IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received April 7, 2019, accepted May 8, 2019, date of publication May 15, 2019, date of current version May 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917152

A Survey of Object Co-Segmentation

ZHOUMIN LU[1,2], HAIPING XU[3], AND GENGGENG LIU[1,2]

[1]College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China
[2]Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350116, China
[3]College of Mathematics and Data Science, Minjiang University, Fuzhou 350108, China

Corresponding author: Genggeng Liu (liu_genggeng@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 61877010 and Grant 11501114.

ABSTRACT It is widely acknowledged that object segmentation is a significant research field for computer vision and a key process for many other visual tasks. In the past unsupervised single-image segmentation, there are often cases where the segmentation result is not good. In the current supervised single-image segmentation, it is necessary to rely on a large number of data annotations and long-term training of the model. Then, people attempted to segment simultaneously the common regions from multiple images. On the one hand, it does not need to use a large amount of labeled data to train in advance. On the other hand, it utilizes the consistency constraint between images to better obtain the object information. This idea can generate better performance than the traditional one did, resulting in many methods related to object co-segmentation. This paper reviews some classic and effective object co-segmentation methods, including saliency-based approaches, joint-processing-based approaches, graph-based approaches, and others. For different methods, we select two or three related models to elaborate, such as a model based on random walks. Moreover, in order to exhibit and evaluate these methods objectively and comprehensively, we not only summarize them in the form of flowcharts and algorithm summaries, but also compare their performance with visualization methods and evaluation metrics, such as intersection-over-union, consistency error, and precision-recall rate. From the experiment, we also attempt to clarify and analyze the existing problems. Finally, we point out the challenges and directions and open new venues for future researchers in the field.

INDEX TERMS Computer vision, semantic segmentation, object co-segmentation, joint processing, saliency, model evaluation.

I. INTRODUCTION

There is no doubt that vision is the most important means of human perception, and that images are the basis of vision. Therefore, image processing and analysis are widely used in fields, such as physiology, computer science, and scenes, such as military operation, remote sensing and meteorological scenarios. As a way of image processing, object segmentation is the technique and process of dividing an image into specific parts with unique properties and proposing objects of interest. It is also a key step from image processing to image analysis, as well as an important basis for image retrieval, object recognition and video tracking.

However, if segmentation is merely based on the brightness and color of the pixels in the image, there may arise various difficulties, resulting in segmentation errors, such as uneven illumination, noise and shadows. Therefore, people hope to introduce some knowledge-oriented and artificial intelligence methods that can effectively help correct the errors during the segmentation process.

Object co-segmentation is essentially a special type of object segmentation, which exploits both inter and intra-image priors, utilizes the consistency information of shared objects between images and divides multiple images simultaneously, in order to improve the segmentation effect. It has the same purpose as image segmentation.

Since Rother *et al.* [1] first proposed an image co-segmentation model based on markov random field, this issue has been increasingly focused on by researchers. In the beginning, it was only viewed as the extension of markov random field, such as [2], [3] and [4], based on which there have emerged many methods [5] with various characteristics. Object co-segmentation can be divided into image pair co-segmentation and image group co-segmentation according to the number of processed images, while it can be categorized into single foreground co-segmentation and multiple

The associate editor coordinating the review of this manuscript and approving it for publication was Farid Boussaid.

foreground co-segmentation according to the number of object classifications. In addition, there are many categories of object co-segmentation methods, such as saliency-based, joint-processing-based, graph-based and other models.

The following section will review the existing object co-segmentation methods based on the above perspectives, to gain a comprehensive understanding of the current co-segmentation technology and its progress, and provide a useful reference for subsequent research.

Compared with other existing reviews of co-segmentation [6], [7], this paper has the following advantages.

- More comprehensive
  In the previous reviews, only 3-4 methods were introduced and their types are poor. The latest year of the methods is only until 2013. However, in this paper, nine methods are introduced and their types are rich, including saliency-based model, sketch-based model, skeletonization-based model and so on. The years of the methods can cover from 2013 to now.
- More detailed
  The previous reviews only roughly introduced some methods or 1-2 steps in the methods. However, this paper describes the entire process of the methods in detail. Additionally, the flowcharts and algorithm summaries are provided for better understanding.
- More objective
  The previous reviews only described the algorithm in a textual manner. However, this paper enriches the experimental contrast section with visualization and a variety of evaluation metrics, which can facilitate the intuitive exploration of distinctions in performance between different methods.
- More profound
  Through observation of experimental results, this paper attempts to discover and analyze some existing problems. Moreover, it also expounds the challenges at this stage and introduces the directions of our future work, trying to give more references and insights to initiates and future researchers in this field.

II. METHODS

The co-segmentation methods have made great progress, including saliency-based, joint-optimization-based, graph-based and other models, which are introduced as follows.

A. SALIENCY-BASED MODEL

When faced with a scene, humans automatically process the regions of interest but selectively ignore regions of no interest. These regions of interest are referred to as salient regions. The saliency detection [8] aims to extract salient regions of the image by simulating human visual characteristics. This obviously coincides with the purpose of object segmentation, because the regions of interest are usually the objects to be segmented, which have higher saliency.

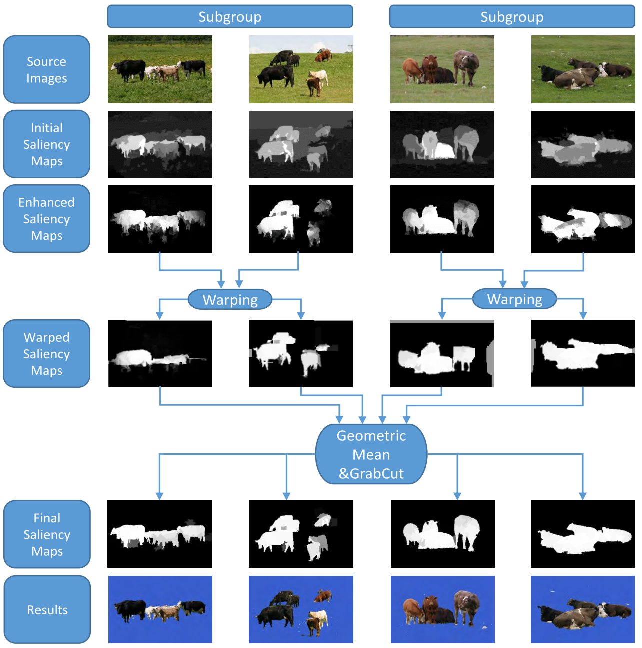

FIGURE 1. **An example for GMS. The saliency of other images can help to improve the saliency detection of a single image without detecting a common object.**

Nowadays, saliency has been widely used in various visual tasks. Also, people have begun to perform co-segmentation [9] via salient and common region discovery, such as [10], [11] and [12]. Next, three well-performed methods based on saliency will be introduced, and they have an inheritance relationship. The former method is the basis of the latter one.

1) GMS

In order to replace the complex co-labelling process, Jerripothula *et al.* [13] used the geometric mean saliency (GMS) to perform co-segmentation. It firstly obtains the global saliency maps by transmitting and merging the saliency information between the images in the group, and then gains the combined saliency map of each image. Finally, it utilizes these saliency maps to perform single-image segmentation. Its greatest advantage is that the saliency of other images helps to improve the saliency detection of a single image without detecting a common object. See Figure 1 for an example.

The initial saliency maps can be obtained by [14] and then converted into the binary maps $T = \{T_1, T_2, \cdots, T_m\}$ by means of Otsu's method [15]. In order to ensure that highly salient regions can adequately cover the objects to be segmented, the foreground needs to be more emphasized with the aid of saliency enhancement methods.

First, for the continuity of T_i, the saliency values of the background pixels need to be updated. When $T_i(p) = 0$, let $T_i(p) = T'_i(p)$, and $T'_i(p)$ is defined as follows.

$$T'_i(p) = \sum_{q \in D_i} \left| I'_i(p) - I'_i(q) \right| e^{\frac{-d_{pq}}{\sigma}} \tag{1}$$

where D_i denotes the domain of image I_i and I'_i indicates the gray image. d_{pq} represents the distance between p and q. In particular, set $\sigma = 25$.

Then, for avoiding the excessive penalties as a result of the low saliency values, the entire saliency maps need to be brightened by the following transformation.

$$M_i(p) = \log_{(1+\mu)}(1 + \mu T_i(p)) \tag{2}$$

where M_i denotes the brightened saliency map and $\mu = 300$.

After that, the weigh of each image is represented by GIST descriptor [16], [17], and then all images are clustered into subgroups by K-Means algorithm, respectively. The images within the subgroup have the higher similarity. As the number of subgroups, K is computed by $\lfloor m/10 \rfloor$. Then, the Dense-SIFT descriptor [18] is employed to match the corresponding pixels between images within each subgroup.

Next, for transmitting M_j to I_i within the subgroup C_k, a warping technique [18], [19] is adopted and the warped saliency map U_i^j can be obtained by $U_i^j(p) = M_j(p')$, where p and p' are the matched pixels in advance. For each image, by fusing these warped saliency maps with its own saliency map, the geometric mean saliency can be obtained as follows.

$$G_i(p) = \sqrt[|C_k|]{M_i(p) \prod_{\substack{j \in C_k \\ j \neq i}} U_i^j(p)} \tag{3}$$

Finally, based on the geometric mean saliency, the GrabCut algorithm [20] is performed to get segmentation results.

2) GSP

In order to solve the expensive calculation cost of GMS, Jerripothula *et al.* [21] performed co-segmentation by group saliency propagation (GSP). This method can be seen as an improvement of GMS. Its main idea is to select a key image to represent the entire group to reduce the number of information transfer between images. See Figure 2 for an example.

First, the similar preprocessing is carried out to ensure that highly salient regions can adequately cover the objects to be segmented, such as adding spatial contrast saliency and brightening the saliency maps. Then all images are clustered into K groups, which are represented by the images closest to the cluster centers, respectively.

For each group, by aligning saliency maps of other images into the key image, its group saliency map can be formed. After that, the saliency map of the key image is merged into the warped saliency maps of other images by geometric mean method, respectively.

As Figure 3 shows, instead of pairwise matching, this method only matches each of the other images with the key image. In GMS, for n images, the related Dense-SIFT computation of other $n-1$ images is needed for each image segmentation, resulting in $n \times (n-1)$ calculation. However, in GSP, the Dense-SIFT computation for the key image is performed $n-1$ times and other images only need to warp the group saliency map back, resulting in $2(n-1)$ calculation.

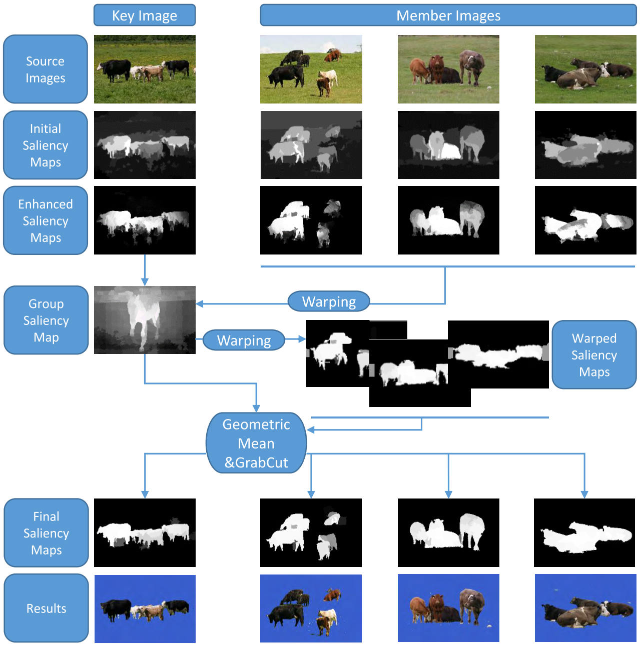

FIGURE 2. An example for GSP. A key image is selected to represent the entire group to reduce the number of information transfer between images.

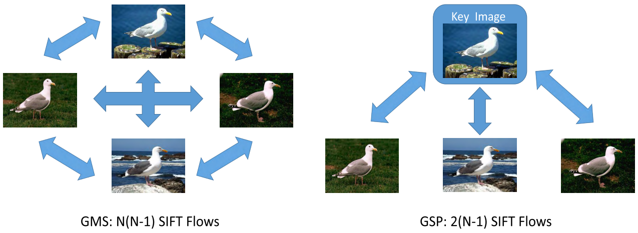

FIGURE 3. Comparison between GMS and GSP. GMS adopts pairwise matching, but GSP only matches each of the other images with the key image.

In particular, for co-segmenting a new image quickly, the image is assigned into a group on the basis of GIST. The group saliency map is warped for this image and then fused into the saliency map of this image. Moreover, for the object prior, the saliency map of the image is converted as follows.

$$O(p) = ((U_I^{C_k}(p))^{|C_k|} \times M(p))^{\frac{1}{|C_k|+1}} \tag{4}$$

where $U_I^{C_k}$ represents the k-th warped group saliency map with regard to the image I.

