connectivity and semantic similarity. In the merging tree, all candidates $\{q_j^m\}$ are treated as leaf nodes. We sequentially merge two neighboring candidates with closest semantic similarity into a new candidate. The generated new candidate is added into the collection $\mathbf{Q}^i$ if its "objectness" score is not smaller than those of its child nodes. Here we ignore the category label corresponding to the maximal response from semantic feature vector because of the existence of unseen objects. The new turn of merging is executed until there exists no neighboring candidates. We summarize the object candidate generation in Algorithm 1.

Algorithm 1 Object Candidate Augmentation

Input: FCN segmentation $\mathbf{L}^i$ and threshold γ.
Output: All candidates in $\mathbf{Q}^i$ with semantic vectors $\{\mathbf{s}_m^i\}$ where $m \in \{1, \ldots, |\mathbf{Q}^i|\}$.
Procedure:

```
Extract individual regions from L^i and build initial candidate set Q^i;
Extract semantic vector s^i_m for candidate q^i_m where m ∈ {1,...,|Q^i|};
Initialize Q^i_temp = Q^i;
while adjacent candidates exist do
    Select two adjacent candidates q^i_m and q^i_m' from Q^i_temp with highest feature similarity;
    Sim(q^i_m, q^i_m') = 1 − sigmoid( 1/2 (KL(s^i_m || s^i_m') + KL(s^i_m' || s^i_m)) );
    Merge two candidates into a new one q^i_m'' and extract its semantic vector s^i_m'';
    if s^i_m'' > MAX(s^i_m, s^i_m') − γ then
        Q^i = Q^i ⋃ {q^i_m''} ;
    Remove q^i_m and q^i_m' from Q^i_temp;
Employ Non-Maximum Suppression (NMS) to the candidate set.
```

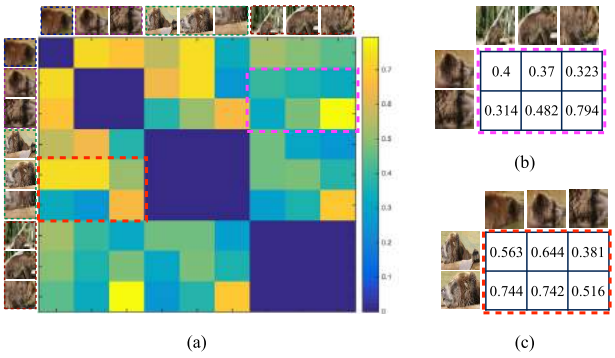

Fig. 4. Similarity matrix of selected candidates from 4 images is shown in (a). Candidates surrounded with different bounding boxes come from different images. In (b) and (c) we show the selected similarity scores from (a) and find that discovery of co-occurring candidates may focus on candidates covering part of the objects.

C. Co-Occurring Candidate Discovery

Based on the candidate collections of multiple images, we dedicate to explore co-occurring objects among these images. Previous works [9], [11] focus on exploring a single optimal matching clique from the graph to represent co-occurring foreground. However, since candidates covering part of the objects may perform higher similarity than that between two covered object and part or object and object, the single optimal solution may select incomplete discovery candidates of the common objects.

To specifically describe matches of incomplete objects, we present a similarity matrix of candidates from four different images in Fig. 4. We find that two candidates with highest similarity only cover part of the objects, as shown in Fig. 4 (b). The coverage with part of the objects induces incomplete discovery and hence influences the final co-segmentation results. Thus, to address the aforementioned drawbacks, we propose to explore multiple matching cliques to represent the co-occurring relationships of common objects.

To explore multiple matching cliques from candidates, we construct an N-partite complete graph each part of which indicates one image. Each part consists of candidates of the corresponding image as shown in Fig. 5. Weights of edges between parts are defined as feature similarity of pairs of candidates from different parts, i.e., inter-link. We delete the edges between candidates from one image, i.e., intra-link. The exclusion of intra-links in the N-partite graph can inherently separate candidates into different cliques without additional constraints. Based on the constructed graph we devote to globally discovering multiple maximum weighted matching cliques, each of which represents co-occurring relationship among multiple images. We seek a global optimal solution on candidate selection to satisfy the maximum of energy function. Besides, considering that there may not have similar candidates in some images, we add a virtual node (the gray

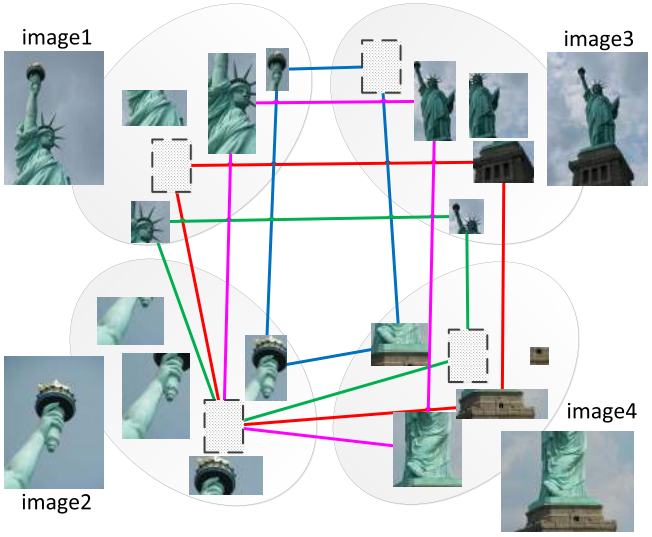

Fig. 5. The schematic diagram of our multiple matching cliques. We draw four matching cliques to demonstrate the effectiveness of our method. The gray blocks with dotted line represent virtual nodes. Each clique is a complete subgraph of the N-partite graph and for clarity we omit the edges between candidates from nonadjacent parts.

block as shown in Fig. 5) for each part in the graph. When similarity between two candidates is too small, we prefer to link a candidate belonging to one image to the virtual node belonging to the other one. Therefore, we can obtain multiple cliques, as indicated by different colors in Fig. 5. For clarity for each clique, we do not draw links between regions from disjoint parts in Fig. 5. The details on exploring multiple cliques are described in follows.

1) N-Partite Graph Construction: Based on the candidate collections $\mathcal{Q} = \{\mathbf{Q}^1, \mathbf{Q}^2, \ldots, \mathbf{Q}^N\}$ where $\mathbf{Q}^i = \{q_1^i, q_2^i, \ldots, q_{|\mathbf{Q}^i|}^i\}$ is the candidates of image $\mathbf{I}^i$, we construct an N-partite graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{S}, \mathcal{W})$. $\mathcal{V} = \bigcup_{i=1}^{N} \mathbf{V}^i$ is the set of nodes in the graph $\mathcal{G}$. $\mathbf{V}^i = \{v_1^i, v_2^i, \ldots, v_{|\mathbf{Q}^i|}^i, v_v^i\}$ is the i^{th} part of graph $\mathcal{G}$ corresponding to the collection $\mathbf{Q}^i$ of image $\mathbf{I}^i$. v_v^i is the virtual node assigned to the i^{th} part. Node v_m^i represents the candidate q_m^i and the "objectness" score s_m^i is assigned as the weight of node v_m^i, i.e., $\mathcal{S} = \{s_m^i\}$. $\mathcal{E} = \{e_{jn}^{im}\}$ is the set of edges in the graph where e_{jn}^{im} is the edge connecting node v_m^i and v_n^j. In the N-partite complete graph, each node is connected with others from different parts. $\mathcal{W} = \{w_{jn}^{im}\}$, where w_{jn}^{im} is the weight of edge e_{jn}^{im} and assigned as feature similarity between nodes v_m^i and v_n^j.

