Pattern Recognition 103 (2020) 107297

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

A survey on image and video cosegmentation: Methods, challenges and analyses

Yan Ren[a,b,*], Adams Wai Kin Kong[b], Licheng Jiao[a]

[a] Xidian University, China
[b] Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 14 February 2019
Revised 29 November 2019
Accepted 21 February 2020
Available online 24 February 2020

Keywords:
Image cosegmentation
Video cosegmentation

ABSTRACT

Image and video cosegmentation is a newly emerging and rapidly progressing area, which aims at delineating common objects at pixel-level from a group of images or a set of videos. Plenty of related works have been published and implemented in varied applications, but there lacks a systematic survey on both image and video cosegmentation. This paper provides a comprehensive overview including the existing methods, applications, and challenges. Specifically, different cosegmentation problem settings are described, the formulation details of the methods are summarized and their potential applications are listed. Moreover, the benchmark datasets and standard evaluation metrics are also given; and the future directions and unsolved challenges are discussed.

1. Introduction

Digital media and social networking have become ubiquitous in our daily lives. The development of inexpensive storage, the prevalence of digital cameras and video recorders, the accessibility of community-sharing websites and the proliferation of mobile devices have led to an enormous flood of unorganized images and videos. The huge amount of data, sorely lacking manual annotations, promotes the emergence and growth of image and video cosegmentation methods, which aim at simultaneously segmenting common foregrounds from two or more images/videos and providing pixel-level delineations. The common foregrounds for cosegmentation can refer to the same object captured from different viewpoints (e.g. photos of the same cheetah in different poses Fig. 2(a)), or different object instances from the same class (e.g. videos of different kinds of birds Fig. 3(a)). The cosegmentation problem could become more challenging if input images/videos are noisy, or have multiple common foregrounds or multiple object parts being consideration (Figs. 2(b–e) and 3(b–d)). For relieving the burden on manual labeling, image and video cosegmentation methods can improve the performance of other image processing methods like image representation [1,2], classification [3,4], recognition [5], and so on. Meanwhile, cosegmentation exhibits a huge potential for various applications, such as image retrieval [6], 3D model construction [7], video summarization [8], and motion recognition [9], etc.

Compared with unsupervised segmentation, cosegmentation does not aim at segmenting every pixel in the images or videos. Because of the extra information from the common foregrounds, including their correspondence in different images/videos and the traditional methods do not rely on training, cosegmentation is generally considered as a weakly supervised segmentation problem with some exception. Deep-learning-based methods mentioned in Section 3.1.4, which are supervised methods, do not build large models based on training data but directly analyze the correspondence among the input images or videos to segment the common foregrounds.

In Fig. 1(a), the publication statistics of the existing cosegmentation methods are given, which show the growing tendency of cosegmentation papers. Compared with image cosegmentation, video cosegmentation is still an underdeveloped research area. In the past two decades, image and video cosegmentation have attracted increasing attention. However, few related surveys have been published. Zhu et al. [10] and Biradar et al. [11] give brief introductions on image cosegmentation. Meanwhile, Daryanto et al. [12] provide some discussions on image cosegmentation without mentioning video cosegmentation. The application potentials, research opportunities and rapid development in this field motivate us to systemically summarize the existing cosegmentation research and list down future research directions. This paper intends to give a more comprehensive overview of this fast-growing area, which provides method taxonomy, fundamental frameworks, es-

* Corresponding author.
E-mail address: yan.ren@ntu.edu.sg (Y. Ren).

https://doi.org/10.1016/j.patcog.2020.107297
0031-3203/

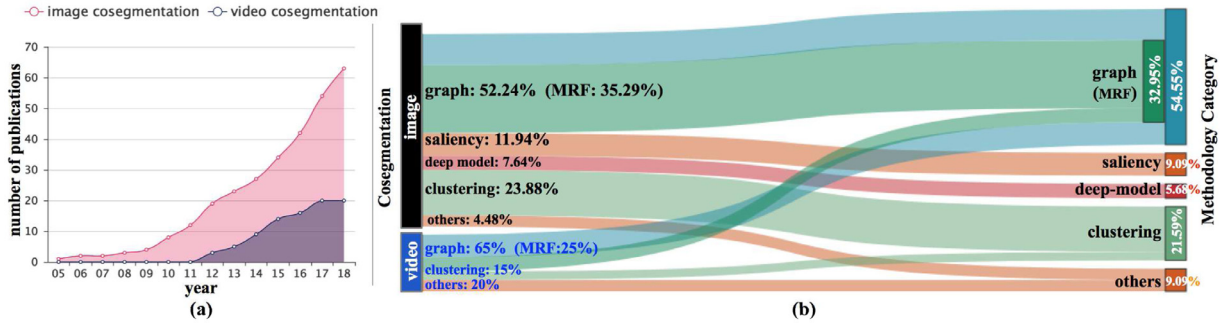


Fig. 1. (a) The growth tendency of publication from 2005 to 2018. (b) In this paper, existing image/video cosegmentation methods are classified into 4 main methodologies: graph-based, clustering-based, saliency-based and deep-model-based. This statistics on the left represent the class-wise ratios of image and video cosegmentation separately, while the statistics on the right represent the class-wise ratios of all the cosegmentation methods. Especially, the MRF-based models, as a subcategory of the graph-based methods, occupy a large share.

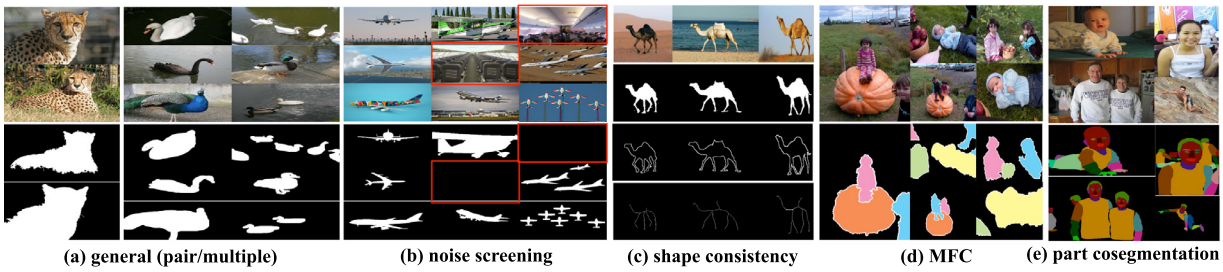


Fig. 2. Illustration of the problem categories of image cosegmentation. The color images in (a)-(c) are input images and the binary images are the ground truth. The natural images in (d)-(e) are input images and the rest images are the ground truth.