Finally, the segmentation results are also obtained by GrabCut, whose foreground and background seed locations are decided by

$$p \in \begin{cases} \text{Foreground} & \text{if } O(p) > \tau \\ \text{Background} & \text{if } O(p) < \phi \end{cases} \tag{5}$$

where the global threshold value ϕ can be automatically decided by the Otsu's method and τ is an adjustable parameter.

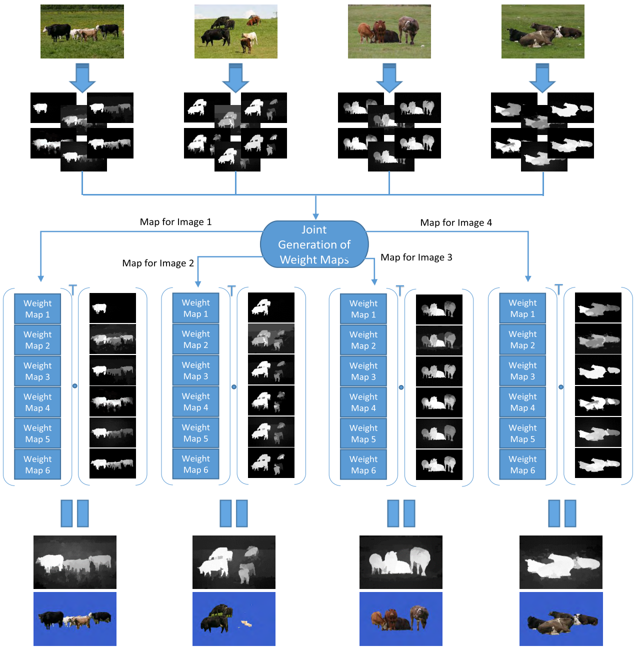

FIGURE 4. An example for SCF. The enhanced saliency maps are generated by fusing diverse saliency maps via weighted summation to exploit the inter-image information.

3) SCF

Based on their previous work, Jerripothula *et al.* [22] held that the single saliency extraction methods had limitations of their own. However, fusing multiple saliency extraction methods was able to overcome their respective defects. Hence, they proposed a method based on saliency co-fusion (SCF), whose objectives include suppressing the saliency of background and boosting the saliency of foreground. This method generates an enhanced saliency map by fusing diverse saliency maps via weighted summation and these weights can be optimized through exploiting the inter-image information. See Figure 4 for an example.

As mentioned above, the method's most important step is to obtain a saliency co-fusion map, and the most important step of obtaining the map is to find out the optimal weight for each of various saliency maps. These saliency maps can be obtained by diverse saliency extraction methods. Thus, the saliency co-fusion approach can be regarded as a weight selection problem. In detail, it hopes that elements with higher confidence can be assigned higher weights and neighboring elements have certain consistency. Moreover, the saliency co-fusion map values have to occur in the range [0, 1].

Given an image group $\mathcal{I} = \{I^1, I^2, \cdots, I^N\}$, $\mathcal{B}^n = \{B_1^n, B_2^n, \cdots, B_M^n\}$ denotes the set of M saliency maps for image I^n acquired by different saliency extraction approaches, and $\mathcal{P} = \left\{P_1^n, P_2^n, \cdots, P_{|\mathcal{P}^n|}^n\right\}$ denotes the superpixel set in image I^n acquired by [23]. $z(n, k, m)$ is the weight related to element $e(n, k, m)$ which belongs to image I^n, superpixel P_k^n and saliency map B_m^n. In addition, all weights will be stacked into the vector $z = [z_1, z_2, \cdots, z_{N_e}]^t$ and $N_e = \sum_{n=1}^{N} M\,|\mathcal{P}^n|$. Eventually, the task can be conceived as following quadratic programming problem.

$$\begin{aligned} &\min_z \; D^t z + \lambda z^t G z \\ &s.t.\; 0 \leqslant z_u \leqslant 1, \quad \forall u \in [1, N_e], \\ &\sum_{m=1}^{M} z(n, k, m) = 1, \quad \forall I^n \in \mathcal{I},\; P_k^n \in \mathcal{P} \end{aligned} \tag{6}$$

where λ trades off two terms as a balancing parameter. The first term actualizes global commonness and co-saliency as a prior term, where the coefficient vector $D \in \mathbb{R}^{N_e \times 1}$. The second term inspires neighborhood elements to seize similar weights as a pairwise smoothness term, where the coefficient matrix $G \in \mathbb{R}^{N_e \times N_e}$. After z is determined via minimizing, the saliency co-fusion map $\mathcal{J}^n$ can be compute as

$$\mathcal{J}^n(p) = \sum_{m=1}^{M} z(n, k, m) \times B_m^n(p) \tag{7}$$

where pixel $p \in P_k^n$.

Once the saliency co-fusion map becomes available, many single-image segmentation methods can be used for segmentation. Specifically, the authors adopted the classical Otsu's method [15] and an improved GrabCut algorithm [20], [24].

B. JOINT-OPTIMIZATION-BASED MODEL

There are links between many visual tasks, which can provide each other with effective information, and even be jointly optimized. The idea of joint processing has begun to show its merits over the individual processing. Next, two methods based on joint optimization will be introduced, which respectively resort to the optimization process of sketch and skeletonization.

1) CST

Sketch is of value for visual tasks such as image retrieval [25], action representation [26] and face synthesis [27]. Also it can provide effective information for segmentation.

Dai *et al.* [28] set forth a method associated with what they called 'co-sketch' (CST), which employs explicit models for sketchable patterns, like [29], [30]. The co-sketch aims to learn deformable shape templates which are shared by images, and to sketch such images by these templates. The shape templates can catch distinct image patterns and each of them is related with a segmentation template. The sketch help establish correspondence among images and the related segmentation templates supply critical bottom-up information for sketch. See Figure 5 for an example.

The learned model is composed of sketch model, region model and coupling model.

The sketch model attempts to encode the sketchable patterns via shape templates. The sketchable patterns consist of region boundaries, non-boundary edges and lines. Each of shape templates is described via an active basis model [31].

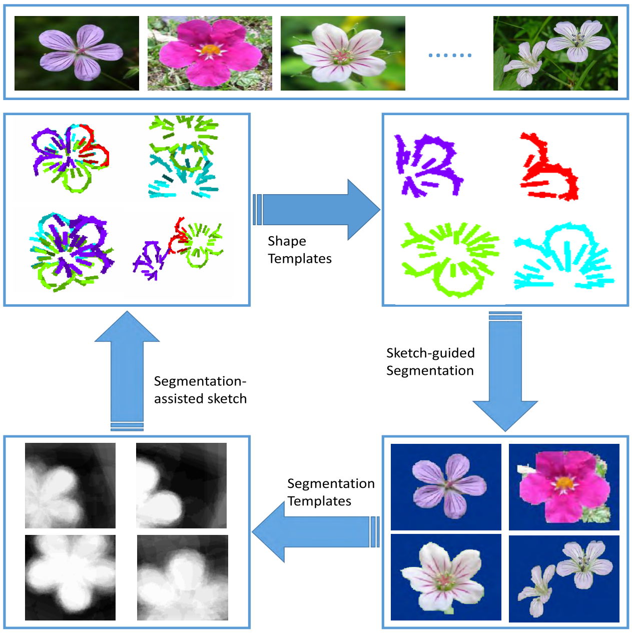

FIGURE 5. An example for CST. The sketch renders top-down information for segmentation and the related segmentation templates supply bottom-up information for sketch.

After a series of calculations and transformations, the energy function of sketch model can be defined as

$$\mathcal{E}(I_m|W_m^S, \Theta_S) = -l(I_m|W_m^S) \tag{8}$$

where I_m denotes the m-th image and W_m^S denotes the sketch representation of I_m. Θ_S represents the parameter of sketch model and $l(I|W)$ is the template matching score.

The region model attempts to encode the non-sketchable patterns via marginal distributions and pairwise similarities. The non-sketchable patterns include region interiors and shapeless patterns. After a series of calculations and transformations, the energy function of region model can be determined as

$$\begin{aligned}\mathcal{E}(I_m|W_m^R, \Theta_R) &= \sum_x \phi_1(I_m(x)|\delta_m(x)) \\ &\quad + \sum_{x\sim y} \phi_2(I_m(x), I_m(y)|\delta_m(x), \delta_m(y))\end{aligned} \tag{9}$$

where the first term is the unary potential and the second term is the pairwise potential. W_m^R denotes the region representation of I_m and Θ_R represents the parameter of region model. $I_m(x)$ is a vector in the color space and $\delta_m(x)$ is the label of pixel x for segmentation.

The coupling model attempts to associate shape templates with segmentation templates. As the probability maps, the segmentation templates offer pixel labels the top-down prior information. On the contrary, the pixel labels as data provide sketch representation the bottom-up information. After a series of calculations and transformations, the energy function of coupling model can be defined as

$$\mathcal{E}(W_m^R|W_m^S, \Theta_C) = -\sum_{k=1}^{K} \sum_{x\in\mathcal{D}_{X_{m,k}}^{(t_{m,k})}} \log P^{(t_{m,k})}(x - X_{m,k}, \delta_m(x)) \tag{10}$$

where Θ_C, $\mathcal{D}$ and P represent the segmentation templates, bounding boxes and probability maps, respectively.

Finally, the combined energy function can be drawn as

$$\begin{aligned}\mathcal{E}(I_m, W_m|\Theta) = \gamma\mathcal{E}(I_m|W_m^S, \Theta_S) + \mathcal{E}(I_m|W_m^R, \Theta_R) \\ + \mathcal{E}(W_m^R|W_m^S, \Theta_C)\end{aligned} \tag{11}$$

where γ is a weighting parameter to balance terms and $\mathcal{E}(I_m, W_m|\Theta)$ can define a joint probability by the Gibbs distribution.

In order to fit model via energy minimization, a relaxation algorithm is presented to alternate the following two steps.

(I) Image parsing. This step consists of sketch-guided segmentation and segmentation-assisted sketch.

(I.1) Sketch-guided segmentation. This substep segments images with the associated segmentation templates while the current sketches of images are given by the shape templates. That is, it goes to minimize $\mathcal{E}(I_m|W_m^R, \Theta_R) + \mathcal{E}(W_m^R|W_m^S, \Theta_C)$.

(I.2) Segmentation-assisted sketch. This substep sketches images by matching the shape templates while the current pixel labels of images are given by the segmentation templates. That is, it goes to minimize $\gamma\mathcal{E}(I_m|W_m^S, \Theta_S) + \mathcal{E}(W_m^R|W_m^S, \Theta_C)$.

(II) Re-learning. This step is similar to [32], [33] and re-learns the model parameters, shape templates and segmentation templates by the current sketches and segmentations. It can be further divided into three sub-steps: re-learn shape templates, re-learn marginal distributions of regions and re-learn segmentation templates.

2) CSZ

Similarly, skeletonization is of use for visual tasks such as shape matching [34], action recognition [35] and body pose recovery [36]. Also it can furnish efficacious intelligence for segmentation even carry out a joint optimization process.

Inspired by the scribble-supervised convolutional networks [37], Jerripothula *et al.* [38] set forth a method associated with what they called 'co-skeletonization' (CSZ) in the basis of [39]. The co-skeletonization aims to extract skeleton of common objects and can serve nice scribbles for segmentation. Alternately, the skeletonization also asks for nice segmentation. So a joint framework is proposed that they can inform and benefit each other. See Figure 6 for an example.

Given an image group $\mathcal{I} = \{I_1, \cdots, I_m\}$, $\mathcal{K} = \{K_1, \cdots, K_m\}$ and $\mathcal{O} = \{O_1, \cdots, O_m\}$ represent skeleton and segmentation, respectively, in which $K_i(p)$ indicates whether a pixel p belongs a skeleton pixel and $O_i(p)$ indicates whether a pixel p belongs a foreground pixel. Then, the overall

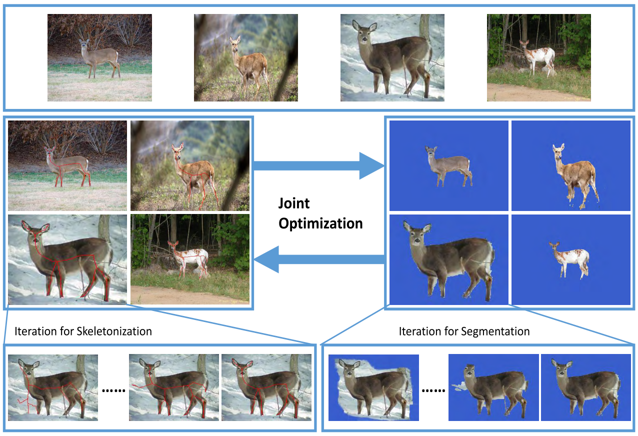

FIGURE 6. An example for CSZ. The skeletonization can serve nice scribbles for segmentation. Conversely, the nice segmentation is also asked for skeletonization.