2) Multiple Clique Matching Using Mixed-Binary Integer Program: Given the constructed N-partite graph, we seek to explore multiple co-occurring relationships among images, i.e., finding co-occurring candidates via exploring multiple matching cliques $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_K\}$, where K is the number of cliques. The cliques must take into account all candidates extracted before. We set K as 50 to be able to encompass the maximal number of candidates contained by one image. We only consider the clique that at least has one candidate to be a valid estimation of co-occurring relationship. Otherwise it will be ignored.

To obtain multiple cliques each of which contains N components from N different images, the energy function of selecting candidates and edges can be formulated as:

$$E(\mathbf{v}, \mathbf{e}) = \underbrace{\sum_{i,m} s_m^i v_m^i + \sum_{i,j,m,n} w_{jn}^{im} e_{jn}^{im}}_{candidate\ nodes} + \underbrace{\sum_{i} s_v^i v_v^i + \sum_{i,j,m} w_v e_{jd}^{im}}_{virtual\ nodes}. \quad (1)$$

The first two terms describe the unary and binary energies on the selection of candidates and edges between candidates, respectively. The last two describe selection of the virtual nodes. i and j are the indices of images and range from 1 to N. m and n are the indices of candidates that come from image $\mathbf{I}^i$ and $\mathbf{I}^j$, respectively. $v_m^i \in \{0, 1\}$ and $v_v^i \in \{0, 1\}$ are binary variables to indicate whether the corresponding node is selected or not. s_m^i and s_v^i are the corresponding weights of the nodes v_m^i and v_v^i. $e_{jn}^{im} \in \{0, 1\}$ and $e_{jd}^{im} \in \{0, 1\}$ indicate the selection of edges and w_{jn}^{im} and w_v are the corresponding weights.

In the formulation we ignore the unary term acted in energy function. Unary term is always assigned with "Objectness" confidence score of the candidate. The score in our framework is defined as maximal response of semantic feature that presents probability of belonging to an object category. However, compared with confidence scores employed in previous works, which reflect probability of being an object corresponding to others, scores of candidates have no relation with each other in our paper. For example, a candidate belonging to cat category with 0.5 probability only indicates that it has higher probability of being a cat than being a dog. We cannot conclude that a candidate with 0.5 score of being a cat has higher probability than one with 0.3 of being a dog. Since the magnitude of confidence score only reflects prominence among categories, but not among candidates. Therefore, we ignore the costs generated by nodes, i.e., s_m^i and s_v^i are 0, in our formulation.

In the above-mentioned binary based formulation, the number of virtual nodes has to be set as a large value since the components selected from one part have exclusive property, i.e., components are different from each other. The upper bound of the number of virtual nodes added into each part is computed as follow. The number of virtual nodes added into each part is equal to the summation of candidates from the other parts, i.e., $N_v^i = \sum_{j \neq i} |\mathbf{Q}^j|$. N_v^i is the number of virtual nodes added into part i and $|\mathbf{Q}^j|$ is the number of candidates in part j. Increment of the number of nodes in each part would lead to aggravation on the computational complexity when solving optimal solution based on the graph.

To reduce the increasing computation complexity brought by adding virtual nodes, we relax the binary enforcement on variables and force instead the integer solution to the selection of virtual nodes. We replace the binary indication with integer indication on selection of virtual nodes to allow any times of selections, i.e., replace $v_v^i \in \{0, 1\}$ to $v_v^i \in \{0, 1, \ldots, int\}$. The value of int reflects the number of connections linked

to the virtual node of part i. Thus we only need to add one virtual node to each part without aggravating much complexity. Selections between nodes from other parts and current virtual node is defined as the number of selections of this virtual node, i.e., $\sum_{i,j} e_{jn}^{im} = v_v^i$ for any $j \in \{1, \ldots, N\}$, $j \neq i$ and $n \in \{1, \ldots, |\mathbf{Q}^j|\}$.

To explore multiple maximum weighted matching cliques, we introduce three constraints to control the selection of components and edges in each clique.

Constraint 1 enforces that the number of selections in part i must be equal to the product of number of cliques (K) and that of the rest parts ($N-1$).

$$\sum_{j=1, i\neq i}^{N} \sum_{n=1}^{|\mathbf{Q}^j|} \sum_{m=1}^{|\mathbf{Q}^i|} e_{jn}^{im} + v_v^i = (N-1)K, \ i \in \{1, \ldots, N\}, \quad (2)$$

where N is the number of parts and n is the index of candidate in part j. $|\mathbf{Q}^j|$ and $|\mathbf{Q}^i|$ are the number of candidates corresponding to part j and i, respectively.

Previously we demonstrate that all candidates of each part must be included in the set of matching cliques. And the number of selected nodes for each part must be equal to K. Besides, we enforce that once a candidate is selected, $N-1$ of its edges must also be selected. Further, due to the existence of virtual nodes, the selections for each part not only include links between candidates, but also include that between a candidate and a virtual node. Thus the number of selections for each part is presented as sum of the number of candidates links and virtual node as shown in Eq. 2.

Constraint 2 enforces that the number of selected edges between one candidate and candidates from any other part is no more than one.

$$\sum_{n=1}^{|\mathbf{Q}^j|} e_{jn}^{im} \leqslant 1, \forall i, j \in \{1, \ldots, N\}, i \neq j, m \in \{1, \ldots, |\mathbf{Q}^i|\}. \quad (3)$$

Once a candidate is selected, enforcing the number of selected edges between this candidate and candidates of another part can be employed to define the number of selected candidates from the corresponding part. Thus this constraint can efficiently control the number of candidates selected from this part. Besides, since we constrain that each candidate must be presented once and only once in the set of matching cliques, it is obviously that one candidate belongs to one clique. Thus, by limiting the number of selected edges to be no more than one, we can prevent one candidate from existing in more than one clique.

Constraint 3 requires that edges in each clique must be formulated as a cycle.

$$e_{jn}^{im} + e_{kt}^{jn} \leqslant 1 + e_{kt}^{im}, \forall i, j, k \in \{1, \ldots, N\}, i \neq j \neq k. \quad (4)$$

The global consistency restricts the correspondences among all selected candidates in each clique. We constrain that the selection of edges for each clique must formulate a cycle. That is, if there exist connections both linking nodes v_m^i and v_n^j ($e_{jn}^{im} = 1$) and nodes v_n^j and v_k^t ($e_{kt}^{jn} = 1$), there must exist a connection between nodes v_m^i and v_k^t, i.e. $e_{kt}^{im} = 1$. If there does not exist connection between nodes v_m^i and v_k^t, i.e. $e_{kt}^{im} = 0$, either the connection between nodes v_m^i and v_n^j or that between nodes v_n^j and v_t^k is disconnected, i.e. either $e_{jn}^{im} = 0$ or $e_{kt}^{jn} = 0$.