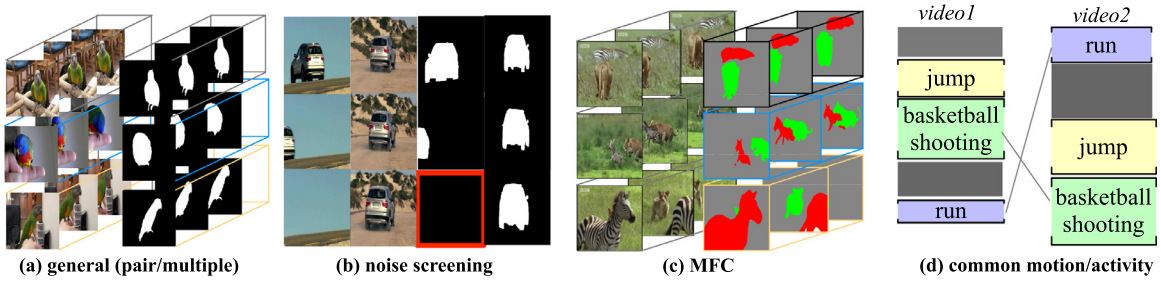


Fig. 3. The figure illustrates the problem categories of video cosegmentation. The color figures in (a) and (b) represent input videos and the binary figures are the ground truth. In (c), *MFC* represents the multiple foreground video cosegmentation problem. The first column represents the input videos and the second column represents the ground truth. The color blocks in (d) represent varied paragraphs of videos.

sential formulation details, possible applications, and future tendencies. The contributions can be highlighted as follows: 1) To the best of our knowledge, this paper is the first comprehensive survey covering both image and video cosegmentation. 2) To clarify the essence of cosegmentation methods, typical methodologies and formulations are summarized and outlined from a wide range of existing methods. 3) Through unveiling the relevance as well as the difference between the two co-related research areas, common ideas and challenges are identified and summarized for future study.

1.1. Related research areas

The related research areas of image and video cosegmentation are **image and video segmentation, cosaliency and region proposal**. These research directions all focus on the discovery of the object-of-interests.

Image and video segmentation has been studied for decades with a wide range of application [13]. Recently, **semantic segmentation** has become a hot research topic, which partitions images or videos into semantical parts with pre-defined class labels. In the past, probabilistic graphical models [14,15] were used to design semantic segmentation methods but most recent methods are based on deep learning [16–18]. In addition, some other related trending topics are **image instance segmentation** and **video object segmentation**. Image instance segmentation aims at labeling the pixels within an image with instance labels [19]. Compared with semantic segmentation, image instance segmentation pays more emphasis on distinguishing occluded object instances. Video object segmentation aims at discovering meaningful objects from a video [20]. Compared with these segmentation methods, cosegmentation methods intend to capture correspondence among the common foregrounds without pixel-level manual annotations. Meanwhile, most of the segmentation methods focus on detecting and delineating object regions from one image or video, while cosegmentation methods intend to find out coherent objects among multiple images or videos.

Image and video cosaliency aim at segmenting common and salient regions from images and videos [21], which derive similar salient regions from multiple relevant images or videos. For ex-

ample, Song et al. [22], Cong et al. [23] and Han et al. [24] use respectively bagging-based clustering, refinement-cycle model and multi-constraint feature matching for cosaliency detection. Zhang et al. [25] construct a deep learning framework for cosaliency detection with auxiliary annotated datasets. The two main differences between cosaliency and cosegmentation are given as follows: 1) Compared with cosegmentation, cosaliency methods concentrate on common saliency regions rather than coherent object regions. Cosegmentation methods can exploit common foregrounds from images and videos even though the common foregrounds are not salient; 2) The results of cosaliency are continuous foreground probability values of pixels, while the results of image cosegmentation are discrete pixel-level labels.

Segment-based region proposal methods generate potential object regions from images and videos without manual annotations. For example, Endres et al [26] generate a set of object segmentation proposals from images by learning region affinity. Li et al. [27] obtain a pool of segments for each video frame by a parametric min-cut figure-ground segmentation algorithm. Compared with video cosegmentation, region proposal methods provide potential object candidates rather than co-occurrence objects.

The rest of this paper is organized as follows. Section 2 discusses 6 different cosegmentation problem settings. Section 3 summarizes the methodology taxonomy for both image and video cosegmentation first and then gives more details on existing methods. Section 4 provides potential applications. Section 5 lists respectively benchmark datasets. Section 6 provides evaluation metrics and comparison results. Section 7 gives some conclusive remarks and describes some future research directions.

2. Problem categories

The original concept of cosegmentation was first discussed in [28], which intends to use a pair of images with common objects to enrich the information for image segmentation. Since then, cosegmentation has been extensively studied over a decade, the scope of which has been broadened from image-pairs cosegmentation [29–31] to multiple image cosegmentation [32–34] and video cosegmentation [35,36]. Meanwhile, more difficult cosegmentation challenges have been identified by researchers [2,8]. The existing cosegmentation problems can be divided into the following categories: **General (pair/multiple)**: In this setting, cosegmentation methods take a pair or multiple images [37–39] /videos [40,41] as input and the expected output is the segmented common foregrounds (Figs. 2(a) and 3(a)). Note that there exists a few interactive image cosegmentation methods, which introduce user scribbles as another input [6,42]. Actually, most of the interactive methods are still dealing with the general(multiple) cosegmentation problem. **Noise screening**: For image cosegmentation, considering a large volume of data, the class labels of images may be wrongly annotated due to the inevitable pressures of manual works. Therefore, screening out the noisy images [43,44] is considered for multiple images. Fig. 2(b) shows the input and the ground truth of this problem, where the red boxes indicate noisy images. For video cosegmentation, some frames of the videos might not have objects. Thus, screening irrelevant/noise frames is considered [40,45,46] (Fig. 3(b)). **Multiple foreground cosegmentation (MFC)**: Other than only considering one type of common objects, the MFC problem aims at segmenting multiple kinds of common foregrounds from images [47,48] and videos [49,50] (Figs. 2(d) and 3(c)). **Shape consistency** (only for image cosegmentation): Some common objects in images have repetitive shape patterns. Thus, shape consistency [51,52] focusses on shape models and avoids the distraction of varied color distributions of objects (Fig. 2(c)). **Part cosegmentation** (only for image cosegmentation): The problem of part cosegmentation [53,54] requires not only common objects but also detailed common local part regions of the objects. Compared with MFC, part cosegmentation focusses on multiple kinds of common local parts of the same type of common foregrounds (Fig. 2(e)). **Common motion/activity** (only for video cosegmentation): In addition to appearance commonality, a few works study common motion/activity in videos [8,9], which is still an under-investigated and challenging research area (Fig. 3(d)).