Algorithm 1 CSZ Algorithm

```
Input: An image set $\mathcal{I}$ containing same category images
Output: Segmentation set $\mathcal{O}$
Initialization: let $O_i^{(0)}$ = Otsu thresholded saliency map
    and $K_i^{(0)} = \mathrm{ma}(O_i^{(0)})$
while $(\lambda\psi_{pr} + \psi_{in} + \psi_{sm})^{(t+1)} \leq (\lambda\psi_{pr} + \psi_{in} + \psi_{sm})^{(t)}$ do
    1)$\mathcal{O}^{(t)} \rightarrow \mathcal{O}$ and $\mathcal{K}^{(t)} \rightarrow \mathcal{K}$
    2)Attain $O_i^{(t+1)}$ through resolving (14) utilizing [20]
      with $\mathcal{O}^{(t)}$ and $K_i^{(t)}$.
    3)Attain $K_i^{(t+1)}$ through resolving (13) utilizing [39]
      with $\mathcal{K}^{(t)}$ and $O_i^{(t+1)}$, s.t. $K_i^{(t+1)} \in \mathrm{ma}(O_i^{(t+1)})$
end while
return segmentation set $\mathcal{O}$
```

objective function can be defined as

$$\begin{aligned} \min_{K_i,O_i}\ & \lambda\psi_{pr}(K_i,O_i|\mathcal{N}_i) \\ & + \psi_{in}(K_i,O_i|I_i) + \psi_{sm}(K_i,O_i|I_i) \\ s.t.\ & K_i \subseteq \mathrm{ma}(O_i) \end{aligned} \tag{12}$$

where the first term is a prior term from neighbor images, the second term is an interdependence term between skeleton and segmentation, and the third term is a smoothness term for smoothness. λ is a balance parameter and the constraint implies that the skeleton has to be a subclass of medial axis (ma) [40] of the shape templates.

In order to solve Equation (12), a classic selective optimization strategy can be adopted to divide it into two sub-problems then resolve them by turns. In detail, one sub-problem can be described as giving the shape O_i to resolve co-skeletonization through

$$\begin{aligned} \min_{K_i}\ & \lambda\psi_{pr}^k(K_i|\mathcal{N}_i) + \psi_{in}^k(K_i|O_i) + \psi_{sm}^k(K_i) \\ s.t.\ & K_i \subseteq ma(O_i) \end{aligned} \tag{13}$$

And another sub-problem can be described as giving the skeleton K_i to resolve co-segmentation through

$$\min_{O_i} \lambda\psi_{pr}^o(O_i|\mathcal{N}_i) + \psi_{in}^k(O_i|K_i, I_i) + \psi_{sm}^o(O_i|I_i) \tag{14}$$

For solving these two sub-problems iteratively, a good initialization is required. Hence the $\mathcal{O}$ and $\mathcal{K}$ are both considered to initialize with the help of Otsu saliency maps and medial axis mask [40], respectively. The specific process can be summarized by Algorithm 1.

C. GRAPH-BASED MODEL

At the pixel or superpixel level, the image is naturally graph-structured because the semantic features are more similar between adjacent pixels or superpixels. Analogously, at the object level, if each image is divided into multiple segments and constructed as a digraph, it is clear that there is greater similarity between the better segments. Next, two graph-based methods will be introduced, which are at the superpixel level and the object level, respectively.

1) MRW

Random walk is a common way in image processing, and some segmentation and co-segmentation methods are also involved, such as [41], [42] and [43]. The basic thought of the random-walker-based methods is as below: as agents walk and exclude each other, eventually each agent has a stationary distribution.

Lee *et al.* [44] held that a random walk model can be applied to exploiting the underlying information via graph structure and its properties can be settled via the optimization tools efficiently, which are quantifiable algebraically via the graph theory. Nonetheless, the conventional random walk method adopts a single agent to simulate the movements, which is inadequate to segment real images reliably. Hence, a model based on the multiple random walkers (MRW) was put forward to describe the walkers' traversal on a graph synchronously, in accordance with a transfer probability matrix. And these walkers can implement the interaction with others to accomplish an expectation. As the random walk progresses, each walker repels others and forms their own ruling zones. Eventually, the balance can be satisfied among the walkers, whose distributions are determined.

For each image, a weighted and undirected graph is constructed independently of the other images. The nodes consist of SLIC super-pixels [23] and the edge connection scheme abides by [45]. Specifically, each node is not only linked to its neighbours, but also connected to the adjacent nodes of its neighbours, and whole boundary nodes are linked to one another. Moreover, the weight w_{ij} of edge e_{ij} can be described as

$$w_{ij} = \begin{cases} exp(-\dfrac{d^2(x_i,x_j)}{\sigma^2}) & \text{if } e_{ij} \in E \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where σ^2 is a parameter and $d(x_i, x_j) = \sum_l \lambda_l d_l(x_i, x_j)$ is employed as the dissimilarity function, where d_l represents five dissimilarities of node features such as bag-of-visual-words histograms of LAB and RGB colors [46], LAB and RGB super-pixel means, boundary cues, and λ_l is the weights to average those dissimilarities. Normalizing $W = [w_{ij}]$, $A = [a_{ij}]$ is computed according to the transition probability $a_{ij} = w_{ij} / \sum_k w_{kj}$.

Additionally, two random walkers, foreground walker and background walker, are employed respectively in I_u for bilayer segmentation, probability distributions of which can be indicated with the help of $p_{f(u)}$ and $p_{b(u)}$. And they interact in the light of

$$\begin{aligned} p_{f(u)}^{(t+1)} &= (1-\epsilon)Ap_{f(u)}^{(t)} + \epsilon r_{f(u)}^{t} \\ p_{b(u)}^{(t+1)} &= (1-\epsilon)Ap_{b(u)}^{(t)} + \epsilon r_{b(u)}^{t} \end{aligned} \tag{16}$$

To utilize the relevance among images, the concurrence distribution of the foreground walker is computed, which denotes the resemblance of each node in image I_u to foreground in the other images. Similar to that previously described, the transfer matrix A_{uv} from I_v to I_u can be obtained through normalizing W_{uv}, whose (i, j)-th elements indicates the affinity from node j in image I_v to node i in image I_u. And it can transfer the foreground distribution $p_{f(v)}$ in I_v to I_u. Hence $A_{uv}p_{f(v)}$ is employed as the restart distribution r. Over integrating the inter-image estimation of all images, then the concurrence distribution of the foreground walker in I_u can be obtained by

$$c_{f(u)} = \frac{1}{Z} S_u \sum_v A_{uv} p_{f(v)} \tag{17}$$

where $S_u = \epsilon(I - (1-\epsilon)A_u)^{-1}$ and Z is the number of input images. And the concurrence distribution $c_{b(u)}$ of the background walker can be obtained in the same way.

To exploit jointly both the intra and inter information, a MRW clustering process is performed to refine $p_{f(u)}$ and $p_{b(u)}$ by $c_{f(u)}$ and $c_{b(u)}$, and the hybrid restart rule can be defined as

$$\phi_{f(u)} = \gamma\alpha Q_{f(u)} p_{f(u)} + (1-\gamma)c_{f(u)} \tag{18}$$

where the elements of diagonal matrix $Q_{f(u)}$ are the posterior probabilities of the foreground walker. α is a normalizing parameter and γ is a balance parameter. And the hybrid restart rule $\phi_{b(u)}$ for the background walker can be defined in the same way.

Finally, the stationary distributions $\pi_{f(u)}$ and $\pi_{b(u)}$ will be obtained by performing the iterative MRW process, which can be summarized by Algorithm 2.

2) SPA

A lot of image tasks tend to rely on graph structure owing to the intra- and inter-image association such as saliency detection [47]. For object co-segmentation, the graph structure is also used in pixel-level semantics such as [48] and [49], as well as superpixel-level semantics such as [50] and [51].

Algorithm 2 MRW Algorithm

```
Input: An image set $\mathcal{I}$
Output: Segmentation maps $\mathcal{C}$
Initialize $\mathcal{P}_{(u)} = \{p_{f(u)}, p_{b(u)}\}$ for each $I_u$
repeat for each image $I_u$
    Inter-image concurrence computation
    Intra-image MRW clustering
    Foreground extraction $\mathcal{C} = \{C_1, \cdots, C_z\}$
    Compute the foreground distance $\sum_{u,v} d_f(C_u, C_v)$
until the foreground distance stops decreasing
Pixel-level refinement
return segmentation maps $\mathcal{C} = \{C_1, \cdots, C_z\}$
```

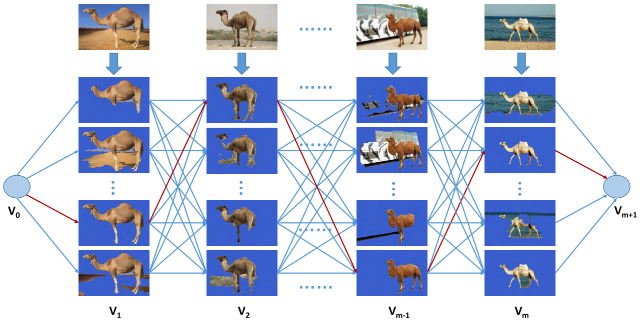

FIGURE 7. An example for SPA. Multiple object-like regions are utilized to construct a graph and then their similarity are measured by shortest path algorithm.

These methods utilize graph structure in a similar way to MRW. However, the graph structure based on object-level semantics is rarely employed for co-segmentation, one of which will be introduced.

Meng *et al.* [52] held that color bottom feature is hard to cope with common objects of different colors. Thus, they proposed a co-segmentation model based on salient spectral information and shortest path algorithm (SPA), which measures the middle-level semantic similarity of the object-related regions so that it can segment common objects under more middle-level semantic features.

First, each original image is segmented into a set of object proposals by using over-segmentation, saliency detection and object detection methods. Then the directed graph is designed to describe the local region similarities in accordance with feature distance and saliency map. And the saliency map can be improved via co-saliency strategy. Finally, the co-segmentation process can be transformed into the problem of selecting a set of nodes with maximum sum of weights, that is, the shortest path problem of the directed graph, which can be quickly solved through dynamic programming. See Figure 7 for an example.

The method's first step is to generate multiple local regions by original image and the set of local regions R consists of three subsets, that is, $R = \{R_1, R_2, R_3\}$. R_1 represents an unsupervised over-segmented local region subset to consider marginal information, which is obtained by [53].

R_2 is a subset of locally salient regions based on salient object detection [14] for introducing salient information to locate the object-oriented region. For R_3, its elements are formed from the possible object regions with the help of object detection method [54].

The method's second step is to construct a directed graph based on the set of local regions R. Given an image group $I = \{I_i, \cdots, I_m\}$, $p_{ij}(j = 1, \cdots, n_i)$ indicates the j-th local regions of image I_i, where n_i is the number of local regions generated by image I_i. In the node generation, node v_{ij} is generated for local region p_{ij} and makes up a node set $V = \left\{v_{ij} | i = 1, \cdots, m, j = 1, \cdots, n_i\right\}$. Then, V is divided into $V = \{V_1, \cdots, V_m\}$ in accordance with image I_i, where $V_i = \left\{v_{ij} | j = 1, \cdots, n_i\right\}$. Furthermore, $V_0 = \{v_{01}\}$ and $V_{m+1} = \left\{v_{(m+1)1}\right\}$ are added into the node set in order to facilitate the extraction of common objects. Next, for edge generation, any node pair v_{ij} and v_{kl} are connected as edge $e = (v_{ij}, v_{kl})$, where $k = i+1$. Last, a weight $w_{ij,kl}$ is assigned to each edge to describe the similarity between local regions. And the weight is computed by

$$w_{ij,kl} = w^1_{ij,kl} + \alpha \cdot w^2_{ij,kl} \tag{19}$$

where $w^1_{ij,kl}$ indicates the region similarity as region term and $w^2_{ij,kl}$ indicates the saliency values of the two nodes as saliency term to overcome the effect of changes in features between common objects. α is a balance parameter. The region term is computed by

$$w^1_{ij,kl} = d(f_{ij}, f_{kl}) \tag{20}$$

where f_{ij} and f_{kl} represent the features of the local regions p_{ij} and p_{kl}, respectively, by color histogram or shape descriptor [55]. $d(\cdot, \cdot)$ represents the distance between the two features. In particular, set $w^1_{0j,1l} = 1$ and $w^1_{mj,(m+1)l} = 1$. The saliency term is computed by

$$w^2_{ij,kl} = s_{ij} + s_{kl} \tag{21}$$

where s_{ij} and s_{kl} denote the saliency values of the local regions p_{ij} and p_{kl}, respectively, by co-saliency model [56], and the co-saliency model is improved.

Once the digraph is completely constructed, the final step can carry out. According to the characteristics of the constructed digraph, a time point t is assigned to each layer. Then, the shortest path problem belongs to a dynamic decision, which can be attributed to the typical problem of dynamic programming optimization. Therefore, dynamic programming is used to solve the problem. Additionally, the constructed digraph only considers the relationship of adjacent image layers, so this method can quickly construct and search for the shortest path.