By incorporating the above-mentioned constraints, we reformulate the energy of function $E(\mathbf{v}, \mathbf{e})$ and seek optimal configuration on $\mathbf{e}$. We concatenate all binary and integer variables for edges and virtual nodes into a vector $\mathbf{x}$. Considering the dimension consistency in matrix operation, we add intra-edges for each candidate in $\mathbf{x}$ and restrict the value of variables on intra-edges ($e_{ij'}^{ij}$) to be 0. Thus the variable $\mathbf{x}$ is defined as $\mathbf{x} = [e_{11}^{11}, e_{12}^{11}, \ldots, e_{N|\mathbf{Q}^N|}^{11}, v_v^1, \ldots, e_{N|\mathbf{Q}^N|}^{N|\mathbf{Q}^N|}, v_v^N] \in \mathbb{R}^{(|V|^2+N)\times 1}$. $|V| = \sum_{i=1}^{N} |\mathbf{Q}^i|$ is the total number of candidates in graph $\mathcal{G}$. The weights for edges and virtual nodes are concatenated as the same and constitute the vector $\mathbf{c}$. The final objective function is represented as:

$$\begin{aligned} &\arg\max_{\mathbf{x}} \ \mathbf{c}^T\mathbf{x} \\ &\text{s.t. } \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}, \\ &\quad \mathbf{C}\mathbf{x} \leq \mathbf{d}. \end{aligned} \quad (5)$$

$\mathbf{A}$ and $\mathbf{b}$ are components of the unified equality constraint by integrating constraints in Eq. 2. $\mathbf{C}$ and $\mathbf{d}$ are components of the inequality constraint rewritten from Eqs. 3 and 4. The objective function is solved by using Mixed-Binary Integer Program (MBIP).

D. Co-Occurring Measurement Construction

Based on the discovery of multiple matching cliques, we can obtain the rough estimation of co-occurring relationships among candidates. Although candidates in the explored multiple matching cliques can not perform "good" segmentations, they can still provide a rough estimation about the location of co-occurring objects. We propose to estimate the probability of co-occurrence for candidates via weighting the co-occurrence of cliques. We simply consider the co-occurring weight of each clique as the number of candidates existed in this clique in Eq. 6. We also ignore the cliques which contain less than two candidates.

$$S(c_c) = \sum_{i}^{N} \Pi(c_c^i). \quad (6)$$

$\Pi\ (c_c^i)$ is a indication function and equals to 1 when c_c^i indicates a candidate. Then we initialize the co-occurring score for each candidate as weight of the corresponding clique, i.e., $S(q_i) = S(c_c)$ and q_i belongs to c_c.

For each candidate we additionally integrate the estimated global scores with spatial scores to formulate a regularized weight. The score of a candidate q_i is computed as:

$$S(q_i) = \frac{1}{|\text{Aff}(q_i)| + 1}\big(S(q_i) + \sum S(q_{i'})\big). \quad (7)$$

$q_{i'}$ is the i'^{th} candidate in the affinity set of q_i. The affinity set $\text{Aff}(q_i)$ consists of candidates with compositional relationships in the corresponding merging tree, i.e., the parent or child

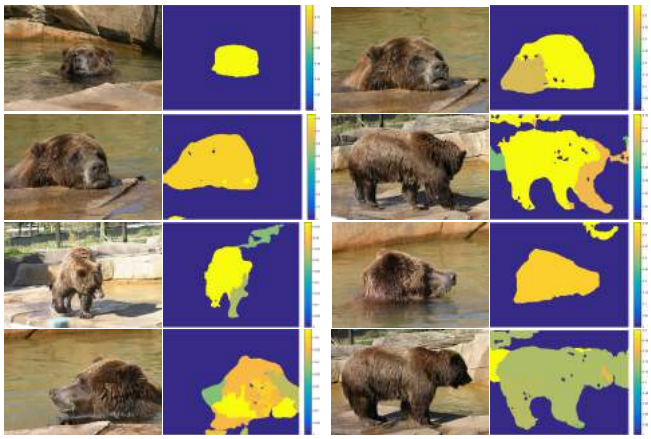

Fig. 6. Co-occurring score maps for images from the same category. The brighter color indicates higher probability of co-occurrence.

candidates in the merging tree. $|\mathrm{Aff}(q_i)|$ is the number of candidates in this set.

To obtain pixel-wise co-occurrence map, for each pixel I_p^i, we take into account the score of all candidates that include this pixel. We estimate the co-occurring score for pixel S_p^i by averaging the scores of all relevant candidates.

$$S_p^i = \sum_{t=1}^{T} S(q_t), \tag{8}$$

where t is the index of a candidate that contains pixel I_p^i and T is the number of candidates. Then for each image we threshold the co-occurring scores via the median value of that image and obtain final co-occurrence map, as shown in Fig. 6.

The co-occurrence map can roughly locate the co-occurring objects in images. To get accurate binary segmentation for each image, we additionally employ Grab-cut segmentation method and visualize the final segmentation in the second column of Fig. 2 (c).

IV. Experiments

To evaluate the performance, we employ the proposed method on four public datasets, i.e., iCoseg [17], MSRC [37], PASCAL-VOC [8] and CosegRep [38]. The iCoseg dataset contains 38 categories, each of which includes images with the same or similar object instances, and totally has 643 images. The shared objects in each class always have similar color properties, but precisely segmenting this dataset is still very challenge due to variation of viewpoint, diversity of background, lighting, shadow, and object deformations and poses within each class. The MSRC dataset is introduced by [37]. It contains 14 classes and totally 420 images. Most images in this dataset have cluttered background that may distract the attention on foreground objects. The objects in each category also have large differences on color and deformation. PASCAL-VOC dataset consists of 20 classes from the PASCAL-VOC 2010 dataset and totally has 1,037 images. This dataset is more challenging due to the existence of distractive co-occurring objects and background, such as classes of person and potted plant. CosegRep dataset contains 23 classes and 572 images. 22 of 23 classes have different animals and flowers. The rest of the class, named 'Repetitive', has repeated similar shape patterns within each image, such as tree leaves.

We employ two common used measurements, Precision (**P**) and Jaccard index (**J**) to evaluate the performance of methods. Precision measures the number of accurately labeled pixels and is calculated as ratio of correctly labeled pixels and total number of image pixels. Jaccard index is obtained by computing the percentage between intersection and union of the segmentation result and groundtruth. Compared with Precision, Jaccard index is more reliable to show the precision of segmentation and provides more suitable evaluation.

To perfectly present the improvement of our proposed framework, we construct a simple baseline upon segmentation results from the employed FCN model. Firstly, the straightforward outputs are sent as the original input to the baseline. Subsequently, we explore a common object category which is contained by the most images. There is no requirement that this common object category is contained by all images. Finally, regions labeled by the common object category are regarded as common foreground for co-segmentation. A image that has no pixels assigned with this label will be entirely treated as wrong segmented one. The quantitative performance of different datasets of the baseline is shown in the penultimate row in Table III, IV, VI and VII, respectively.

A. Comparison on the Generated Candidates

We estimate the capability of our extracted candidates in three folds, i.e., Detection Rate (DR), Intersection over Union (IoU) of foreground covered by candidates (Foreground Ratio (FR)), and IoU of background covered by candidates (Background Ratio (BR)), on three public datasets, iCoseg, MSRC and PASCAL-VOC. We compare with four widely used candidate generation methods, i.e., "Bing" [29], "Proposals" [13], "EdgeBoxes" [35] and "SSD" [36]. For "Bing" and "EdgeBoxes", we use suggested parameter settings and select top 100 candidates with descending confidences. For "Proposals", we follow the suggested parameter setting. For "SSD", we employ the released 300 × 300 model trained on PASCAL VOC 2007 and 2012 dataset and preserve the given parameters. We state the superiority of generated candidates in two folds, i.e., image-level and box-level. For image-level, we perform IoU of foreground and background over all extracted candidates in Table II. For box-level, we plot DR curve with increasing number of candidates in Fig. 7.

It can be seen in Table II that candidates extracted by our method can almost obtain higher performance on foreground ratio and tremendous decrement on background ratio. We prominently outperform "Bing", "EdgeBoxes" and "Proposals" with less number of candidates. For example, in Table II we achieve 61.3%, 54.5% and 65.8% improvements of foreground ratio on iCoseg dataset, respectively.