3. Methodology taxonomy

The basic flowchart of cosegmentation is usually composed of three main steps: representation generation, correspondence estimation and outline refinement (Fig. 4). Representation generation is the initialization step for cosegmentation. For image cosegmentation, input images are usually represented with image subregions like pixels [28], superpixels [55], hierarchies [56] and region proposals [44] as the basic units for cosegmentation. Varied image features are extracted from the subregions, such as color/intensity histograms [30], SIFT features [43], HOG [44], bag of words [55], visual dictionary [57], adaptive features [38,58] and CNN features [39,48], etc. To further reduce the computational complexity, Wang et al. [59] and Jerripothula et al. [60] use global GIST features and cluster images into small groups and Li et al. [61] rank the images by their complexities and deal with simple images first. For video cosegmentation, the videos can either be represented with the 2D image subregions based on each video frame, or 3D video subregions like supervoxels [36], trajectories [62], and tracklets [49], etc.

Correspondence estimation is the core step of cosegmentation. For image cosegmentation, two kinds of correspondence are analyzed: intra-image correspondence in a single image and inter-image correspondence among images. The intra-image correspondence encodes the correspondence between the potential foreground and background of the same image. Gaussian Mixture Model [63], normalized Euclidean distances among subregion features [64] and saliency detection methods [65] are frequently utilized for intra-image correspondence. The inter-image correspondence describes the relationships among different images, which can be computed by multiple similarity measurements (e.g. SIFT flow [66], similarity propagation [67], etc.). For video cosegmentation, three kinds of correspondence are considered: intra-frame correspondence in a single frame, inter-frame correspondence among frames in the same video and across-video correspondence among videos. Intra-frame correspondence focusses on the spatial relationships among subregions in a unary frame. Approaches such as objectness [35,68], saliency detection [46,50] and background subtraction [62] are utilized to encode the spatial information. Inter-frame correspondence reveals the temporal relationships among frames in each video. Optical flow [40,69] and motion cues [49,70] are frequently utilized to exploit temporal information. Cross-video correspondence studies the spatial and temporal relationships among videos, which can be formulated with varied models, such as trajectory consistency [62], spatiotemporal SIFT flow [40], and distance-dependent Chinese Restaurant Process [71], etc. Compared with image cosegmentation, video cosegmentation pays more attention to the spatiotemporal properties, therefore, video cosegmentation is not just a simple extension of image cosegmentation.

Outline refinement aims at refining the object boundaries, which can efficiently improve the precision of image and video cosegmentation. Methods such as bilateral filter [55], Otsu's thresholding [72], GrabCut [45,60], and pixel-level spatiotemporal graph-based segmentation [68,73] are considered for refinement.

As aforementioned above, correspondence estimation plays an important role in cosegmentation methods. Thus, a comprehensive analysis of the correspondence models can outline a clear picture of cosegmentation. Existing image and video cosegmenta-

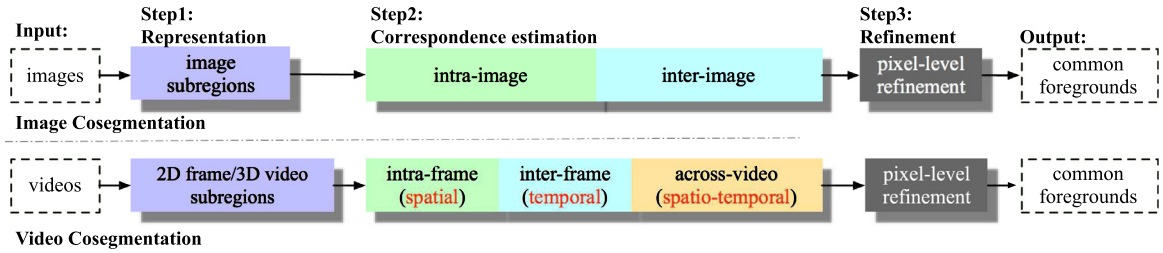


Fig. 4. Brief flowcharts of image and video cosegmentation methods. For cosegmentation, there are 3 key steps: subregion representation, correspondence estimation, and outline refinements. Compared with image cosegmentation, video cosegmentation methods take spatiotemporal information into consideration.

Table 1
The pros and cons of the method categories for both image and video cosegmentation methods.

Category	Pros	Cons
graph-based	rich geometrical information	high computational complexity
clustering-based	relatively higher efficiency	sensitive to initialization
saliency-based	rich prior information	the missing of non-salient but common regions
deep-model-based	high-level features	lack of training data

tion methods are divided into four main categories based on correspondence estimation models: **graph-based, clustering-based, saliency-based, and deep-model-based**. The existing video cosegmentation methods appear only in graph-based and clustering-based categories. The percentages of the categories are given in Fig. 1(b). The pros and cons are summarized in Table 1 and some rough comparisons among the categories are provided in Figs. 7 and 9(a). General formulations of each category and more details of the methods are given in the following sections.

3.1. Image cosegmentation methodology taxonomy

3.1.1. Graph-based image cosegmentation

The graph-based category is the most popular one among the existing methods (Fig. 1(b)). It formulates the image cosegmentation problem as a labelling problem of common foregrounds by graph construction and matching. More precisely, given a group of n images $\{I^i\}_{i=1}^n$, each image I^i is divided into a set of p_i small image subregions $V^i = \{v_k^i\}_{k=1}^{p_i}$ (e.g., a set of pixels, superpixels or object proposals). Then based on the subregions in all the images $V = \{V^i | i = 1, \ldots, n\}$, a graph $G = (V, Ed, L)$ can be constructed, where Ed represents the edges connecting the vertices V and $L = \{l_k^i | i = 1, \ldots, n, k = 1, \ldots, p_i\}$ represents the labels of vertices to be estimated. The labels could be either binary or multiple values based on the problem category. For general cosegmentation, shape consistency and noise screening problems, binary labels $l \in \{1, 0\}$ are utilized. For part cosegmentation and MFC problems, multiple labels $l \in \{1, \ldots, m\}$ are considered for multiple foreground instances. Details of the graph-based image cosegmentation methods are summarized in Tables 2 and 3.