D. OTHER MODEL

In addition to the above approaches, two other methods will be introduced, both of which attempt to capture objects' shape features. The former mainly depicts objects' contour by minimizing the energy function, while the latter determines complete objects by matching the local shape features.

1) AC

Different from other methods, active-contour-based methods attempt to segment objects by capturing the contour features of the objects, such as [57], [58] and [59].

Meng *et al.* [60] put forward a co-segmentation method which incorporates color reward strategy and active contour (AC) model. This method adopts a linear similarity measurement, which can avoid the shortcomings of the traditional reward strategy. It considers both the inter-image foreground consistency and the intra-image background consistency.

On the one hand, the foreground similarity is usually measured through penalizing the dissimilarities between foregrounds, such as [1]. However, the penalizing strategy will bring about NP-hard optimization problem. Rather than penalizing the dissimilarities, [3] makes a choice to reward the similarities by $\sum_l |h_1(b) \cdot h_2(b)|$. Moreover, the energy function constructed by the reward strategy is usually a convex function, such as the submodular function, so the model solution is simpler than the one by the penalty strategy.

On the other hand, the active-contour-based segmentation model describes the boundary of the segmented region by curve, and constructs an energy function that can reflect the characteristics of the segmented region for the curve. Hence, the objects' contours correspond to the smallest energy function value of all curves. Then, the object segmentation can be viewed as the optimization problem to obtain the curve with the minimum value of the energy function, which can be solved by the partial differential equation. According to the adopted curve features, the existing active contour methods can be divided into a boundary-based method and a region-based method. The boundary-based and region-based active contour methods define the energy function by the region edge gradient and the properties of the region, respectively. Compared to the former, the region-based active contour method is more robust to the initial contour setting.

For curve C_k, the energy function is formed as

$$\begin{aligned} E_k(C_k) = \mu \cdot Length(C_k) &+ \nu \cdot Area(\omega^i_k) \\ &- \lambda^i_k \int_{\omega^i_k} f[I_k(x,y), g(\omega^i_{1-k})]\, dxdy \\ &- \lambda^o_k \int_{\omega^o_k} f[I_k(x,y), g(\omega^o_k)]\, dxdy \end{aligned} \tag{22}$$

where I_k is the k-th image, C_k indicates the curve in I_k, and $E_k(C_k)$ denotes the energy function relevant to C_k. ω^i_k and ω^o_k describe the regions inside and outside the C_k, respectively. $g(\omega)$ represents the region ω and $f(p, g(\omega))$ is utilized to estimate the resemblance between pixel p and region ω. $Length(C_k)$ represents the length of C_k and $Area(\cdot)$ indicates the region area. The third term, called interior term, depicts the foreground similarity. The last term, named exterior term, describes the background consistency.

The model is based on the assumption that the backgrounds of the images are different and the initial contour contains most of the object region. For one foreground pixel $I_k(i,j)$ in image I_k, there is $f[I_k(i,j), g(\omega^i_{1-k})] > f[I_k(i,j), g(\omega^o_k)]$. Conversely, for a background pixel, there is $f[I_k(i,j), g(\omega^i_{1-k})] < f[I_k(i,j), g(\omega^o_k)]$. Therefore, any pixel-level label exchange between foreground and background will result in an increase of the energy value. In other words, the energy value is minimal only when the curve accurately segments the common object. Thus, for image I_k, the co-segmentation problem is expressed as

$$C_k^* = \arg\min_{C_k} E_k(C_k) \tag{23}$$

In the process of optimization, the level-set technique is first used to describe the contour, and the curve optimization can be transformed into the optimization of the level set function. Then, the Euler-Lagrange formula is employed to optimize the energy function described by the level set. After a series of treatments, the final model is expressed as

$$\begin{aligned}\phi_k^{n+1}(i,j) = \phi_k^n(i,j) + \delta(\phi_k^n(i,j)) \cdot [-\mu \cdot \kappa(i,j) - \nu \\ + \lambda_k^i \cdot f[I_k(i,j), g(\omega^i_{1-k})] \\ - \lambda_k^o \cdot f[I_k(i,j), g(\omega^o_k)]]\end{aligned} \tag{24}$$

This approach makes use of the dynamic mode to achieve the segmentation of common objects as Algorithm 3.

Algorithm 3 AC Algorithm

```
Input: An image set $\mathcal{I}$
Output: The required locations
Initialize: Curves $\phi_k^0$ and other parameters
repeat for each image $I_k$
    Calculate $g(\omega^i_{1-k})$ and $g(\omega^o_k)$
    Obtain $\phi_k^{n+1}$ by (24) over solving PDE
until convergence criterion has been satisfied
return the locations with $\phi_k^{n+1} > 0$
```

2) COMP

The composition-based idea is very useful so that some visual problems attempt to employ it, such as [61] and [62]. It depends on shape templates and local regions.

Faktor and Irani [63] proposed an approach by composition (COMP), which does not depend on any common and simple model, but the framework improved in [64] and [65]. It shows that non-trivial image parts induce statistically meaningful affinities among other image parts when they re-occur in other images. In addition, the method expects that a co-segment share large non-trivial regions with other co-segments, and can be well composable from other co-segments, yet not be easily composed of parts outside the co-segments. See Figure 8 for an example.

The first step of this method is to induce affinities between image parts. For the shared regions, the rarer and larger they are, the higher their affinities. And these regions can offer a rough localization of co-objects. In terms of [64], there is a definition about the affinity $\mathcal{A}$ of shared region R between images I_1 and I_2 as

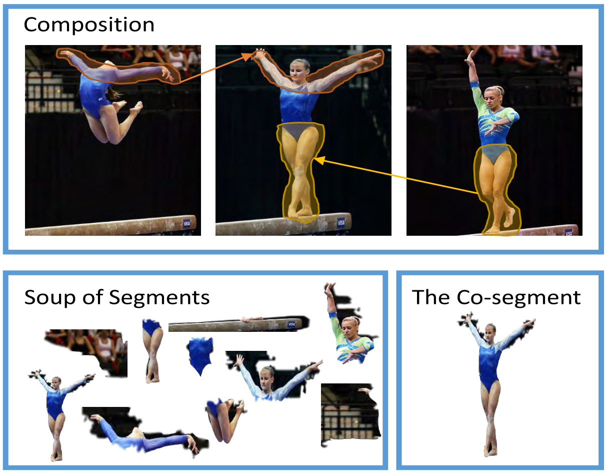

FIGURE 8. An example for COMP. The non-trivial image parts induce statistically meaningful affinities among other image parts when they re-occur in other images.

$$\mathcal{A}(R|I_1, I_2) = log \frac{p(R|I_1, I_2)}{p(R|H_0)} \tag{25}$$

where $p(R|I_1, I_2)$ represents the similarity of the regions found in I_1 and I_2. $p(R|H_0)$ denotes the likelihood of the region to occur by a random process H_0. If a region matches well but is trivial, it will still induce a low affinity. With a series of approximations, the affinity $\mathcal{A}$ can be simply expressed as

$$\mathcal{A}(R|I_1, I_2) = \sum_{d_i \in R} |\Delta d_i(H_0)|^2 - |\Delta d_i(I_1, I_2)|^2 \tag{26}$$

where the densely sampled descriptors d_i describes the region $R \subset I_1$, whose likelihood is generated from I_2. $\Delta d_i(I_1, I_2)$ denotes the l_2 distance between d_i, called matching error. the random process H_0 is approximated via a descriptor codebook $\widehat{D}$ generated by K-Means clustering to all descriptors. $\Delta d_i(H_0)$ represents the l_2 distance between d_i and its nearest descriptor in $\widehat{D}$, called error of descriptor with regard to codebook. In other words, the induced affinity parallels to the total matching error and descriptor error. The image parts with high affinities tend to coincide with unique parts of the co-objects and good for co-segments.

Generally speaking, detection of large non-trivial shared regions is very hard and the size and shape of regions may be arbitrary. Therefore, a randomized search algorithm [65] was proposed and it can make sure the efficient detection with high reliability. Usually, regions can be depicted by HOG descriptors or other densely sampled descriptors. The region matching algorithm is equal to an extension about 'PatchMatch' [66]. It drives each descriptor to choose the best match in another image and then propagate its match to neighbor. Only if the propagated match is better, its neighbor will replace current match. Furthermore, in order to solve the quadratically growing complexity as a result of searching shared regions between all image pairs, a way of image collaboration is adopted to maintain linear complexity.

The second step is to make use of the detected regions and their induced affinities to seed co-segments as well as estimate the co-segment likelihood of each pixel. The detected regions afford a rough assessment about the location of co-objects, although they only cover parts of co-segments and may across boundaries. In other words, they cannot structure good segments on their own. Hence, a 'soup of segments' can be exploited to refine the co-segments and the 'soup of segments' consists of multiple overlapping segment candidates. The specific process is as follows.

1. Extract a 'soup of segments' $\{S_l\}$ from each image by employing the hierarchal segmentation of [67].

2. Calculate the co-segment score for each segment S_l via induced affinity as

$$Score(S_l) = \frac{1}{|S_l|} \sum_{m} \mathcal{A}(R_m|I, I_{\chi(m)}) \tag{27}$$

where $\{R_m\}$ represents the detected shared regions with high intersection with segment S_l. $\chi(m)$ denotes the index of image in which R_m searches for its region match. $|S_l|$ is the size of segment to normalize region contributions.

3. Find out the K segments $\{S_k\}$ with the highest scores for each pixel p and assess their co-segmentation likelihood as follows.

$$CSL(p) = \frac{1}{K} \sum_{k} Score(S_k) \tag{28}$$

4. Normalize the likelihood map of entire image and let its all values be in [0, 1].

The final step is to collaborate and share information with each other to improve the quality of co-segmentation, including consensus scoring and acquisition of the final binary co-segmentation maps. Let p be a pixel in image, $p_1, \cdots, p_M \in Neighborhood(p)$ and $q_1, \cdots, q_{M'}$ represent corresponding pixels to p in all other images. Then, the co-segmentation likelihood will be updated along with iterations as follows.

$$\log CSL^{(t+1)}(p) = \frac{1}{2M} \cdot \sum_{i=1}^{M} \log CSL^{(t)}(p_i) + \frac{1}{2M'} \cdot \sum_{j=1}^{M'} \log CSL^{(t)}(q_j) \tag{29}$$

By performing several iterations of consensus re-scoring, the true co-segments may be revealed. And they can be obtained by using Grab-cut [20] or modified Grab-cut [68].

III. EXPERIMENT

In an attempt to observe the actual effects of these co-segmentation methods and make an objective assessment, an experiment will be introduced.

A. DATASET AND BENCHMARK

For the sake of a fair comparison, some publicly available datasets will be adopted by the experiment because they are widely used by a variety of co-segmentation methods to evaluate performance.

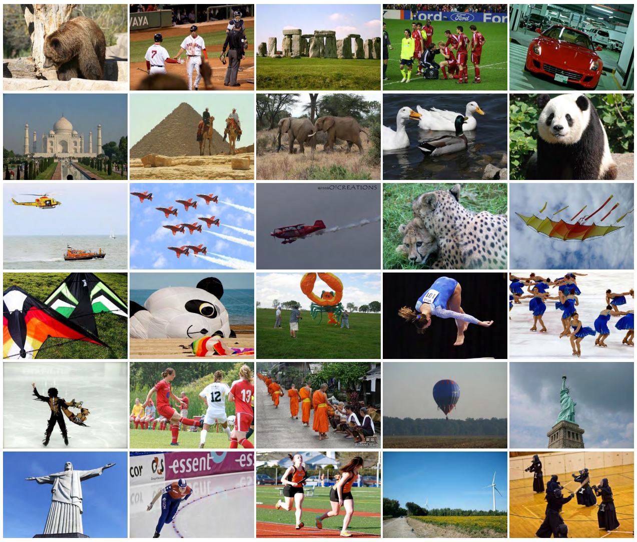

FIGURE 9. Samples on CMU-Cornell iCoseg dataset.

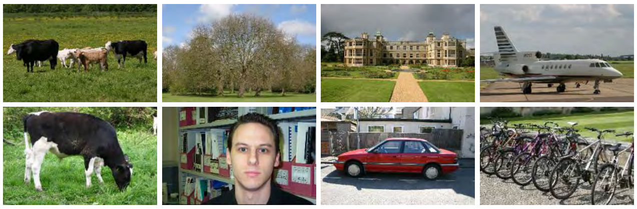

FIGURE 10. Samples on MSRC dataset.

iCoseg dataset [69] was first proposed for co-segmentation and it is still widely employed by the majority of co-segmentation methods. It contains 38 object classes (643 images) such as kites, hot balloons, animals and sport players (see Figure 9), which is full of challenge due to varying locations, appearances and cluttered backgrounds.

MSRC dataset [70] was first proposed for recognition and segmentation but its modified version is frequently adopted by salient object detection and co-segmentation. It contains 8 object categories (233 images) such as cows, trees, cars and faces (see Figure 10).