We also perform the performance with the increased number of candidates in Fig. 7. We estimate detection rate of foreground with increasing number of candidates. Detection rate given #WIN candidates (DR-#WIN) is computed as ratio of

TABLE II

ANALYSIS ON REPRESENTATIVE ABILITY OF THE GENERATED CANDIDATES. WE EVALUATE OUR PROPOSED CANDIDATE GENERATION METHOD ON THREE CRITERIA, i.e., NUMBER OF EXTRACTED CANDIDATE, RATIO OF FOREGROUND OCCUPIED ON CANDIDATES AND THAT OF BACKGROUND OCCUPIED ON CANDIDATES. WE COMPARE OUR METHOD WITH FOUR WILDLY USED CANDIDATE GENERATION METHODS AND STATE THAT OUR METHOD HAS OUTSTANDING PERFORMANCE

	iCoseg			MSRC			PASCAL		
	#WIN	FG	BG	#WIN	FG	BG	#WIN	FG	BG
Objectness [29]	100	22.50%	74.20%	100	32.40%	65.20%	100	17.03%	79.91%
Proposals [13]	38	21.90%	78.10%	98	31.60%	68.50%	98	16.48%	83.52%
EdgeBoxes [35]	100	23.50%	49.30%	100	34.30%	41.10%	100	18.30%	56.80%
SSD [36]	2	22.51%	28.94%	1	47%	42%	2	19.40%	33.71%
Ours	8	36.60%	36.30%	4	46.30%	38.30%	6	25.55%	51.57%

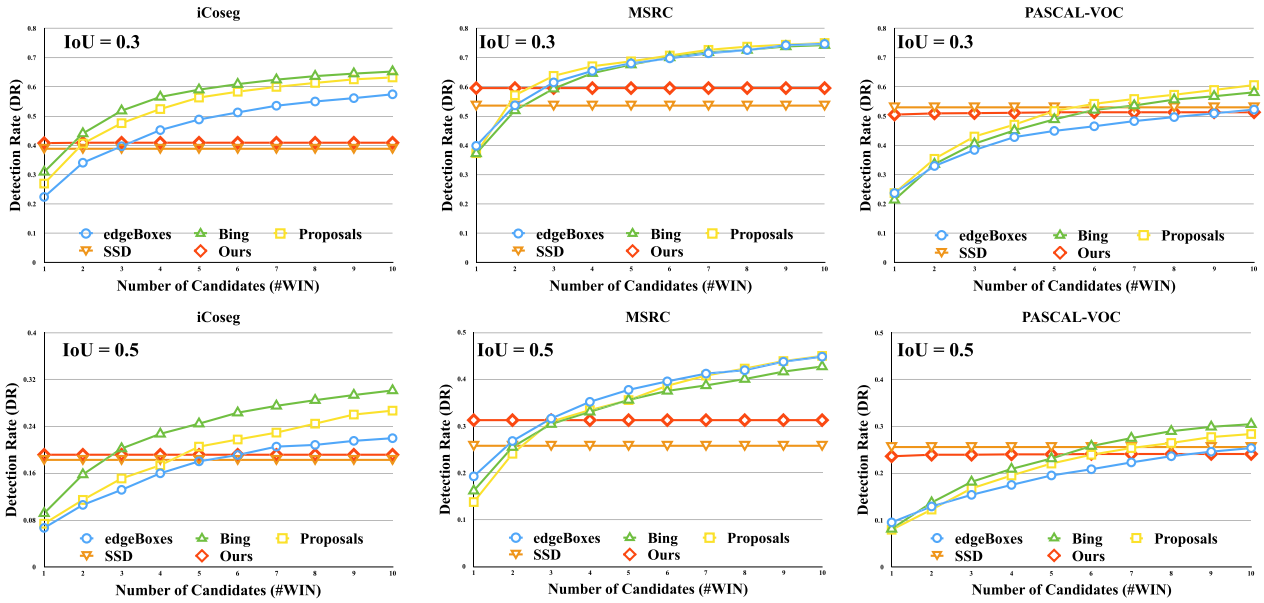

Fig. 7. Tradeoff between number of candidates (#WIN) and Detection Rate (DR) on three datasets. The first row shows performance of DR-#WIN under 0.3 IoU and the second row is DR-#WIN under 0.5 IoU. Three columns correspond to iCoseg, MSRC and PASAL-VOC dataset, respectively.

detected foreground under #WIN candidates with a given IoU threshold. We measure performance of candidates under two IoU thresholds, 0.3 and 0.5. Different from object detection which has bounding box groundtruth for each object, we only have pixel-level groundtruth indicating whether a pixel belongs to foreground. The area of pixel-level groundtruth is larger than bounding box groundtruth when multiple instances exist in the image. Thus we select two lower IoU thresholds, 0.3 and 0.5. For clear observation, we show the performance of the top 10 candidates. To obtain DR for methods with less than 10 candidates, we conduct the same operation as mentioned in "Bing" by assigning with the previous result. Columns in Fig. 7 indicate performance on different datasets and rows are different measurement criteria. The first row shows DR-#WIN results with setting IoU threshold as 0.3. The second column is DR-#WIN under 0.5. We obtain superior performance, i.e., higher DR-#WIN, on iCoseg and MSRC datasets. The generated candidates can be more exact to catch foreground and eliminate background.

Our method performs better on iCoseg and MSRC compared with "SSD", as shown in Fig. 7, but gets inferior results on PASCAL-VOC dataset. We ascribe these inferior results to the large gap of training data. "SSD" uses PASCAL VOC 2007 and 2012 datasets and forms a training set with 16, 551 images. While our FCN uses part of PASCAL context dataset which only contains 3, 480 images. Such a large distance of training data will lead to inferior performance. Besides, "SSD" has inferior generalization ability to unseen object categories. "SSD" gets poor performance on some categories which are absent in training dataset, such as "Stonehenge" and "Pyramid" categories in iCoseg dataset. This limitation is inherent. Because "SSD" aims to conduct object detection task and treats problem of object detection as object classification. Thus it will have good performance on seen object categories and get inferior results on unseen objects.

B. Effect of Matching Based on N-partite Graph With Augmented Virtual Nodes

In this section we illustrate the performance of matched results based on *N*-partite Graph. The multiple matched cliques aim to enforce inclusion of all candidates and exclusion of irrelevant candidates dissimilar to the rest in each clique. To clearly perform the ability of multiple matched cliques,

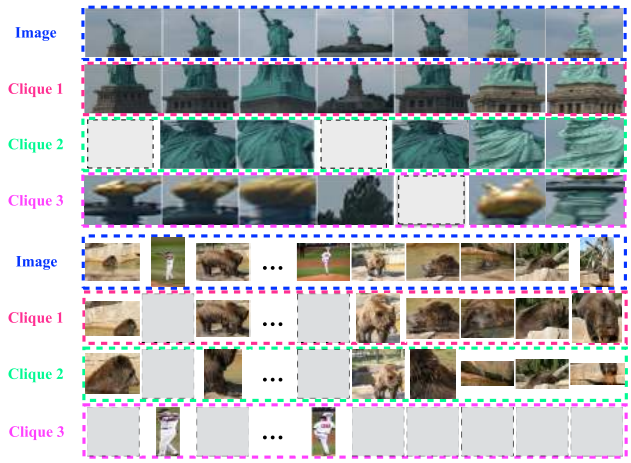

Fig. 8. Matching results for images selected from two sets. The top group performs matching results of "Statue of Liberty" to show that they can separate candidates with different parts into different cliques. The bottom group presents matching results of adding outliers to Category "AlaskanBear." It shows the discriminative ability to exclude outliers containing no co-occurring objects.

we present some matching results in Fig. 8. As shown in the top group of Fig. 8, the matched cliques of "Statue" category effectively discriminate co-occurring candidates depicted different parts of the objects. The matched cliques (2^{nd} to 4^{th} row in the top group of Fig. 8) illustrate the superiority of exploring multiple co-occurrence without suffering the deficiency induced by consideration of part of the candidates and irrelevant matches. The clique in 2^{nd} row connects all candidates representing the bottom of "Statue" among all images. The cliques in 3^{rd} and 4^{th} rows present the ability to prevent dissimilar matches, i.e., the first and fourth images do not contain candidates depicting the upper body of "Statue".