The topology of the graph is determined by Ed, and the correspondence information can be imbedded in the graph through edge/vertex weight assignment. Four types of graph structures have been commonly utilized for image cosegmentation: complete, tree, directed and n-partite. More specifically, **Complete graph** [28,53] is constructed with all vertices connecting to each other by edges. The complete graph gives more flexibility to model feature relationships, which is the most popular graph in the graph-based methods; **Tree** [56,65] is a bipartite graph with a hierarchical structure, which manages to mine richer spatial information for cosegmentation. Meanwhile, the hierarchical structure makes it possible to exploit overlapping image subregions with different scales as the vertices. **Directed graph (digraph)** [81] is constructed by directed edges with weights connecting the vertices within one image to those in another. The shortest path of digraph indicates the strongest similarities among the common foregrounds. However, the shortest path assumes the existence of common foregrounds in all images, which is not be suitable for the noise screening problem; **N-partite graph** [82] divides vertices into n subsets. Each subset involves a group of similar vertices. Thus N-partite graph can be utilized to model the problem of MFC with n common foregrounds.

The initial image cosegmentation methods [28,37] introduce Markov Random Fields (MRF) for the undirected graphical model. Since then, MRF has nearly dominated the graph-based category (Table 2). Under unsupervised conditions, MRF models infer the labels of vertices by analyzing the foreground and background probability distributions, which can be formulated as an energy minimization problem:

$$\min_l \underbrace{E^{\text{intra}}}_{unary} + \underbrace{E^{smooth} + E^{\text{inter}}}_{pairwise}. \tag{1}$$

E^{intra} represents the intra-image discriminability for each unary image. Some methods define E^{intra} as a ballooning term, which biases to foreground labelling: $E^{\text{intra}} = \sum_{i=1}^n \sum_{k=1}^{p_i} w_k^i l_k^i$, where the weights w_k^i reflects the foreground possibilities such fixed constants [28,74], deviation penalties with a RBF kernel [30,75]. Assuming that $\{f_k^i\}_{k=1}^{p_i}$ are features extracted from the vertices, E^{intra} can also be defined as Gaussian Mixture Models (GMMs): $E^{\text{intra}} = \sum_{i=1}^n \sum_{k=1}^{p_i} -\log(l_k^i P(f_k^i|\theta^{F_i}) + (1 - l_k^i) P(f_k^i|\theta^{B_i}))$, where θ^{F_i} and θ^{B_i} are GMMs trained on foregrounds and backgrounds and $P(\cdot)$ represents the probabilities of label assignments. The E^{intra} can encode various feature distributions through $P(\cdot)$, such as color distributions [63], color and saliency priors [5] and local and global shape feature distributions [79]. In addition, E^{intra} can also be defined by saliency detection approaches [1,65], which utilize saliency priors for unary image foreground inferring. E^{inter} in Eq. 1 represents the pairwise correspondence between images, which can be represented as: $E^{\text{inter}} = \lambda \sum_{i,j=1}^n \sum_{k=1}^{p_i} \sum_{m=1}^{p_j} D(f_k^i, f_m^j)$, where $D(\cdot)$ measures the consistency of two features from different images, and λ is the trade-off to balance the three terms in Eq. (1). The consistency between features can be defined with varied kinds of formulations such as the L_1-norm distance between empirical un-normalized histograms: $D(f_1, f_2) = \|f_1 - f_2\|_1$ [28], squared L_2-norm as a relaxation of L_1-norm: $D(f_1, f_2) = \|f_1 - f_2\|_2^2$ [30], and the spectrum of normalized Laplacian matrix [78]. Moreover,

Table 2
Details of graph-based image cosegmentation methods with Markov Random Field models. L represents the graph labels, V represents the vertices of graph, E^{intra} and E^{inter} are two energy terms for graph edge weight assignments. ** represents the interactive image cosegmentation methods with user scribbles as an extra input [28,30,74,75,56,63,31,76,6,42,77,1,78,57,5,66,61,65,37,51,52,79,53,80].

paper, year	problem	graph	L	V	E^{intra}	E^{inter}	optimization	feature	dataset
[28], 2006	general(pair)	complete	binary	pixel	likelihood weights	L_1 distance	graph cut	RGB	self-built
[30], 2009					deviation penalty	L_2 distance	Pseudo-Boolean optimization	RGB, gradients	
[74], 2010					likelihood weights	Boykov-Jolly model	Dual Decomposition	RGB	
[75], 2010					deviation penalty	L_2 distance, similarity reward	maximum flow	intensity, Gabor filter	
[56], 2016		tree		super -pixel	GMM	optical flow, hierarchy rebuild	iterated graph cuts	Bag of Word, HOG	PASCAL, Internet dataset, FG3Dcar
[63], 2010	general (multiple)	complete		pixel	GMM, confidence	density -clustering	min-cut	Harris, SIFT	self-built
[31], 2011					likelihood weights	low conditional entropy (scale invariant)	Pseudo-Boolean optimization	RGB, texture, SIFT	self-built, MSRC Weizman horse, Oxford flowers,
**[76], 2016				super -pixel	GMM	L_1 distance, higher order cliques	graph cut	scribbles	self-built
**[6], 2010						contrast sensitive Potts model		color, scribbles	iCoseg
**[42], 2011									
**[77], 2015						GMM, spline regression	hybrid optimization		
[1], 2011					cosaliency	visual vocabulary learning	graph cut	RGB, visual words	MSRC, Weizman horse
[78], 2012				pixel, super -pixel	GMM (multi-scale)	χ^2 distance, normalized Laplacian		RGB, SIFT	MSRC, iCoseg
[57], 2012				pixel	GMM	subspace structure	maximizing submodular functions	RGB	
[5], 2014						joint GMM	GrabCut	RGB, SIFT	MIT, Caltech-28
[66], 2016						SIFT flow (multiple groups)	expectation maximization (EM)		iCoseg, Cat-Dog, Caltech-UCSD birds, Internet dataset
[61], 2017					GMM +image ranking	L_2 distance	Grabcut	color, SIFT	iCoseg
[65], 2016		tree		super -pixel	saliency detection	hierarchical spatial dependency	Pseudo-Boolean optimization	RGB, texture	
[37], 2005	shape	complete		pixel	GMM	layered pictorial structures	graph cut	outline, texture	self-built
[51], 2013						active basis model, L_2 distance		RGB	MSRC,iCoseg, Coseg-Rep
[52], 2017				region proposal	skeleton rebuild, GMM	neighbour-based consistency	skeleton pruning, Grabcut	RGB, SIFT	CO-SKEL
[79], 2018	noise				GMM + shape priors +attentiveness	L_2 distance	graph cut	CNN, Lab	MSRC, Internet dataset, Graz02
[53], 2017	part		multi	pixel	GMM	part and structure consistency	normalized cut	shape context	PASCAL, Cat-Dog, Caltech-UCSD birds, UCF Sports Actions
**[80], 2017					seed matching	region consistency	EM	color, objectness, shape, scribbles	