Coseg-Rep dataset [28] was later proposed for co-segmentation. It contains 23 groups (572 images) such as a variety of flowers and animals (see Figure 11). Additionally, there is a special group called 'repetitive' among them. In this group, each of images exists similar shape patterns repeating themselves, respectively, such as tree leaves.

B. EVALUATION METRICS

For the objectivity and comprehensiveness of experiment, a variety of evaluation metrics are employed from the most influential conferences and journals such as [71], [72]. Next they will be introduced in brief.

Suppose $S = \{S_1, S_2, \cdots, S_M\}$ is the segmented image where S_i is the ith segment and $G = \{G_1, G_2, \cdots, G_N\}$

FIGURE 11. Samples on Coseg-Rep dataset.

is the ground truth where G_j is the jth partition (if it is a binary segmentation, $M = N = 2$). B_S and B_G indicate the boundaries of S and G, respectively. $|\cdot|$ denotes the number of image pixels and $n = \sum_{i=1}^{M}\sum_{j=1}^{N}|S_i \cap G_j|$ represents the total intersection between S and G. Let $\mathcal{P} = \{(p_i, p_j) \in I \times I | i < j\}$ be all pairs of pixels in the image, then divide $\mathcal{P}$ into four classes:

$\mathcal{P}_{00}$: in different regions both in S and G,
$\mathcal{P}_{01}$: in the same region in G but different in S,
$\mathcal{P}_{10}$: in the same region in S but different in G,
$\mathcal{P}_{11}$: in the same region both in S and G.

According to the relative overlap of regions, each region can be categorized into object candidates, part candidates or fragmentation candidates. Let oc, fc and pc be the number of object candidates, part candidates and fragmentation candidates in S, respectively. Similarly, oc', fc' and pc' represent the counterparts in G. Specifically, if $O_S^{ij} > \gamma_o$ and $O_G^{ij} > \gamma_o$, S_i and G_j are both classified into object candidates. If $O_S^{ij} > \gamma_p$ and $O_G^{ij} > \gamma_o$, S_i is sorted into fragmentation candidate while G_j is viewed as part candidate. If $O_S^{ij} > \gamma_o$ and $O_G^{ij} > \gamma_p$, S_i is sorted into part candidate while G_j is viewed as fragmentation candidate. γ_o is an object threshold and γ_p is a part threshold. They are set to 0.95 and 0.25 in [72], respectively. Further, the O_S^{ij} and O_G^{ij} can be defined as

$$O_S^{ij} = \frac{|S_i \cap G_j|}{|S_i|}, \quad O_G^{ij} = \frac{|S_i \cap G_j|}{|G_j|} \tag{30}$$

1) INTERSECTION-OVER-UNION

The intersection-over-union (IoU) is widely adopted to measure image segmentation problem, which is defined in various literatures [73]. It computes the similarity between the segmented region and the ground-truth region (the higher, the better), which can be defined as

$$\text{IoU} = \frac{\text{Segmentation} \cap \text{Ground truth}}{\text{Segmentation} \cup \text{Ground truth}} \tag{31}$$

2) CONSISTENCY ERROR

In order to measure the differences between S and G from different perspectives, a range of consistency errors are employed (the lower, the better) and described as follows.

First, the error between S_i and G_j can be defined as

$$P_{ij} = \frac{|S_i \setminus G_j|}{|S_i|} \times |S_i \cap G_j| = (1 - \frac{|S_i \cap G_j|}{|S_i|}) \times |S_i \cap G_j| \tag{32}$$

and the error between G_j and S_i can be defined as

$$Q_{ij} = \frac{|G_j \setminus S_i|}{|G_j|} \times |S_i \cap G_j| = (1 - \frac{|S_i \cap G_j|}{|G_j|}) \times |S_i \cap G_j| \tag{33}$$

Further the global consistency error (GCE) can be defined as

$$\text{GCE}(S, G) = \frac{1}{n} \min \left\{ \sum_{i=1}^{M}\sum_{j=1}^{N} P_{ij}, \sum_{i=1}^{M}\sum_{j=1}^{N} Q_{ij} \right\} \tag{34}$$

The local consistency error (LCE) can be defined as

$$\text{LCE}(S, G) = \frac{1}{n} \sum_{i=1}^{M}\sum_{j=1}^{N} \min(P_{ij}, Q_{ij}) \tag{35}$$

The bidirectional consistency error (BCE) can be defined as

$$\text{BCE}(S, G) = \frac{1}{n} \sum_{i=1}^{M}\sum_{j=1}^{N} \max(P_{ij}, Q_{ij}) \tag{36}$$

The object-level consistency error can be defined as

$$\text{OCE}(S, G) = \min(E_{s,g}, E_{g,s}) \tag{37}$$

where the partial error measure $E_{s,g}$ can be defined as

$$E_{s,g}(S, G) = \sum_{i=1}^{M} \left[1 - \sum_{j=1}^{N} \frac{|S_i \cap G_j|}{|S_i \cup G_j|} \times W_{ij} \right] W_i \tag{38}$$

where W_{ij} weighs each G_j related to all segments in S and W_i weighs each S_i related to all segments in G.

3) PRECISION AND RECALL

The precision and recall are able to reflect that a segmentation result is coarse (high precision, low recall) or fragmented (low precision, high recall) and usually summarized by F-score [74] (the higher, the better).

The precision and recall for boundaries can be defined as

$$P_b = \frac{|B_S \cap B_G|}{|B_S|}, \quad R_b = \frac{|B_S \cap B_G|}{|B_G|} \tag{39}$$

and summarized by F-score as

$$F_b = \frac{2P_b \cdot R_b}{P_b + R_b} \tag{40}$$

The precision and recall for regions can be defined as

$$P_r = \frac{|\mathcal{P}_{11}|}{|\mathcal{P}_{11}| + |\mathcal{P}_{10}|}, \quad R_r = \frac{|\mathcal{P}_{11}|}{|\mathcal{P}_{11}| + |\mathcal{P}_{01}|} \tag{41}$$

and summarized by F-score as

$$F_r = \frac{2P_r \cdot R_r}{P_r + R_r} \tag{42}$$

TABLE 1. Evaluation metrics for object co-segmentation.

Measure Representative	References	Notation
Precision-Recall for Boundaries	[75] [76]	P_b, R_b
Precision-Recall for Regions	[76]	P_r, R_r
Precision-Recall for Objects and Parts	[72]	P_{op}, R_{op}
Variation of Information	[77]	VoI
Probabilistic Rand Index	[78] [79]	PRI
Segmentation Covering	[80]	SC
Directional Hamming Distance	[81] [82]	D_H
Van Dongen Distance	[83]	d_{vD}
Bidirectional Consistency Error	[76]	BCE
Global Consistency Error	[84] [85]	GCE
Local Consistency Error	[84] [85]	LCE
Object-Level Consistency Error	[84] [86]	OCE

The precision and recall for objects and parts can be defined as

$$P_{op} = \frac{oc + fc + \beta pc}{|S|}, \quad R_{op} = \frac{oc' + fc' + \beta pc'}{|G|} \tag{43}$$

where β is set to 0.1 in [72] and summarized by F-score as

$$F_{op} = \frac{2P_{op} \cdot R_{op}}{P_{op} + R_{op}} \tag{44}$$

4) OTHER METRICS

Although IoU is widely applicable, there are many other evaluation metrics that are recognized and employed depending on different needs and perspectives. Some of them will be introduced and more details can be seen in Table 1.

The probabilistic rand index (the higher, the better) can be defined as

$$\mathrm{PRI}(S, G) = \frac{|\mathcal{P}_{00}| + |\mathcal{P}_{11}|}{|\mathcal{P}|} \tag{45}$$

The directional hamming distance (the lower, the better) can be defined as

$$D_H(S \Rightarrow G) = n - \sum_{G_j \in G} \max_{S_i \in S} |S_i \cap G_j| \tag{46}$$

The van dongen distance (the lower, the better) can be defined as

$$d_{vD}(S, G) = D_H(G \Rightarrow S) + D_H(S \Rightarrow G) \tag{47}$$

The segmentation covering (the higher, the better) can be defined as

$$\mathrm{SC}(S \to G) = \frac{1}{n} \sum_{G_j \in G} |G_j| \cdot \max_{S_i \in S} \frac{|S_i \cap G_j|}{|S_i \cup G_j|} \tag{48}$$

The variation of information (the lower, the better) can be defined as

$$\mathrm{VoI}(S, G) = H(S) + H(G) - 2I(S, G) \tag{49}$$

where the $H(S)$ is the entropy of a discrete random variable and $I(S, G)$ is the mutual information.

C. RESULT AND ANALYSIS

The actual effects of these co-segmentation methods can be seen through the final segmented images, some of which are shown in Figures 12, 13 and 14. In general, different algorithms perform differently for the same image group, and even far from each other. Similarly, for the different image groups, the performance of the same algorithm also shows a big gap.

As can be seen in Tables 2, 4 and 6, the IoU performance of most methods shows the ups and downs, which is more or less dependent on the adjustment in some parameters and characteristics of selected image sets. For example, in the iCoseg dataset, the worst IoU performance of AC, COMP and CST are merely 0.079, 0.158 and 0.114 respectively, while their best performance can reach 0.918, 0.959 and 0.863. Among these methods, SCF and CSZ have relatively stable IoU performance, MRW works well except on one of the collections, AC and COMP also have passable IoU performance, and others have good IoU performance on some images, while their overall IoU performance is not satisfactory. Especially, in the MSRC dataset, the best scores of these methods are not high because of ground truth with rough annotation and background similar to the foreground. See Table 8 for more details.

Tables 3, 5 and 7 show that, notwithstanding some methods may perform worse in IoU evaluation than others, they may be more effective on other metrics, which are proposed for other considerations introduced in the previous subsection, and vice versa. On the average, precision-recall for regions, segmentation covering and probabilistic rand index of each method all have good scores, while none of their consistency errors has good scores.

In general, from the experimental results, these methods have the following problems.

- Segmentation accuracy and boundary
  Some methods perform well in the metrics that measure the segmented regions, but not in the metrics that measure the boundaries of segmented regions. The regions are usually trivial and coarse due in part to the extra energy term hired to enforce inter-image consistency, which often gives rise to unsmooth segmentations.
- IoU metric and parameter regulation
  Even if IoU metric is only considered, its performance still has a lot of room for improvement, and the performance of most methods depends on the setting and adjustment of parameters.
- Generalization and feature selection
  These methods perform well in some image groups, but not in others, partly because the common features in different image groups are not the same, and these methods tend to adopt only one fixed feature so that they are not well self-adapted.
- Robustness and object-level semantics
  Faced with complex backgrounds, backgrounds similar to foregrounds, and too small objects, these methods

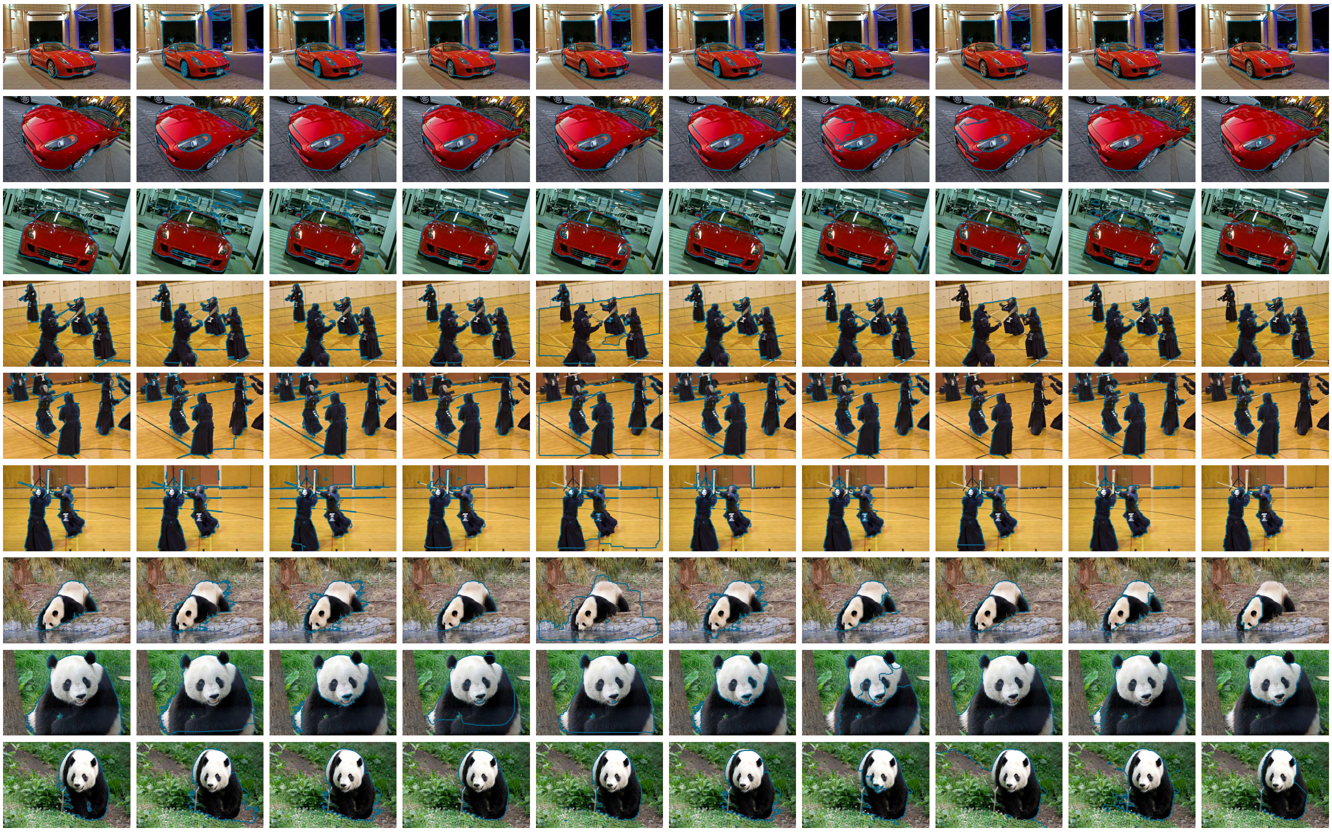

FIGURE 12. Segmentation cases on iCoseg dataset. The first column represents ground truth, the other columns represent GMS, GSP, SCF, CST, CSZ, MRW, SPA, AC and COMP from left to right. Some algorithms perform better on one image than others, but on another image may reverse.