The matched cliques can not only handle dissimilar matches from candidates but also exclude irrelevant matches from outlying images. We present the matched results for the condition that adding outliers to the original image set. We additionally build a category consisting of the original images from "Bear2" category and outliers from "Baseball" category, as shown in the first row of the bottom group in Fig. 8. We present three matched cliques shown in 2^{nd} to 4^{th} rows of the bottom group. The first two cliques focus on discovering co-occurrence of bear-related candidates and effectively exclude candidates from outliers. And the third clique considers the co-occurrence of candidates from outliers and also can prevent inclusion of bear-related candidates. The separation of candidates from original images and outliers can distinctly identify the co-occurring response for each image.

C. Qualitative and Quantitative Results

1) Comparison on iCoseg Dataset: In this section, we perform the evaluation on the iCoseg dataset. The iCoseg dataset contains 38 categories with 643 images. To obtain final binary segmentation results, we employ Joint-Grab-cut [3], which computes color models to all images together, incorporated with the computed co-occurring maps.

TABLE III
COMPARISON WITH THE-STATE-OF-THE-ARTS ON ICOSEG DATASET

	Jou10 [14]	Kim11 [39]	Jou12 [21]	Rub [15]	Fak13 [8]
P	61.0%	66.6%	70.2%	89.9%	92.8%
J	0.39	0.38	0.43	0.69	0.73
	Lee15 [40]	**Quan16 [28]**	**Tao17 [41]**	**FCN [42]**	**Ours**
P	91.2%	93.3%	90.8%*[1]	84.7%	**93.8%**
J	0.7	0.76	0.74*	0.64	**0.77**

[1]The experiments are conducted on part of iCoseg dataset, which has 31 categories compared with the original 38 categories.

We compare results on iCoseg dataset with recent proposed methods for object co-segmentation [8], [14], [15], [21], [28], [39]–[41], all of which obtain good segmentation quality. We also show improved performance compared with results of baseline. Zhiqiang *et al.* * [41] conducts experiments on part of iCoseg dataset, which has 31 categories with 530 images. Numerical results are shown in Table III. Our method obviously takes improvements on Precision and Jaccard index criteria. We obtain 10.7% and 20.3% higher performance on Precision and Jaccard index over results of FCN [42], respectively. Compared with the second-best method [28], our method has a slightly increase on Precision and a larger improvement on Jaccard index. Additionally, we perform category-level Precision and Jaccard index evaluation for some classes in iCoseg dataset in Fig. 9(a) and Fig. 9(b), respectively, to clearly illustrate the advantages of our method. Our method not only performs better Precision and Jaccard index on categories existed in FCN training data, such as "Baseball" (person category in FCN, which obtains improvements of 6% on Precision and 16% on Jaccard index compared with [8]) and "Ferrari" (car category in FCN, which gets improvements of 4.5% on Precision and 9% on Jaccard index compared with [8]). We also have higher performance on those which do not presented in FCN training, for example, "Bear2" (which has increase of 2% on Precision and 3% on Jaccard index) and "panda2" (which has increase of 14% on Precision and 23% on Jaccard index). The high performance demonstrates that the category gap between data in FCN training and co-segmentation can be weakened by the proposed method. Although categories like "bear2" and "panda" do not exist in procedure of training FCN model, their segmentations can be compensated by some related categories like "dog" and "cat" since their pixels have similar appearances.

Moreover, we conduct the comparison with other more recent proposed methods [19], [38], [43], [44]. Data used by these methods contains 26 of the 38 categories in iCoseg dataset. Detail information of these data can be seen in [43]. We present comparison of Precision in the second row of Table V and obtain large improvement over other methods.

Some qualitative results are presented in Fig. 11(a). The original and groundtruth images are shown in 1^{st}, 2^{nd} row, respectively. We also show the qualitative results of [15] in 3^{rd} row. The final segmentation results of our method are presented in 4^{th} row. From the visualization of segmentation, it can be seen that our method can not only distinguish the boundaries between foreground and background, like

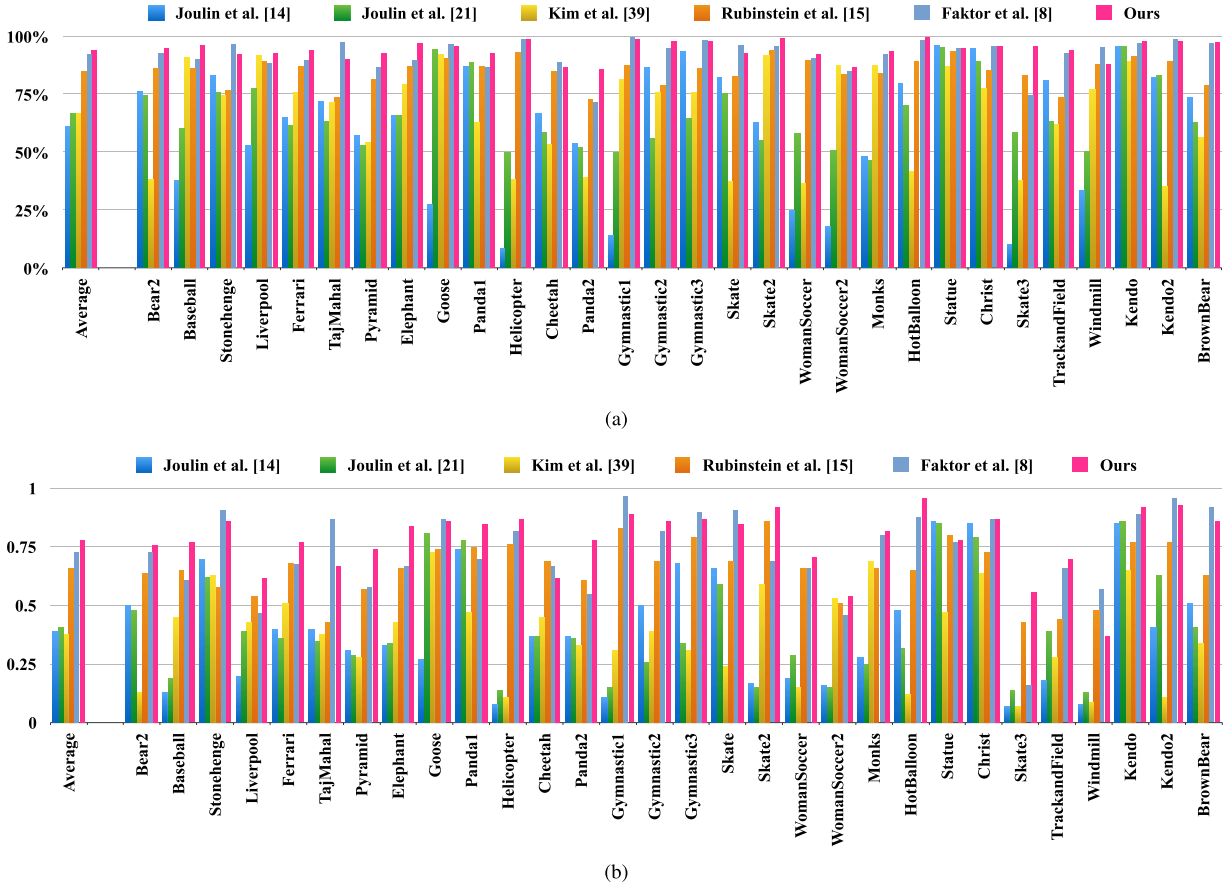

Fig. 9. Performance of categories in iCoseg dataset. We present the Precision and Jaccard index results for each category in (a) and (b), respectively. Our proposed method performs better on most of categories, not only on categories seen in FCN, but also those have not been known like "Panda" and "Bear". (a) Precision on Each Category in iCoseg Dataset. (b) Jaccard index on Each Category in iCoseg Dataset.