Rubinstein et al. [43] and Meng et al. [66] introduce SIFT flow to measure the distances between co-related features: $D(f_1, f_2) = \|f_1(v) - f_2(v + w(v))\|_1$, where v is a vertex in one image and $v + w(v)$ is a vertex corresponding to v in another image based on SIFT flow method [83]. E^{smooth} in Eq. (1) is a contrast sensitive smoothness term, which penalizes assigning neighbouring subregions in an image with different labels [5,53,74]. It can be defined as: $E^{smooth} = \beta \sum_{i=1}^{n} \sum_{(k,m)\in\mathcal{N}} \delta(l_k^i \neq l_m^i) \exp(-\|f_k^i - f_m^i\|_2^2)$, where $\mathcal{N}$ represents a neighbourhood of subregions, β is a scaling parameter, and δ is the Dirac delta function. Under the smoothness term, similar neighbors are more likely to be assigned with same labels.

In addition to MRF, some other graphical methods have also been considered for cosegmentation (Table 3). For instance, Vicente et al. [7] train Random Forest regressor for general image cosegmentation. Kim et al. [85] combine GMM with spatial pyramid matching(SPM) for MFC problem. Meng et al. [54] inject part seed information into digraph for part segmentation. Han et al. [84] use close-loop graph with graph optimized-flexible manifold for part segmentation. In graph-based methods, the topology of graphs and the definition of their edge weights are two critical elements to achieve outstanding performances. Compared the other categories, graph-based methods exploit richer geometrical information through graph representation and matching. However, complicated graphical models would result in higher computational complexity.

3.1.2. Clustering-based image cosegmentation

The clustering-based category focusses on partitioning image subregions into clusters, so that common foregrounds could be grouped into same clusters. More specifically, given n images $\{I^i\}_{i=1}^n$ and the corresponding subregions $\{v_m^i\}_{m=1}^{p_i}$, clustering methods aim at dividing the subregions into K classes. For each image, a label vector $l_m^i \in R^K$ is defined for each subregion. If the mth subregion in the ith image belongs to the kth class, $l_m^i(k) = 1$, otherwise $l_m^i(k) = 0$, $k = 1, \ldots, K$, $m = 1, \ldots, p_i$ and $i = 1, \ldots, n$. For all the images, a label matrix $L \in R^{P \times K}$ ($P = \sum_{i=1}^{n} p_i$) is constructed to denote all subregions. The cluster number K is a predefined value. In most of the existing methods, two clusters ($K = 2$) are considered: common foregrounds and backgrounds. Meanwhile, to fully exploit correspondence among subregions, some methods [88,89] consider multiple clusters ($K > 2$) first and then utilize prior knowledge to reduce the number of clusters to 2. For the MFC cosegmentation problem [47,90], K larger than 2 is considered for multiple common foregrounds. To find out the cluster labels, classic clustering methods are integrated with cosegmentation, such as discrimina-

Table 3
Details of graph-based image cosegmentation methods with non-MRF graphical models. *L* represents the graph labels, *V* represents the vertices of graph [7,33,81,39,82,84,29,43,85,86,87,54].

paper, year	problem	graph	*L*	*V*	edge weight assignment	optimization	feature	dataset
[7], 2011	general (multiple)	complete	binary	pixel	random forest regressor	loopy belief propagation	color, texton, SIFT	iCoseg, MSRC
[33], 2013		complete			affinity density, consensus scoring	graph cut	color, HOG	MSRC, iCoseg, PASCAL
[81], 2012		directed		super pixel	color, shape and saliency similarity	dynamic programming for shortest path	color, shape	MSRC, ETHZ shape
[39], 2016		close -loop			Graph optimized-flexible manifold ranking algorithm	GrabCut	color, texture, CNNs	iCoseg, Internet dataset, PASCAL
[82], 2017		n-partite			multiple clique matching	Mixed-Binary Integer Program	FCN, R-CNN	iCoseg, MSRC, PASCAL, Coseg-Rep
[84], 2018		self-modified close-loop			manifold ranking +re-assessing strategy	iterative optimi -zation process	color, texture, dense SIFT, CNN	iCoseg, Internet dataset, PASCAL
[29], 2008	shape	complete		pixel	level-set with biased shape dissimilarity	bayesian statis -tical inference	average gray level	self-built
[43], 2013	noise	complete			saliency detection, SIFT flow, foreground likelihood	GrabCut	saliency, RGB, SIFT flow	MSRC, iCoseg, Internet dataset
[85], 2012	MFC	adjacency	multiple	super -pixel	subtree assignment with GMM+SPM	dynamic programming	RGB, spatial pyramid of gray, HSV, SIFT	MFCflicker, ImageNet
[86], 2014	part	tree		pixel	deformable part models nonrigid mapping	energy minimization	HOG	MFCflicker, iCoseg
[87], 2016		complete directed		pixel, boundary fragment	global shape and local part propagation	dynamic programming, belief propagation	SIFT, shape context	PASCAL, Caltech-UCSD birds, Cat-Dog
[54], 2018		directed		pixel, region proposal	dense pixel and shape matching, seed-based shape decomposition	dynamic programming	shape context	PASCAL, Caltech -UCSD birds, Cat-Dog, UCF Sports Actions

tive clustering [32], random walker clustering [64], Anisotropic diffusion based clustering [88], etc. Details of the clustering category are summarized in Table 4.