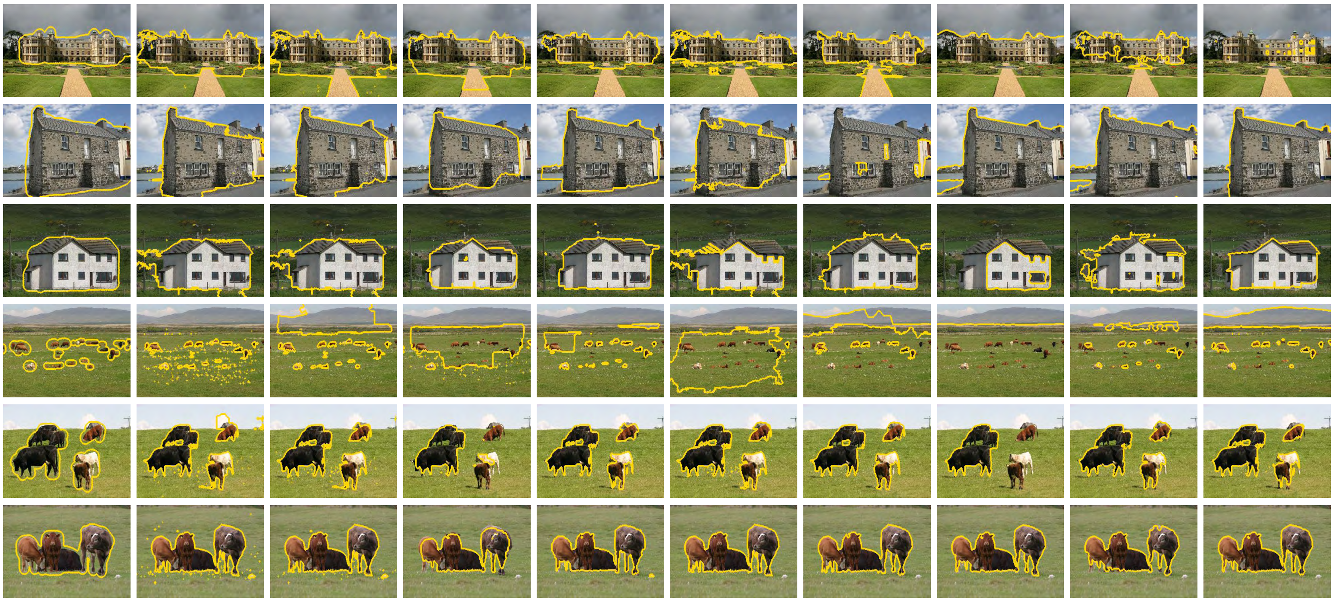

FIGURE 13. Segmentation cases on MSRC dataset. The first column represents ground truth, the other columns represent GMS, GSP, SCF, CST, CSZ, MRW, SPA, AC and COMP from left to right. Some algorithms perform better on one image than others, but on another image may reverse.

are not often good at distinguishing objects. Part of the reason is that these methods always use low-level features such as color, which do not capture the semantic information of objects well and are susceptible to noise.

IV. CHALLENGE AND DIRECTION

In recent years, due to the emergence of large-scale datasets and the combination of the advantages of unsupervised and supervised segmentation, co-segmentation has attracted more

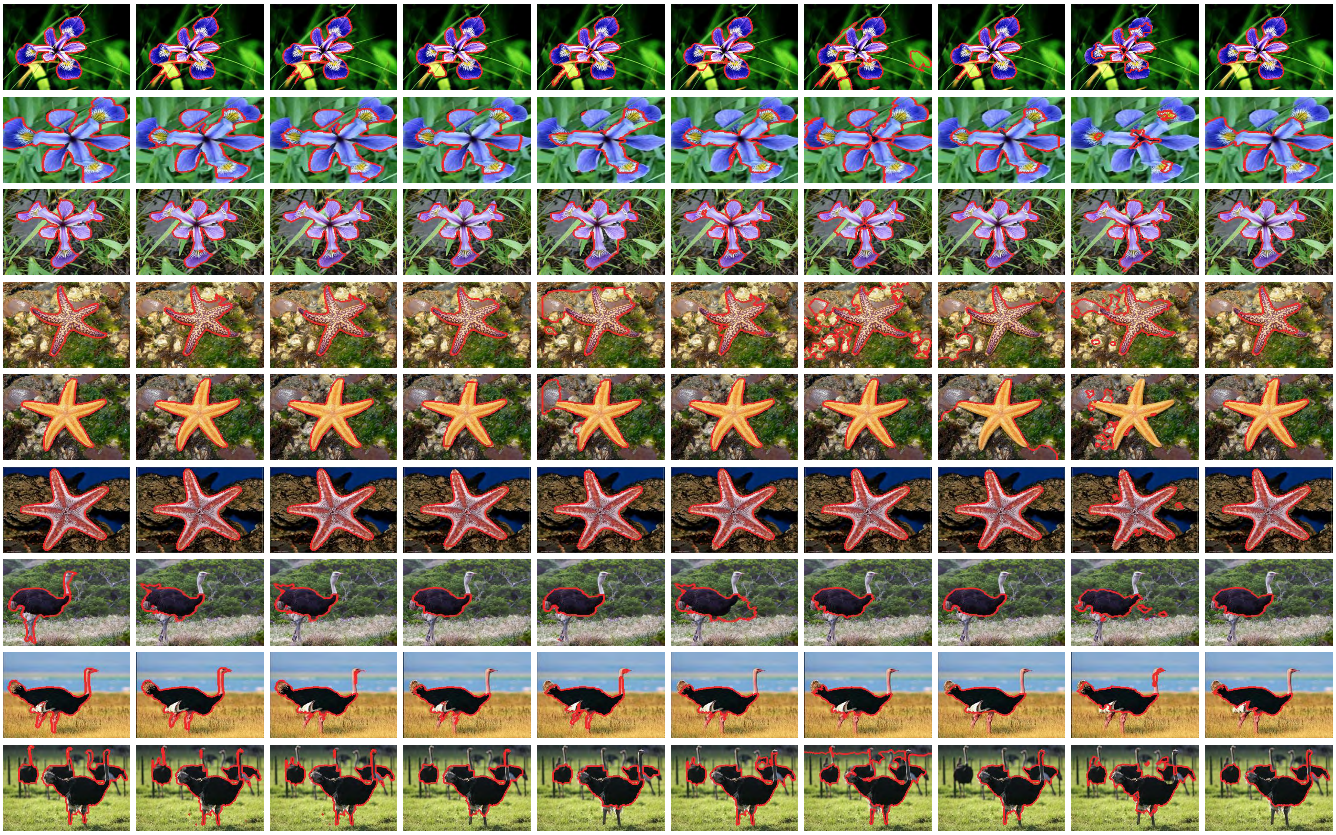

FIGURE 14. Segmentation cases on Coseg-Rep dataset. The first column represents ground truth, the other columns represent GMS, GSP, SCF, CST, CSZ, MRW, SPA, AC and COMP from left to right. Some algorithms perform better on one image than others, but on another image may reverse.

and more attention from researchers and made some progress, but still faces the following challenges.

- Theoretical analysis and model construction
  For the time being, it is difficult to find a general co-segmentation algorithm that can adapt to all scenarios and datasets. Therefore, how to reasonably analyze and construct an effective co-segmentation model is a basic and challenging problem in the field for the complexity and variability of the foreground and background presented by different scenarios and datasets.
- Consistency measurement of regional semantics
  Inter-regional consistency measurement has always been a key step in the segmentation problem. Existing co-segmentation models often use a fixed measure of consistency to extract common objects. However, the common features in different image groups are not the same, which may be color, shape, texture or others, and are usually not known in advance in practical applications. Therefore, how to self-adaptively identify common features and learn regional consistency measurement is also a challenge and direction in this field.
- Mining and utilizing higher-level semantic features
  Most co-segmentation models use low-level semantic features such as color, which are often difficult to distinguish when the background is too complex or even similar in some respects to the foreground. At this time, the more discriminative middle- and high-level semantic features can provide better foreground information and ensure the integrity of the object to a certain extent. Thus, how to reasonably and effectively mine and utilize the middle- and high-level semantic features has become another challenging problem.
- Model optimization and solution
  Object co-segmentation can be transformed into a solution process for the model. When dealing with large-size images and complex middle- and high-level semantic features, it often has great time overhead and memory loss. Hence, how to optimize the model to reduce the complexity and choose the appropriate solution method is also a major problem in the field.
- Discovery and segmentation of multiple foregrounds
  The multi-foreground co-segmentation model develops slowly, mainly because the ambiguity between the foreground and the background creates a major difficulty when an appropriate prior is not given. In addition, there are also differences in the number of object categories between images, making the inter-image consistency information more difficult to use. It often requires artificially setting the foreground number to adjust and faces

TABLE 2. The effectiveness and efficiency of the compared methods using the IoU metric on iCoseg dataset.