TABLE IV
COMPARISON WITH THE-STATE-OF-THE-ARTS ON MSRC DATASET AND ITS SUBSET

	MSRC		MSRC Subset	
Methods	**P**	**J**	**P**	**J**
Joulin et al. [14]	70.5%	0.45	-	-
Rubinstein et al. [15]	87.7%	0.68	92.2%	0.75
Joulin et al. [21]	73.3%	0.51	-	-
Kim et al. [39]	54.4%	0.34	-	-
Factor et al. [8]	89.2%	0.73	92%	0.75
Lee et al. [40]	84.4%	0.6	83.8%	0.52
FCN [42]	78.0%	0.64	87.7%	0.48
Ours	**90.9%**	**0.73**	**93.6%**	**0.76**

TABLE V
COMPARISON ON PART CATEGORIES OF ICOSEG AND MSRC PRESENTED IN [43]. THE EXPERIMENTAL RESULTS OF [43], [44], [19] AND [38] ARE REPORTED IN [43]. THE DETAILS OF CATEGORIES USED HERE CAN BE SEEN IN [43]

Methods	**Cate-gories**	**Ours**	[43]	[44]	[19]	[38]
iCoseg(P)	26/38	**94.6%**	92.2%	-	88.7%	76.1%
MSRC(J)	8/14	**0.74**	0.63	0.66	0.57	0.65

"baseball" and "skate3", but also can precisely preserve the objects like "panda".

2) Comparison on MSRC Dataset: We also evaluate our method on MSRC dataset and compare it with representative methods [8], [14], [15], [21], [39]. Numerical results are presented in Table IV. We achieve obvious improvements over [15] and [14] with 4% and 24% advancement, respectively, and get slightly superiority over [8]. We also examine the effectiveness on subset of MSRC dataset which contains 7 categories with 145 images in total. We compare the results with Rubinstein *et al.* [15] and Faktor and Irani [8]. The detail numerical results can be seen in Table IV.

For clarity we present performance for each category in Fig. 10. Fig. 10(a) shows the performance on Precision and Fig. 10(b) shows the performance on Jaccard index.

Additionally, we also conduct comparison on Jaccard index with other more recent proposed methods [19], [38], [43], [44] in Table V. The experiment is conducted on 8 of 14 categories in MSRC dataset as mentioned in [43]. The numerical results of four comparing methods are obtained from [43]. Salient improvement, with a maximal increase of 17.5%, can be seen in Table V.

Some qualitative results for MSRC dataset are presented in Fig. 11(b). It is shown that our method can correctly reject the disturbance from cluttered background and preserve integrated object information after co-segmentation.

3) Comparison on PASCAL Dataset: In this section we conduct the proposed method on a more challenging cosegmentation dataset, PASCAL dataset, which has large variation on object and cluttered background. PASCAL dataset

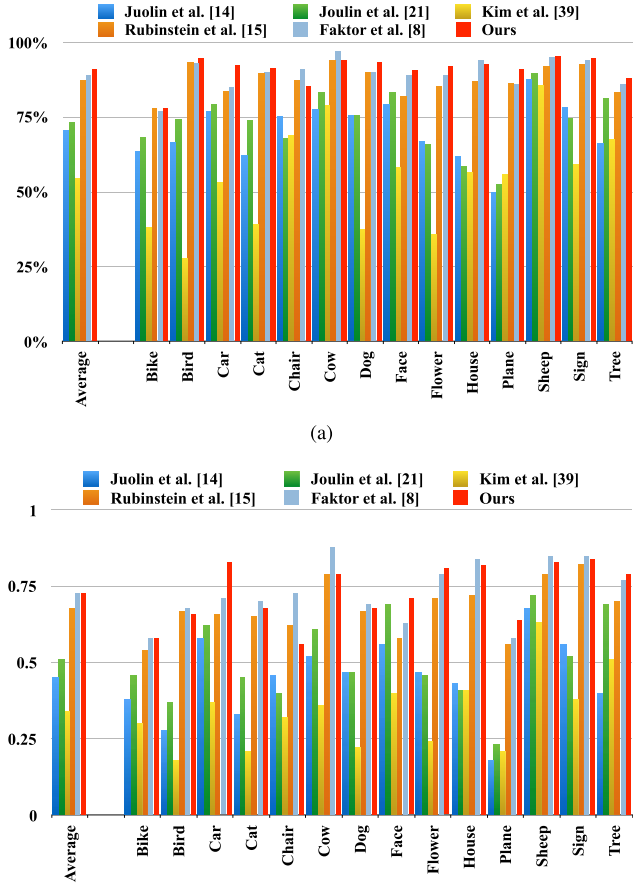

Fig. 10. Performance of categories in MSRC dataset. We present the Precision and Jaccard index results for each category in (a) and (b), respectively. Our method performs superior results on most categories. (a) Precision for Each Category in MSRC Dataset. (b) Jaccard index for Each Category in MSRC Dataset.

TABLE VI

COMPARISON RESULTS ON PASCAL VOC DATASET. WE SLIGHTLY IMPROVE THE PERFORMANCE ON PRECISION AND HAVE LARGE ASCEND ON JACCARD INDEX

PASCAL VOC	Ours	Faktor et al. [8]	Grab-Cut [45]
P	**84.3**%	84%	76%
J	**0.522**	0.429	0.38

is constructed based on PASCAL-VOC 2010 and contains 20 categories and 1,037 images. Categories in the dataset cover not only objects which have large intra-category variance but also those which have little discrimination with background. We present the comparison results of our method, [8] and [45] in Table VI. As can be seen in the table, our method outperforms the other two methods and get large improvement on Jaccard index. We show the comparisons of qualitative result with [8] in Fig. 12(a). Our method effectively reduces the prediction of background, which is suppresses by semantic information.