Discriminative clustering-based cosegmentation can be formulated as:

$$\min_{L} \; L^T\left(A^{\text{intra}} + A^{\text{inter}}\right)L, \tag{2}$$

where A^{intra} represents the similarity matrix within an image, and A^{inter} represents the correspondence matrix among images. A^{intra} can be a normalized Laplacian matrix [32,92,93] defined as: $A^{intra} = I_P - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, where D is a diagonal matrix and its elements are the sum of the row of W [99], and I_P is a p-dimensional identity matrix. The block-diagonal matrix $W \in R^{P\times P}$ is constructed with $W^i \in R^{p_i\times p_i}$ on diagonal. For each image, assuming that two kinds of features are extracted from the subregions, $\{f_k^i\}_{k=1}^{p_i}$ and $\{g_k^i\}_{k=1}^{p_i}$, $W^i(k, m)$ can be defined as the normalized Euclidean distance between subregions: $W^i(k,m) = \exp(-\lambda_f\|f_k^i - f_m^i\|_2^2 - \lambda_g\|g_k^i - g_m^i\|_2^2)$. $A^{inter} = \lambda_A(I_P - \frac{1}{P}1_P1_P^T)(P\lambda_A I_P + \mathscr{K})^{-1}(I_P - \frac{1}{P}1_P1_P^T)$,where $\mathscr{K} \in R^{P\times P}$ is a kernel derived based on χ^2 distance [32,93]. The formulation of discriminative clustering reserves high flexibility to involve varied consistency constraints. For example, Sun et al. [91] denote A^{intra} as the response maps of part detectors, and A^{inter} as the combination of kernel methods and dense feature matching.

Random walk clustering is a classic graph partitioning method, which simulates random walks starting from each subregion to a set of user predefined foreground or background seed points. The clustering labels of subregions are assigned based on whether the random walks reach a foreground seed l^* first or not. Random walk clustering-based cosegmentation can be formulated as:

$$\min_{L} L^T B^{intra} L + E^{inter}, s.t.\; l^s = l^*, \tag{3}$$

where l^s represents the optimized label of seeds, and l^* represents the pre-known label of seeds. B^{intra} represents the intra-image correspondence. Given the distance matrix W mentioned above, B^{intra} can be defined as a Laplacian matrix [64]. Meanwhile E^{inter} encodes the correspondence among images as the E^{inter} mentioned in Section 3.1.1. For example, Grady et al. [64] formulate $E^{inter} = \lambda\sum_{i=1}^{n}\sum_{k=1}^{p_i}\|f_k^i - \bar{f}\|_2^2$, which encodes the comparison between the foreground feature f_k^i and the global mean histogram $\bar{f}$. The random walk clustering-based cosegmentation intends to reserve the advantages of random walk and inject correspondence information into the formulation. In addition to single random walk, some methods also involve multiple random walks (foreground and background walkers) with hybrid restart rules [8] and alternating random walks based on intra-image and inter-image similarities [94]. Note that although seed points are utilized, cosegmentation based on random walk clustering can still be regarded as a weakly supervised problem, considering no training data has joined the computation.

Anisotropic diffusion methods regard the clustering problem as maximizing the temperature of a system with K heat sources. As for cosegmentation, the temperature represents a submodular function based on the similarities among subregions [88,89]. The submodularity of diffusion methods makes it possible to implement cosegmentation on large-scale image collections. In addition to the aforementioned clustering methods, some other clustering strategies have also been utilized for image cosegmentation. For example, Tao et al. [2] introduce affinity propagation clustering for common shape models. Chen et al. [98] create coherent and aligned clusters as visual subcategories for feature extraction and noise screening. Meng et al. [47] combine directed graph with clustering for enriching spatial information. Tao et al. [34,97] use saliency-guided constrained clustering to evaluate the similarity between saliency-prior-predicted labels and the final resultant labels.

Compared with graph-based methods, the clustering models relieve the burden on the design of graph structures. However, most of the clustering-based methods are sensitive to initialization, such as the choice of seeds for random walk, the assignment of cluster

Table 4
Details of clustering-based image cosegmentation methods. K represents the number of clusters. Intra- and inter-image represent the spatial consistency in one image and the correspondence among images separately [91,92,32,93,64,94,55,95,96,34,97,88,89,2,98,47,90].

paper, year	problem	K	sub-region	clustering	intra-image	inter-image	optimization	feature	dataset
[91], 2016	general (multiple)	binary	pixel	discriminative	part detectors	dense feature matching	graph cut	SIFT	Internet dataset, MSRC, ImageNet
[92], 2017			super-pixel		normalized Euclidean distances	square loss, cosegmentation with co-localization	quadratic program	SIFT, CNN	MSRC, PASCAL
[32], 2010			pixel			χ^2-kernel based square loss	low-rank	RGB, Gabor, SIFT	Oxford flowers, Weizman horse, MSRC
[93], 2012						χ^2-kernel based soft-max loss	expectation maximization	SIFT	iCoseg, MSRC-v2
[64], 2012				random walk		foreground histogram matching	gradient projection	texture, SIFT, optical flow	iCoseg, self-built [63, 75]
[94], 2016						alternating random walk strategy	spectral method	RGB	iCoseg, clinical images
[55], 2015			superpixel		multiple random walkers, hybrid restart rules	concurrence distribution	repulsive restart rule	RGB, LAB, boundary, SIFT, texton	iCoseg
[95], 2012				hierarchical	k-means clustering, χ^2- and Euclidean distance based edge affinity		normalized cut	LAB, HOG, SURF	MSRC, ImageNet
[96], 2015			pixel	global	local clustering, saliency detection	global clustering of local clusters	Grabcut	LAB, Gabor, position	MSRC, iCoseg, Weizmann Horses
[34], 2017			superpixel	saliency guided	cosine distance, cosine utility function		augmented Lagrangian	SIFT, texton, LAB	iCoseg, Internet
[97], 2019					cosine similarity, multi-view weighting		K-means-like solution	CNN	iCoseg,RGB-D, MSRC
[88], 2011		from multi to binary		Anisotropic diffusion	2D graph diffusion based on Gaussian feature similarity,agglomerative clustering		belief propagation	LAB, texture	MSRC, ImageNet
[89], 2017					3D graph co-diffusion based on Gaussian feature similarity				iCoseg, MSRC, FlicrMFC
[2], 2015	shape		pixel	affinity propagation	coherent point drift based shape registration	common shape pattern discovery	Grabcut	shape	CoShape
[98], 2014	noise shape				aligned and homogenous clusters for visual subcategory construction			CHOG	Internet dataset
[47], 2015	MFC	multi			directed graph clustering		graph cut	RGB, gray	iCoseg, FlickrMFC
[90], 2014			superpixel		ensemble clustering for region proposal detection		α expansion	RGB, LAB, LBP	

number K for multiple foreground cases, and the selection of heat sources for diffusion methods, etc. To alleviate the influences of initialization, Li et al. [90] utilize ensemble clustering to enhance the performances of weak clustering methods.