Class	#Image	GMS	GSP	SCF	CST	CSZ	MRW	SPA	AC	COMP
Brown Bear-1	19	0.6421	0.6665	0.6004	0.4284	0.6357	0.6658	0.5239	0.5344	0.6640
Brown Bear-2	5	0.7218	0.7520	0.7058	0.6326	0.7607	0.7622	0.4193	0.7790	0.8490
Red Sox Players	25	0.6411	0.6902	0.7246	0.1276	0.6773	0.7099	0.2865	0.2590	0.6993
Stonehenge-1	5	0.6245	0.5936	0.6354	0.4404	0.5924	0.7232	0.0303	0.6618	0.7375
Stonehenge-2	18	0.7633	0.7871	0.7183	0.8130	0.7416	0.7832	0.2159	0.8359	0.6794
Liverpool FC Players	33	0.4708	0.4581	0.5466	0.4366	0.5490	0.5292	0.4491	0.5542	0.4666
Ferrari	11	0.6555	0.6520	0.6691	0.6364	0.6376	0.7417	0.5273	0.6483	0.6697
Agra Taj Mahal-1	5	0.5446	0.5062	0.5278	0.4656	0.5585	0.7449	0.5674	0.4892	0.5737
Agra Taj Mahal-2	5	0.4284	0.4817	0.4792	0.3716	0.3592	0.4664	0.2716	0.4322	0.3734
Pyramids	10	0.7343	0.6342	0.5449	0.4061	0.5918	0.5849	0.2469	0.6051	0.6271
Elephants	15	0.6807	0.7241	0.6517	0.3087	0.7336	0.7034	0.4229	0.3825	0.5106
Goose	31	0.6681	0.6901	0.5231	0.3959	0.6962	0.7888	0.6157	0.8416	0.8419
Pandas-1	25	0.7367	0.7165	0.7122	0.7832	0.6219	0.6956	0.4283	0.6780	0.6952
Pandas-2	21	0.6412	0.6216	0.5749	0.6676	0.5064	0.7269	0.6974	0.5897	0.2779
Airshows-helicopter	12	0.7841	0.8034	0.7774	0.6908	0.8236	0.7960	0.8839	0.7980	0.7387
Airshows-planes	39	0.5049	0.5148	0.5634	0.5791	0.5049	0.5713	0.2968	0.2275	0.6089
Airshows-Huntsville	22	0.3915	0.4162	0.5237	0.2646	0.5172	0.7757	0.5476	0.4812	0.5816
Cheetah	33	0.7552	0.7647	0.7357	0.7719	0.7152	0.7984	0.5463	0.6303	0.2200
Kite-1	18	0.7378	0.6915	0.8172	0.2940	0.8142	0.8503	0.7758	0.8150	0.8911
Kite-2	10	0.5447	0.4645	0.5596	0.5347	0.5892	0.4687	0.2390	0.5927	0.6302
Kite-3	7	0.7625	0.7151	0.8322	0.5681	0.7647	0.8823	0.4869	0.8677	0.9591
Kite-4	11	0.6831	0.6923	0.7048	0.3859	0.7234	0.8502	0.7359	0.2126	0.6747
Gymnastics-1	6	0.8333	0.8396	0.8359	0.8693	0.8374	0.7852	0.7892	0.0979	0.5174
Gymnastics-2	4	0.6487	0.6659	0.7657	0.3937	0.6757	0.7685	0.8086	0.6387	0.7907
Gymnastics-3	6	0.6224	0.6203	0.8131	0.7512	0.7654	0.7581	0.8030	0.7255	0.7944
Skating-1	11	0.6759	0.7007	0.5889	0.5888	0.6751	0.8380	0.3554	0.7689	0.5928
Skating-2	12	0.8461	0.8113	0.8602	0.1177	0.9139	0.9292	0.8052	0.2271	0.8445
Skating-3	13	0.2233	0.3038	0.4509	0.1136	0.4158	0.1096	0.1740	0.0790	0.1584
Soccer Players-1	36	0.6552	0.6480	0.6680	0.6787	0.6619	0.6399	0.5813	0.6753	0.5462
Soccer Players-2	16	0.5040	0.5249	0.4684	0.4633	0.5062	0.5005	0.2847	0.5241	0.4154
Monks	17	0.7017	0.7200	0.7620	0.7048	0.7134	0.7627	0.5858	0.6992	0.5802
Hot Balloons	24	0.6549	0.6348	0.7516	0.5038	0.7421	0.8493	0.5536	0.6704	0.9378
Statue of Liberty	41	0.7892	0.7772	0.7439	0.6764	0.8234	0.7741	0.8426	0.7918	0.6997
Christ the Redeemer	13	0.7275	0.6968	0.7122	0.6397	0.7813	0.8388	0.6897	0.7316	0.7089
Track and Field	5	0.5066	0.5786	0.4897	0.3671	0.4193	0.5457	0.2839	0.4465	0.3271
Windmill	18	0.3145	0.3170	0.4462	0.1441	0.3386	0.2723	0.4044	0.2524	0.2521
Kendo-1	30	0.7835	0.8297	0.8219	0.1515	0.9022	0.9210	0.6885	0.9027	0.8051
Kendo-2	11	0.8725	0.9127	0.8483	0.4490	0.8922	0.9478	0.6732	0.9184	0.8749
Average	17	0.6441	0.6478	0.6620	0.4899	0.6626	0.7068	0.5142	0.5807	0.6267

TABLE 3. The performance of the compared methods using other metrics on iCoseg dataset.

Metric	GMS	GSP	SCF	CST	CSZ	MRW	SPA	AC	COMP
F_b	0.5737	0.5734	0.5595	0.4734	0.5929	0.6170	0.5179	0.4890	0.5702
F_r	0.8689	0.8750	0.8764	0.7769	0.8768	0.8958	0.8329	0.8465	0.8819
F_{op}	0.5848	0.5983	0.5631	0.4092	0.6107	0.6540	0.4851	0.5053	0.5976
VoI	0.6293	0.6094	0.6122	0.9355	0.5979	0.5306	0.7353	0.7351	0.5755
PRI	0.8301	0.8373	0.8422	0.7179	0.8404	0.8671	0.7767	0.8008	0.8423
SC	0.8252	0.8333	0.8390	0.6987	0.8362	0.8648	0.7792	0.7944	0.8471
D_H	0.8977	0.9064	0.9083	0.8066	0.9081	0.9200	0.8858	0.8741	0.9185
d_{vD}	0.9118	0.9160	0.9160	0.8478	0.9168	0.9290	0.8881	0.8950	0.9194
BCE	0.8014	0.8095	0.8125	0.6791	0.8139	0.8414	0.7540	0.7732	0.8183
GCE	0.9007	0.9040	0.8983	0.8512	0.9043	0.9129	0.8865	0.8818	0.9128
LCE	0.9367	0.9389	0.9345	0.9050	0.9388	0.9442	0.9277	0.9193	0.9440
OCE	0.6825	0.6784	0.6782	0.7337	0.6787	0.6767	0.6898	0.7006	0.6740

more complex modeling, optimization, and calculations. Consequently, the discovery and segmentation of multiple foregrounds is also a challenge and direction in this field.

- Co-segmentation for specific applications

In practical applications, there are often sharp differences between the images to be processed, such as noise images, medical images, and specific scene images. The general model is hard to perform well. Therefore, according to specific occasions and requirements, designing a more targeted co-segmentation model to obtain better solutions has become an important research direction in this field.

TABLE 4. The effectiveness and efficiency of the compared methods using the IoU metric on MSRC dataset.

Class	#Image	GMS	GSP	SCF	CST	CSZ	MRW	SPA	AC	COMP
Animal	23	0.6285	0.5826	0.6194	0.6291	0.6502	0.6606	0.6039	0.5863	0.6545
Tree	30	0.7567	0.7583	0.6889	0.7302	0.6798	0.7374	0.6623	0.6637	0.4896
Building	30	0.7762	0.7117	0.7099	0.7487	0.6834	0.6909	0.4970	0.6282	0.6244
Plane	30	0.5418	0.5416	0.5507	0.4939	0.5309	0.5089	0.4992	0.4802	0.4980
Cow	30	0.7816	0.7681	0.7432	0.7623	0.7678	0.7399	0.7237	0.7105	0.7496
Face	30	0.6172	0.6303	0.6214	0.5830	0.5921	0.4996	0.3592	0.5654	0.5336
Car	30	0.7126	0.6922	0.6672	0.6953	0.5790	0.6782	0.6089	0.5681	0.4668
Bike	30	0.4282	0.4096	0.4691	0.5243	0.3178	0.5280	0.4384	0.4906	0.4262
Average	29	0.6554	0.6368	0.6337	0.6458	0.6001	0.6304	0.5491	0.5866	0.5553

TABLE 5. The performance of the compared methods using other metrics on MSRC dataset.

Metric	GMS	GSP	SCF	CST	CSZ	MRW	SPA	AC	COMP
F_b	0.2146	0.2104	0.1608	0.1807	0.1900	0.1987	0.1641	0.1935	0.1501
F_r	0.8068	0.7984	0.7971	0.7847	0.7885	0.7848	0.7494	0.7683	0.7861
F_{op}	0.4014	0.3695	0.3637	0.3637	0.3511	0.3471	0.3119	0.2882	0.3335
VoI	0.9370	0.9666	0.9597	1.0008	0.9898	1.0081	1.0915	1.0921	0.9849
PRI	0.7767	0.7655	0.7653	0.7563	0.7509	0.7510	0.6996	0.7331	0.7407
SC	0.7704	0.7595	0.7602	0.7443	0.7447	0.7425	0.6913	0.7228	0.7429
D_H	0.8741	0.8675	0.8614	0.8470	0.8608	0.8507	0.8310	0.8385	0.8636
d_{vD}	0.8699	0.8634	0.8596	0.8533	0.8524	0.8511	0.8236	0.8362	0.8488
BCE	0.7365	0.7250	0.7199	0.7112	0.7102	0.7115	0.6564	0.6924	0.6999
GCE	0.8215	0.8157	0.8087	0.8014	0.8101	0.8009	0.7935	0.7818	0.8168
LCE	0.8684	0.8639	0.8642	0.8551	0.8616	0.8537	0.8564	0.8303	0.8681
OCE	0.7468	0.7486	0.7435	0.7550	0.7457	0.7489	0.7524	0.7515	0.7368

TABLE 6. The effectiveness and efficiency of the compared methods using the IoU metric on Coseg-Rep dataset.

Class	#Image	GMS	GSP	SCF	CST	CSZ	MRW	SPA	AC	COMP
Blue Flag Iris	10	0.9112	0.9094	0.8920	0.8617	0.8769	0.8447	0.8973	0.7196	0.9366
Camel	24	0.7308	0.6950	0.7015	0.5578	0.6894	0.6898	0.7312	0.5150	0.6714
Cormorant	14	0.6673	0.6607	0.6696	0.6857	0.6435	0.5979	0.5681	0.5627	0.6590
Cranesbill	18	0.8985	0.8994	0.8815	0.8746	0.8713	0.8689	0.7398	0.3517	0.9052
Deer	19	0.6896	0.6844	0.6215	0.4641	0.6380	0.5118	0.4897	0.5017	0.5544
Desert Rose	49	0.8942	0.8680	0.8635	0.8655	0.8658	0.8323	0.6652	0.5885	0.8292
Dragonfly	14	0.5301	0.4977	0.3705	0.4161	0.2908	0.5539	0.1952	0.2324	0.2983
Egret	20	0.6204	0.5999	0.6089	0.5384	0.6342	0.7545	0.6141	0.6289	0.5696
Fire Pink	15	0.8893	0.8888	0.8464	0.8623	0.9016	0.8971	0.8351	0.7765	0.9064
Fleabane	19	0.8582	0.7889	0.7952	0.8157	0.7661	0.8171	0.5775	0.7116	0.8666
Forget-me-not	47	0.8711	0.8752	0.8124	0.8616	0.8675	0.8231	0.5919	0.4479	0.8542
Frog	20	0.7371	0.7408	0.7316	0.6467	0.6495	0.4769	0.5328	0.4268	0.4441
Geranium	33	0.9083	0.9060	0.9233	0.8951	0.9159	0.8781	0.7916	0.7027	0.9256
Ostrich	22	0.7581	0.7312	0.6821	0.6591	0.6812	0.6074	0.6572	0.6041	0.6099
Pear Blossom	23	0.7944	0.7903	0.8004	0.7337	0.7452	0.7805	0.4000	0.6902	0.7620
Piegon	19	0.6051	0.5886	0.6509	0.5086	0.6387	0.3831	0.5880	0.5776	0.5810
Seagull	14	0.7375	0.6081	0.6716	0.6753	0.5692	0.3800	0.6095	0.6035	0.6200
Seastar	9	0.8079	0.8191	0.7733	0.6310	0.7689	0.4908	0.6534	0.4217	0.5636
Silene Colorata	15	0.7892	0.7896	0.8066	0.7552	0.8335	0.8415	0.7947	0.6278	0.8764
Snow Owl	20	0.7010	0.6924	0.6961	0.5515	0.6988	0.6929	0.4122	0.5453	0.5142
White Campion	18	0.8506	0.8598	0.8701	0.7702	0.8549	0.8748	0.6910	0.7038	0.8972
Wild Beast	14	0.8534	0.8515	0.8352	0.5586	0.8100	0.7556	0.7108	0.6606	0.7040
Repetitive	116	0.7945	0.7683	0.7718	0.7341	0.7619	0.6586	0.5139	0.3250	0.6262
Average	25	0.7782	0.7614	0.7511	0.6923	0.7379	0.6961	0.6200	0.5620	0.7033

V. FUTURE WORK

In the future work, we will first select several suitable co-segmentation algorithms based on the different types of algorithms mentioned in this paper and the experimental results, and then more specifically test their robustness against noise images, complex background images and so on, as well as adaptation to image differences within the same group. Last but not least, we will explore the possibility of dealing with single image segmentation in a co-segmentation approach through a series of data augmentation methods, and attempt to apply constraint relationships between images to regions within the image.

Moreover, it can be seen from the experiment that the segmentation results in one group often have differences, and the

TABLE 7. The performance of the compared methods using other metrics on Coseg-Rep dataset.

Metric	GMS	GSP	SCF	CST	CSZ	MRW	SPA	AC	COMP
F_b	0.7050	0.6901	0.6394	0.6100	0.6746	0.6216	0.5774	0.4599	0.6307
F_r	0.9109	0.9069	0.9010	0.8626	0.9053	0.8699	0.8497	0.8407	0.8947
F_{op}	0.7237	0.7067	0.6831	0.6053	0.6927	0.6131	0.5470	0.4645	0.6412
VoI	0.4887	0.5081	0.5380	0.6774	0.5106	0.6455	0.6927	0.7558	0.5377
PRI	0.8884	0.8820	0.8754	0.8306	0.8763	0.8382	0.8047	0.7835	0.8604
SC	0.8872	0.8816	0.8751	0.8231	0.8776	0.8360	0.8109	0.7984	0.8693
D_H	0.9382	0.9367	0.9317	0.8954	0.9365	0.9094	0.9104	0.9166	0.9367
d_{vD}	0.9408	0.9378	0.9339	0.9069	0.9363	0.9125	0.8985	0.8924	0.9297
BCE	0.8639	0.8574	0.8487	0.7995	0.8534	0.8111	0.7769	0.7562	0.8366
GCE	0.9191	0.9164	0.9105	0.8845	0.9187	0.8878	0.8899	0.8919	0.9191
LCE	0.9469	0.9446	0.9408	0.9220	0.9461	0.9247	0.9301	0.9262	0.9468
OCE	0.7031	0.7032	0.7003	0.7205	0.6965	0.7100	0.7064	0.7031	0.6955

TABLE 8. The best, worst and mean scores of the compared methods using the IoU metric on three datasets.