4) Comparison on CosegRep Dataset: The CosegRep dataset proposed in [38] has 23 categories with 572 images in total. 22 of the 23 categories are different categories

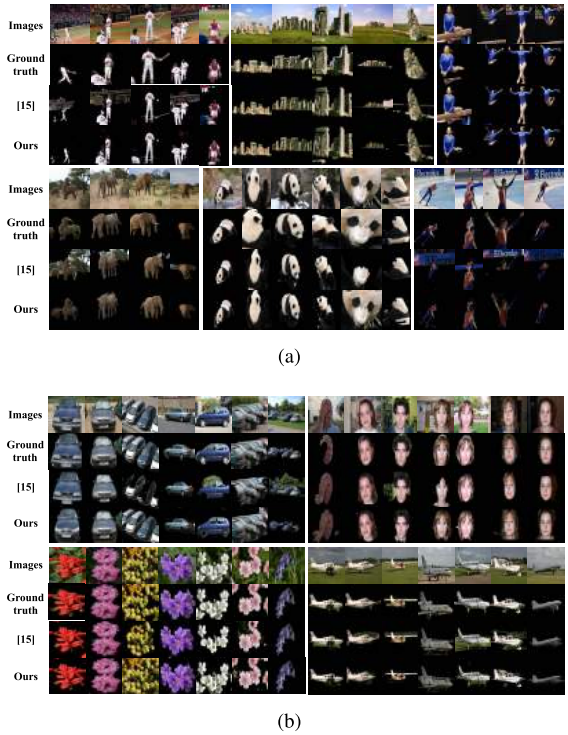

Fig. 11. Qualitative results on the iCoseg and MSRC datasets. We show some comparison with Rubinstein *et al.* [15]. Our method effectively reduce the segmentation of background. (a) Co-segmentation Results of iCoseg Dataset. (b) Co-segmentation Results of MSRC Dataset.

TABLE VII

COMPARISON RESULTS ON REPETITIVE CATEGORY OF COSEGREP DATASET. WE OBTAIN IMPROVEMENTS BOTH ON PRECISION AND JACCARD INDEX

Repetitive	Ours	Dai et al. [38]	Jerripothula et al. [46]	Jerripothula et al. [47]
P	**87.9**%	86.2%	-	-
J	**0.782**	0.754	0.747	0.776

of animals and flowers. Moreover, the remaining category, called "Repetitive", is special since each of 116 images has repetitive instances and similar shape pattern. It also has large object variations like difference between bird and sculpture of bird. Moreover, some images have distractive background such as object shading. Cosegmentation results are presented in Table VII. We make comparison with three the-state-of-the-art methods, [38], [46] and [47], and obtain marked improvement. We also present some qualitative results in Fig. 12(b). Our method preforms better with more accurate object boundaries and less prediction of background.

D. Discussion

In this section, we discuss generalization of the proposed framework, as well as failure cases.

1) Generalization: Our method has good generalization capacity and achieves good performance on unseen object categories, as shown in Fig. 13(a). Although unseen objects cannot be segmented with true category label, most of them

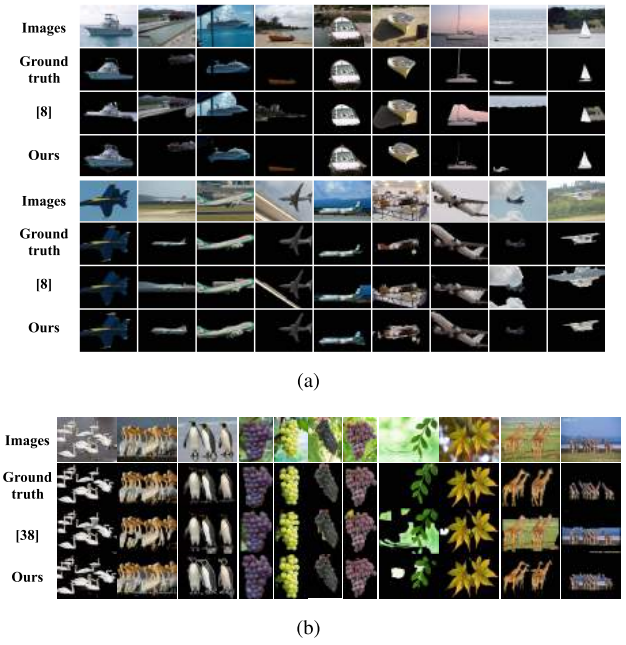

Fig. 12. Qualitative results of PASCAL VOC and CosegRep datasets. (a) Qualitative results on "boats" and "aeroplane" categories of PASCAL VOC dataset, compared with [8]. Our method effectively reduces segmentation of background. (b) Some qualitative results from "Repetitive" category in CosegRep dataset, compared with [38]. Our method achieves better performance on objects with less background.

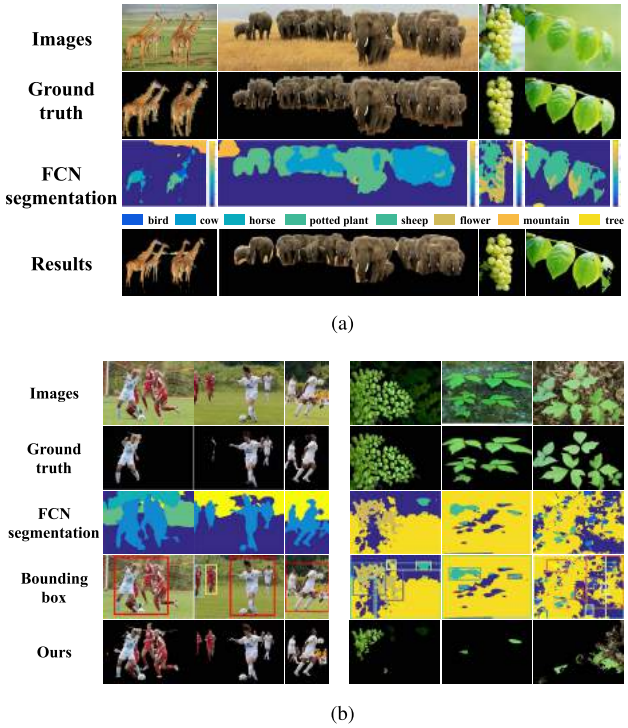

Fig. 13. Segmentation results on unseen categories of FCN model are shown in (a). Although these categories are not presented in FCN training dataset, there always exist similar components between known categories and unseen ones. Some failure cases are presented in (b). Since FCN is not an instance-level framework, we fail to identify adjoining objects from the same category. (a) Results on unseen categories. (b) Failure cases.

are predicted with known categories that have pixels with the same RGB value. As shown in Fig. 13(a), objects of unseen categories "giraffe" and "elephant", are represented by known labels, such as "horse" and "cow". The same as for "grape" and "leaves" categories, whose objects are labeled with "potted plant" and "tree". Since semantic regions focus on local responses corresponding to parts of objects, they are hierarchically merged together to larger regions, which hold larger areas and higher semantic responses. Therefore, unseen objects are extracted as semantic region candidates for the subsequent steps.

2) Failure Case Analysis: There are some failure cases of our method on distinguishing objects with the same category but different appearances like color. For example, separating women wearing white from women wearing red, i.e., "white-woman-soccer" and "red-woman-soccer" categories in iCoseg dataset, is difficult. We ascribe the poor performance to two main reasons. Firstly, our method cannot extract separated candidates for woman wearing white and red, since both have the same semantic response and adjoining spatial location. For example, as shown in the left of Fig. 13(b), since red bounding boxes in the first and third columns cover all women in the images, women wearing red are treated as a part and cannot be removed. Besides, since we estimate similarity between candidates based on semantic feature, candidates with the same semantic response but the different color, i.e., red and cyan bounding boxes in the second column of Fig. 13(b), have higher similarity score and are treated as the same objects.

For images whose background has the same semantic response with foreground, like images shown in the right of Fig. 13(b), our method fails to locate foreground and extract background as candidates. During correspondence exploration, since candidates in these images, which cover background, have similar semantic responses with foreground, they are still treated as common objects. Therefore, our method cannot prevent these candidates covering background from being selected. Incorporating correspondence exploration with other information like edge may improve the performance.