3.1.3. Saliency-based image cosegmentation

The saliency-based category intends to solve cosegmentation problem with image saliency detection methods, which have been extensively studied for decades [106]. Apparently, saliency regions in images have higher possibility to become foregrounds. There are two ways to combine saliency with image cosegmentation. 1) The first one is to directly consider saliency detection results as prior knowledge or cues. For example, some methods choose saliency region maps as the initial foreground estimation [5,44,96]. Chang et al. [1] involve cosaliency prior into cosegmentation. Rubinstein et al. [43] design a saliency term, which encourages the salient pixels to be plausible foregrounds. 2) The second is to integrate saliency detection with correspondence estimation simultaneously. The related methods are summarized in Table 5. Given images $\{I^i\}_{i=1}^n$ and their subregions $\{v_m^i\}_{m=1}^{p_i}$, the saliency-based cosegmentation methods aim at figuring out the foreground probabilities of subregions $l_k^i, i = 1, \ldots, n, k = 1, \ldots, p_i$, which can be formulated as an energy minimization problem:

$$\min_{l} \underbrace{E^{sal}}_{saliency} + \underbrace{E^{cor}}_{correspondence}, \tag{4}$$

where E^{sal} represents the saliency term and E^{cor} represents the correspondence term. The saliency term and correspondence terms are jointly optimized with each other in order to highlight the common foregrounds from images.

The formulation in Eq. (4) can be designed as low-rank matrix recovery based methods. For instance, in [101] and [102], a low-rank matrix decomposition-based saliency detection process is integrated with logistic regression based discriminative learning. Given feature vectors $\{f_k^i\}_{k=1}^{p_i}$ extracted from subregions, Eq. (4) can be re-written as,

$$\min_{L,S} \underbrace{\sum_{i=1}^{N} \left(\left\| L^i \right\|_* + \lambda \left\| S^i \right\|_1 \right)}_{saliency} + \underbrace{\mu_1 E^{cor}}_{correspondence} + \underbrace{\mu_2 \sum_{i=1}^{n} \sum_{k=1}^{p_i} \left(l_k^i - \alpha^i \left\| S_k^i \right\|_{2,1} \right)^2}_{regularization}, s.t.\ F^i = L^i + S^i,\ i \in N, \tag{5}$$

where the first saliency term (equals to E^{sal}) is constructed based on the low-rank matrix recovery framework. $\| \cdot \|_*$ is the nuclear norm representing the low-rank property of a matrix based on the summary of singular values. $\| \cdot \|_1$ is the L_1-norm representing the sparsity of a matrix, which indicates the saliency of subregions. The feature matrix F^i can be represented as a low-rank matrix (non-salient background) L^i plus a sparse noise matrix (salient foreground) S^i. The correspondence term E^{cor} is defined by logistic-regression-based discriminative learning. Meanwhile, an extra term is added to regularize the relationship between saliency and correspondence, in which S_k^i is the saliency value of the kth subregion

Table 5
Details of saliency-based image cosegmentation methods. Saliency detection and correspondence estimation are jointly optimized [60,100,101,102,103,67,104,105].

paper, year	problem	sub-region	saliency	correspondence	optimization	feature	dataset
[60], 2015	general (multiple)	pixel	geometric mean saliency	dense SIFT matching	GrabCut	SIFT	MSRC, iCoseg, Coseg-Rep
[100], 2015				group saliency propagation			MSRC, iCoseg, Coseg-Rep, Internet Dataset, Weizmann Horses
[101], 2013		superpixel	low-rank matrix ecovery	discriminative learning	inexact Augmented Lagrange Multiplier (ALM)	color, Gabor, steerable pyramid	MSRC-v2, iCoseg
[102], 2015				discriminative learning, feature transformation		color, Gabor, steerable pyramid	MSRC-v2, iCoseg, Caltech101
[103], 2016			saliency co-fusion	feature and saliency similarity	Otsu's method, GrabCut	GIST, SIFT	MSRC, iCoseg, Coseg-Rep
[67], 2017			saliency and objectness maps	similarity propagation	shortest path	RGB, LAB, SIFT	BSDS, PASCAL
[104], 2018			tree structured sparse matrix decomposition	subtree isomorphism	inexact ALM	color, Gabor, steerable pyramids	MSRC, iCoseg, Coseg-Rep, Internet Dataset, FlickerMFC
[105], 2018		pixel, superpixel	saliency fusion with quality measurement	dense correspondence	GrabCut	color, SIFT	MSRC, iCoseg, Coseg-Rep, Internet Dataset, ImageNet

Table 6
Details of deep-model-based image cosegmentation methods. The backbone term represents the pre-trained networks for fine-tuning. The mutual learning term represents the correlation information considered in the Siamese networks. * represents that the deep-model-based methods are supervised rather than weakly supervised due to the training process of deep networks [110,112,113,109,111,112,114,115,116,117].

paper, year	problem	backbone	framework	mutual learning	loss function	dataset
[110]*, 2017	general (multiple)	VGG16 on ImageNet [112]	Deep-dense Conditional Random Field		self-defined, cross entropy	iCoseg, Internet Dataset
[113]*, 2017		FCN on PASCAL-VOC [109]	Siamese Network (Fig. 5)	dense correspondence flows	dense CRF compared with network output	VAP people segmentation dataset
[111]*, 2018		VGG16 on ImageNet [112]		correlation layer	cross entropy	PASCAL VOC 2012, MSRC, iCoseg, Internet Dataset
[114]*, 2018		ILSVRC [115]		approximate nearest neighbor	contrastive loss	MSRC, iCoseg
[116]*, 2018		VGG16 on PASCAL-VOC [117]		attention learner	cross entropy	PASCAL VOC 2012, MSRC, iCoseg

in image I^i, and α^i represents the normalized weights to balance the values. λ, μ_1 and μ_2 are parameters determining the trade-offs among the saliency, correspondence and regularization terms, which are determined by experiments. Ren et al. [104] extend Eq. (5) with tree-structured sparsity and subtree isomorphism.