Dataset	Score	GMS	GSP	SCF	CST	CSZ	MRW	SPA	AC	COMP
iCoseg	Best	0.8725	0.9127	0.8602	0.8693	0.9139	0.9478	0.8839	0.9184	0.9591
	Worst	0.2233	0.3038	0.4462	0.1136	0.4158	0.1096	0.1740	0.0790	0.1584
	Mean	0.6441	0.6478	0.6620	0.4899	0.6626	0.7068	0.5142	0.5807	0.6267
MSRC	Best	0.7816	0.7681	0.7432	0.7623	0.7678	0.7399	0.7237	0.7105	0.7496
	Worst	0.4282	0.4096	0.4691	0.4939	0.3178	0.4996	0.3592	0.4802	0.4262
	Mean	0.6554	0.6368	0.6337	0.6458	0.6001	0.6304	0.5491	0.5866	0.5553
Coseg-Rep	Best	0.9112	0.9094	0.9233	0.8951	0.9159	0.8971	0.8973	0.7765	0.9366
	Worst	0.5301	0.4977	0.3705	0.4161	0.2908	0.3800	0.1952	0.2324	0.2983
	Mean	0.7782	0.7614	0.7511	0.6923	0.7379	0.6961	0.6200	0.5620	0.7033

segmentation results of some images are significantly better than others. Therefore, in the next work, we will try to design a post-processing process to improve these bad segmentation performance. First, we need to find a reasonable segmentation quality assessment method to distinguish between good segmentation and bad segmentation, and then make use of the information provided by the good segmentation to repair bad segmentation.

VI. CONCLUSION

Faced with the poor performance of unsupervised segmentation and the large demand of data annotations for supervised segmentation, people attempted to segment simultaneously the common regions from multiple images and discovered that this idea can generate better performance than the traditional one. Thus, such methods have begun to develop and are referred to as co-segmentation.

As a survey of co-segmentation, some classical and effective algorithms were introduced and summarized by means of flowcharts and algorithm summaries. These algorithms covered almost the basic idea of most methods in the field. Then, in order to evaluate these algorithms more intuitively and objectively, an experiment was introduced. In the experiment, three datasets and multiple evaluation metrics were adopted. Through observation of experimental results, we have discovered and analyzed some existing problems. Finally, we expounded the challenges at this stage and introduced the directions of our future work, trying to give more references and insights to initiates and future researchers in this field.

ACKNOWLEDGMENT

First of all, the original version of each code used in the experiment comes from the authors of the corresponding algorithm, which has greatly helped us to carry out the experiment, so we would like to express our sincere gratitude. Secondly, the reviewers provided a lot of useful comments and suggestions for this paper. We accepted them with humility and the paper has been considerably improved. We also express sincere gratitude. Finally, we are also grateful to the editors responsible for this paper.

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 Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 993–1000.

[2] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2028–2035.

[3] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2010, pp. 269–276.

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

[5] S. Wang, H. Zhang, and H. Wang, "Object co-segmentation via weakly supervised data fusion," *Comput. Vis. Image Understand.*, vol. 155, pp. 43–54, Feb. 2017.

[6] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 465–479.

[7] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 12–27, Jan. 2016.

[8] S. Wang and A. Huang, "Salient object detection with low-rank approximation and ℓ2,1-norm minimization," *Image Vis. Comput.*, vol. 57, pp. 67–77, Jan. 2017.

[9] Y. Li, J. Liu, Z. Li, H. Lu, and S. Ma, "Object co-segmentation via salient and common regions discovery," *Neurocomputing*, vol. 172, pp. 225–234, Jan. 2016.

[10] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2129–2136.

[11] H. Chen, P. Wang, and M. Liu, "From co-saliency detection to object co-segmentation: A unified multi-stage low-rank matrix recovery approach," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2016, pp. 1602–1607.

[12] C.-C. Tsai, W. Li, K.-J. Hsu, X. Qian, and Y.-Y. Lin, "Image co-saliency detection and co-segmentation via progressive joint optimization," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 56–71, Jan. 2019.

[13] K. R. Jerripothula, J. Cai, F. Meng, and J. Yuan, "Automatic image co-segmentation using geometric mean saliency," in *Proc. Int. Conf. Image Process.*, 2015, pp. 3277–3281.

[14] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 409–416.

[15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[16] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[17] 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.

[18] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[19] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "SIFT flow: Dense correspondence across different scenes," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 28–42.

[20] 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.

[21] K. R. Jerripothula, J. Cai, and J. Yuan, "Group saliency propagation for large scale and quick image co-segmentation," in *Proc. Int. Conf. Image Process.*, 2015, pp. 4639–4643.

[22] 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.

[23] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "usstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[24] M. M. Cheng, V. A. Prisacariu, S. Zheng, P. H. S. Torr, and C. Rother, "DenseCut: Densely connected CRFs for realtime GrabCut," *Comput. Graph. Forum*, vol. 34, no. 7, pp. 193–201, 2015.

[25] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[26] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 984–989.

[27] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. Int. Conf. Comput. Vis.*, vol. 1, 2003, pp. 687–694.

[28] J. Dai, Y. N. Wu, J. Zhou, and S.-C. Zhu, "Cosegmentation and cosketch by unsupervised learning," in *Proc. Int. Conf. Comput. Vis.*, 2016, pp. 1305–1312.

[29] B. Alexe, T. Deselaers, and V. Ferrari, "Classcut for unsupervised class segmentation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 380–393.

[30] J. Winn and N. Jojic, "Locus: Learning object classes with unsupervised segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 756–763.

[31] Y. N. Wu, Z. Si, H. Gong, and S.-C. Zhu, "Learning active basis model for object detection and recognition," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 198–235, 2010.

[32] N. Ahuja and S. Todorovic, "Extracting texels in 2.1D natural textures," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[33] L. Lin, X. Liu, and S.-C. Zhu, "Layered graph matching with composite cluster sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1426–1442, Aug. 2010.

[34] H. Sundar, D. Silver, N. Gagvani, and S. Dickinson, "Skeleton based shape matching and retrieval," in *Proc. Shape Modeling Int. Symp.*, 2003, pp. 130–139.

[35] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1110–1118.

[36] C. Menier, E. Boyer, and B. Raffin, "3D skeleton-based body pose recovery," in *Proc. Int. Symp. 3D Data Process., Vis., Transmiss.*, 2006, pp. 389–396.

[37] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3159–3167.

[38] K. R. Jerripothula, J. Cai, J. Lu, and J. Yuan, "Object co-skeletonization with co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3881–3889.

[39] W. Shen, X. Bai, X. Yang, and L. J. Latecki, "Skeleton pruning as trade-off between skeleton simplicity and reconstruction error," *Sci. China Inf. Sci.*, vol. 56, no. 4, pp. 1–14, 2013.

[40] W.-P. Choi, K.-M. Lam, and W.-C. Siu, "Extraction of the Euclidean skeleton based on a connectivity criterion," *Pattern Recognit.*, vol. 36, no. 3, pp. 721–729, 2003.

[41] 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.* Cham, Switzerland: Springer, 2014, pp. 695–709.

[42] M. D. Collins, J. Xu, L. Grady, and V. Singh, "Random walks based multi-image segmentation: Quasiconvexity results and gpu-based solutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1656–1663.

[43] Y. Wang, B.-J. Yoon, and X. Qian, "Co-segmentation of multiple images through random walk on graphs," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 1811–1815.

[44] 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.*, Jun. 2015, pp. 3837–3845.

[45] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[46] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[47] F. Meng, H. Li, and G. Liu, "A new co-saliency model via pairwise constraint graph matching," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, 2012, pp. 781–786.

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

[49] F. Meng, B. Luo, and C. Huang, "Object co-segmentation based on directed graph clustering," in *Proc. Vis. Commun. Image Process.*, 2014, pp. 1–5.

[50] 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.*, Jun. 2016, pp. 687–695.

[51] C. Wang, H. Zhang, L. Yang, X. Cao, and H. Xiong, "Multiple semantic matching on augmented N -partite graph for object co-segmentation," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5825–5839, Dec. 2017.

[52] 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.

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

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

[55] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, Feb. 2007.

[56] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.

[57] F. Meng, H. Li, and G. Liu, "Image co-segmentation via active contours," in *Proc. Int. Symp. Circuits Syst.*, 2012, pp. 2773–2776.

[58] F. Meng, H. Li, K. Ngan, B. Zeng, and N. Rao, "Cosegmentation from similar backgrounds," in *Proc. Int. Symp. Circuits Syst.*, 2014, pp. 353–356.

[59] Z. Zhang, X. Liu, and N. Q. Soomro, "An efficient image co-segmentation algorithm based on active contour and image saliency," in *Proc. MATEC Web Conf.*, vol. 54, 2016, Art. no. 08004.

[60] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Image cosegmentation by incorporating color reward strategy and active contour model," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 725–737, Apr. 2013.

[61] A. Ion, J. Carreira, and C. Sminchisescu, "Image segmentation by figure-ground composition into maximal cliques," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2110–2117.

[62] D. Banica, A. Agape, A. Ion, and C. Sminchisescu, "Video object segmentation by salient segment chain composition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, Jun. 2013, pp. 283–290.

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

[64] O. Boiman and M. Irani, "Similarity by composition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 177–184.

[65] A. Faktor and M. Irani, "'Clustering by composition'—Unsupervised discovery of image categories," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 474–487.

[66] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[67] I. Endres and D. Hoiem, "Category independent object proposals," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 575–588.

[68] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in ImageNet," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 459–473.

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

[70] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 1800–1807.

[71] H. H. Chang, A. H. Zhuang, D. J. Valentino, and W. C. Chu, "Performance measure characterization for evaluating neuroimage segmentation algorithms," *NeuroImage*, vol. 47, no. 1, pp. 122–135, 2009.

[72] J. Pont-Tuset and F. Marques, "Measures and meta-measures for the supervised evaluation of image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2131–2138.

[73] S. Nowozin, "Optimal decisions from probabilistic models: The intersection-over-union case," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 548–555.

[74] G. Hripcsak and A. S. Rothschild, "Agreement, the F-measure, and reliability in information retrieval," *J. Amer. Med. Informat. Assoc.*, vol. 12, no. 3, pp. 296–298, 2005.

[75] G. Liu and R. M. Haralick, "Assignment problem in edge detection performance evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2000, pp. 26–31.

[76] D. R. Martin, "An empirical approach to grouping and segmentation," Ph.D. dissertation, Dept. EECS, Univ. California, Berkeley, Berkeley, CA, USA, Aug. 2003.

[77] M. Meilă, "Comparing clusterings: An axiomatic view," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 577–584.

[78] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[79] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 929–944, Jun. 2007.

[80] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[81] Q. Huang and B. Dom, "Quantitative methods of evaluating image segmentation," in *Proc. Int. Conf. Image Process.*, vol. 3. 1995, pp. 53–56.

[82] J. S. Cardoso and L. Corte-Real, "Toward a generic evaluation of image segmentation," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1773–1782, Nov. 2005.

[83] S. Dongen, "Performance criteria for graph clustering and Markov cluster experiments," Centre Math. Comput. Sci., Nat. Res. Inst. Math. Comput. Sci., Amsterdam, The Netherlands, 2000.

[84] M. Polak, H. Zhang, and M. Pi, "An evaluation metric for image segmentation of multiple objects," *Image Vis. Comput.*, vol. 27, no. 8, pp. 1223–1227, 2009.

[85] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.* , vol. 2, Jul. 2001, pp. 416–423.

[86] H. Vojodi and A. M. E. Moghadam, "A supervised evaluation method based on region shape descriptor for image segmentation algorithm," in *Proc. CSI Int. Symp. Artif. Intell. Signal Process.*, 2012, pp. 18–22.

ZHOUMIN LU received the B.S. degree in computer science and technology from Northeastern University, Shenyang, China, in 2017. He is currently pursuing the M.S. degree with the College of Mathematics and Computer Science, Fuzhou University. His research interests include the deep learning, computer vision, and image processing.

HAIPING XU received the Ph.D. degree from the Center for Discrete Mathematics and Theoretical Computer Science, Fuzhou University, Fuzhou, China, in 2018. She is currently a Lecturer with the College of Mathematics and Data Science, Minjiang University, Fuzhou. Her research interests include computer vision, image processing, and partial differential equations.

GENGGENG LIU received the B.S. degree in computer science and the Ph.D. degree in applied mathematics from Fuzhou University, Fuzhou, China, in 2009 and 2015, respectively, where he is currently an Assistant Professor with the College of Mathematics and Computer Science. His research interest includes computational intelligence and its application.