V. Conclusion

In this paper, we explore the task of object co-segmentation in computer vision. We propose a novel object co-segmentation framework to discovery multiple matching cliques among semantic candidates for the complementation of the coverage of part of the common objects and exclusion of irrelevant matches. Firstly, the generated candidates contained semantic and object-level information effectively alleviate the presence of background and reduce the number of candidates. Secondly, the explored multiple maximum weighted matching cliques take into account all candidates so as to avoid the decrement of meaningful information. Based on the N-partite complete graph augmented with virtual nodes, the multiple cliques can be naturally separated without additional constraints due to the nonexistence of edges between candidates from the same part in the graph. And the additional virtual node in each part prevents irrelevant matches from being in one clique. Finally, the superior results demonstrate the

effectiveness of the proposed framework. Moreover, since we have not taken color features into the exploration of multiple cliques, the matched results are indiscriminating of the same objects with different color, such as person worn white shirt and red shirt. Thus although the semantic feature can provide abundant information, low-level features like color are still with high importance for object co-segmentation.

REFERENCES

[1] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching—Incorporating a global constraint into MRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 993–1000.

[2] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *Proc. Eur. Conf. Comput. Vis.*, Heraklion, Greece, Sep. 2010, pp. 465–479.

[3] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in imagenet," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 459–473.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Stateline, NV, USA, Dec. 2012, pp. 1097–1105.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE ICCV*, Dec. 2015, pp. 1440–1448.

[6] 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.*, Columbus, OH, USA, Jun. 2014, pp. 1464–1471.

[7] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2011, pp. 2217–2224.

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

[9] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based RGBD image co-segmentation with mutex constraint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4428–4436.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[11] D. Zhang, O. Javed, and M. Shah, "Video object co-segmentation by regulated maximum weight cliques," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, Sep. 2014, pp. 551–566.

[12] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFS with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 109–117.

[13] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 222–234, Feb. 2014.

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

[15] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in Internet images," in *Proc. CVPR*, Jun. 2013, pp. 1939–1946.

[16] J. C. Rubio, J. Serrat, A. López, and N. Paragios, "Unsupervised co-segmentation through region matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 749–756.

[17] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3169–3176.

[18] Z. Yuan, T. Lu, and P. Shivakumara, "A novel topic-level random walk framework for scene image co-segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 695–709.

[19] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE Trans. Multimedia*, vol. 14, no. 5, pp. 1429–1441, Oct. 2012.

[20] E. Ahmed, S. Cohen, and B. Price, "Semantic object selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 3150–3157.

[21] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 542–549.

[22] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 2294–2301.

[23] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 73–80.

[24] F. Meng, H. Li, K. N. Ngan, L. Zeng, and Q. Wu, "Feature adaptive co-segmentation by complexity awareness," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4809–4824, Dec. 2013.

[25] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[26] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 345–360.

[27] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2016.

[28] R. Quan, J. Han, D. Zhang, and F. Nie, "Object co-segmentation via graph optimized-flexible manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 687–695.

[29] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300 fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3286–3293.

[30] Z. Zhou, Q. M. J. Wu, F. Huang, and X. Sun, "Fast and accurate near-duplicate image elimination for visual sensor networks," *Int. J. Distrib. Sensor Netw.*, vol. 13, no. 2, p. 1550147717694172, 2017.

[31] Y. Chen, C. Hao, W. Wu, and E. Wu, "Robust dense reconstruction by range merging based on confidence estimation," *Sci. China Inf. Sci.*, vol. 59, no. 9, p. 92103, 2016.

[32] J. Li, X. Li, B. Yang, and X. Sun, "Segmentation-based image copy-move forgery detection scheme," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 507–518, Mar. 2015.

[33] X. Qi, J. Shi, S. Liu, R. Liao, and J. Jia, "Semantic segmentation with object clique potential," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2587–2595.

[34] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[35] C. L. Zitnick and P. Dollär, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 391–405.

[36] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[37] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 1–15.

[38] J. Dai, Y. N. Wu, J. Zhou, and S.-C. Zhu, "Cosegmentation and cosketch by unsupervised learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1305–1312.

[39] 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.*, Barcelona, Spain, Nov. 2011, pp. 169–176.

[40] C. Lee, W.-D. Jang, J.-Y. Sim, and C.-S. Kim, "Multiple random walkers and their application to image cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3837–3845.

[41] T. Zhiqiang, H. Liu, H. Fu, and Y. Fu, "Image cosegmentation via saliency-guided constrained clustering with cosine similarity," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 4285–4291.

[42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[43] K. Li, J. Zhang, and W. Tao, "Unsupervised co-segmentation for indefinite number of common foreground objects," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1898–1909, Apr. 2016.

[44] F. Wang, Q. Huang, M. Ovsjanikov, and L. J. Guibas, "Unsupervised multi-class joint image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3142–3149.

[45] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[46] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1896–1909, Sep. 2016.
[47] K. R. Jerripothula, J. Cai, F. Meng, and J. Yuan, "Automatic image co-segmentation using geometric mean saliency," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 3277–3281.

Chuan Wang is currently pursuing the Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Her current research interests include image cosegmentation, matching, and people counting.

Hua Zhang received the Ph.D. degree in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2015. He is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include computer vision, deep learning, multimedia, and machine learning.

Liang Yang received the B.E. and M.E. degrees in computational mathematics from Nankai University, Tianjin, China, and the Ph.D. degree in computer science from the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He is currently an Assistant Professor with the School of Information Engineering, Tianjin University of Commerce, Tianjin, China. His current research interests include community detection, semi-supervised learning, low-rank modeling, and deep learning.

Xiaochun Cao (SM'14) received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, since 2012. He spent about three years as a Research Scientist with Object Video Inc., From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has authored and co-authored over 120 journal and conference papers. He is a fellow of the IET. His dissertation was nominated for the University of Central Florida's university-level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition. He is on the Editorial Board of the IEEE TRANSACTIONS OF IMAGE PROCESSING.

Hongkai Xiong (M'01–SM'10) received the Ph.D. degree in communication and information system from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2003. Since 2003, he has been with the Department of Electronic Engineering, SJTU, where he is currently a Full Professor. From 2007 to 2008, he was with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, as a Research Scholar. From 2011 to 2012, he was a Scientist with the Division of Biomedical Informatics, University of California, San Diego, CA, USA.

His research interests include source coding/network information theory, signal processing, computer vision, and machine learning. He has published over 180 refereed journal/conference papers. He was the recipient of the Top 10% Paper Award at the 2016 IEEE Visual Communication and Image Processing, the Best Student Paper Award at the 2014 IEEE Visual Communication and Image Processing, the Best Paper Award at the 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, and the Top 10% Paper Award at the 2011 IEEE International Workshop on Multimedia Signal Processing.

Dr. Xiong he has been a member of the Innovative Research Groups of the National Natural Science since 2012. He served as a TPC member for prestigious conferences, such as the ACM Multimedia, ICIP, ICME, and ISCAS. He received the Yangtze River Scholar Distinguished Professor from Ministry of Education of China, in 2016, and the Youth Science and Technology Innovation Leader from Ministry of Science and Technology of China. He was a recipient of the Shanghai Academic Research Leader. In 2014, he received the National Science Fund for Distinguished Young Scholar and Shanghai Youth Science and Technology Talent. In 2013, he was a recipient of the Shanghai Shu Guang Scholar. In 2011, he was a recipient of the First Prize of the Shanghai Technological Innovation Award for Network-oriented Video Processing and Dissemination: Theory and Technology. In 2010 and 2013, he received the SMC-A Excellent Young Faculty Award of Shanghai Jiao Tong University. In 2009, he was a recipient of the New Century Excellent Talents in University, Ministry of Education of China.