Eq. (4) can also be formulated as saliency fusion based methods. Assuming that M saliency region maps $S^m, m = 1, \ldots, M$ are extracted for each image, the fused saliency map for each image I^i and each subregion v_k^i can be constructed as follows,

$$S^{fuse}\left(v_k^i\right) = \sum_{m=1}^{M} \underbrace{w(m,i,k)}_{correspondence} \underbrace{S^m\left(v_k^i\right)}_{saliency}. \tag{6}$$

S^{fuse} represents the final co-fused result for cosegmentation, and $w(m, i, k)$ represents the weight for the mth saliency value of subregion v_k^i. The weights w (equal to E^{cor} in Eq. (4)) are learned through analyzing the subregion correspondence. For instance, Jerripothula et al. [60,100] generate geometric mean saliency region maps through fusing saliency region maps with dense SIFT correspondence. Similar saliency co-fusion framework can also be found in [103]. Meanwhile, Huang et al. [67] introduce similarity propagation on saliency information. Jerripothula et al. [105] consider quality measurement system for the fusion of saliency region maps.

The key of the saliency-based category implants feature correspondence into saliency detection frameworks. Compared with other categories, saliency-based methods thoroughly exploit the saliency priors of images. However, the performances of saliency detection methods would directly influence the final cosegmentation results. The common but non-salient foregrounds might be failed to be detected.

3.1.4. Deep-model-based image cosegmentation

Recent years have witnessed the rise of convolutional neural networks (CNNs) in the field of computer vision, such as image classification [107], object detection [108] and semantic segmentation [109]. Some initial attempts have been made to combine image cosegmentation with deep models. The combinations can be understood from two perspectives: 1) utilizing deep features to replace conventional features and 2) building deep frameworks for cosegmentation (Table 6). Quan et al. and Yang et al. extract deep features for cosegmentation. Yuan et al. [110] model a deep-dense conditional random field framework to generate correspondence maps for image cosegmentation. Li et al. [111] construct a Siamese network with mutual correlation layers. Note that the deep-model-based cosegmentation methods involve training data from either cosegmentation benchmarks or other image datasets, which should be regarded as supervised methods. However, rather than learning specific object models, deep cosegmentation networks intend to learn the correspondence among the images.

Almost all the recent deep-model-based cosegmentation methods are based on Siamese networks (Table 6), which share weights among two identical subnetworks. Fig. 5 illustrates the end-to-end flow chart of a Siamese network. The encoder blocks provide high-level intra-image features, the mutual layers ensure the exploration of inter-image correspondence, and decoder blocks mix the intra- and inter-information together to obtain refined cosegmen-

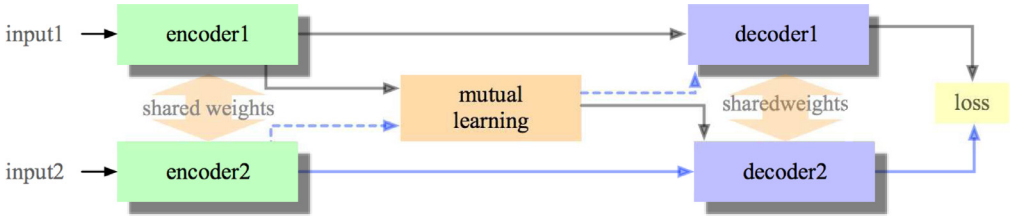


Fig. 5. The basic framework of deep Siamese network with mutual learning layers for cosegmentation.

tation results. More specifically, Choi et al. [113] construct Siamese network for the dense correspondence estimation between multi-spectral images. Mukherjee et al. [114] feed Siamese network with object proposals to obtain high-level features, then use the features to build approximate nearest neighbor library for cosegmentation. Chen et al. [116] leverage channel-wise attention into the Siamese structure to enhance the semantic information.

Meanwhile, considering the lack of large-scale training data for cosegmentation, pre-trained backbone networks on large datasets like VGG16 on ImageNet [112] are introduced as the initializations of the frameworks. The loss functions are also critical for the back propagation of deep models. Other than classic cross entropy functions, contrastive loss function [114] and self-defined loss function [110] have also been considered. Choi et al. [113] define the loss functions in a weakly-supervised and iterative manner. This method directly inputs the output results of the networks into CRF models, and calculates the losses between the output results and the CRF ones. In a brief, the deep-model-based methods have blazed a new trial for image cosegmentation. However, considering the weakly supervised nature of cosegmentation, techniques like transfer learning, data augmentation would be necessary for the construction of deep models.

In addition to the aforementioned methods, some outstanding image cosegmentation methods cannot be classified into the major categories. For instance, Wang et al. [59,118] introduce a consistent functional map to obtain graph correspondence and Meng et al. [119] propose the region-based active contour model which integrates with a rewarding strategy under level-set based energy minimization formulation.

3.2. Video cosegmentation methodology taxonomy

3.2.1. Graph-based video cosegmentation

As with graph-based image cosegmentation (Section 3.1.1), the graph-based video cosegmentation considers the segmentation problem as a labelling problem based on graph structures. Given a set of n videos $\{\mathcal{V}^i\}_{i=1}^n$, each video contains J^i frames $\{\mathcal{F}_j^i\}_{j=1}^{J^i}$. The graph constructed by the existing video cosegmentation methods can be divided into two kinds (represented by G^1 and G^2) depending on their ways to encode temporal information. The first kind of graph $G^1 = (X_{j,r}^i, Ed^1, L^1)$ inherits the spirit of graph-based image cosegmentation, whose vertices $\{X_{j,r}^i\}_{r=1}^{p_j}$ are still 2D subregions extracted from video frames such as pixels, superpixels and region proposals. In order to encode the temporal information in videos, G^1 includes 3 kinds of edges Ed^1: the spatial edges between the subregions in the same frame, temporal edges across the frames in a same video and correspondence edges across different videos. Labels $L^1 = \{l_{j,r}^i\}$, where $i = 1, \ldots, n, j = 1, \ldots, J$ and $r = 1, \ldots, p_i$, are generated for each 2D subregions as the cosegmentation results. The second kind of graph $G^2 = (Y_r^i, Ed^2, L^2)$ defines the vertices $\{Y_r^i\}_{r=1}^{q_i}$ as video 3D subregions rather than image 2D subregions. G^2 involves the temporal information during the generation of vertices. The video 3D subregions can be represented by supervoxels (the video analog to superpixels [120]), trajectories (a series of points tracked over multiple frames [62]) and tracklets (a series of detected/tracked object regions across multiple frames [49]). G^2 includes 2 kinds of edges Ed^2: spatiotemporal edges of the 3D subregions in a same video and correspondence edges across different videos. The labels $L^2 = \{l_r^i\}_{r=1}^{q_i}$ represent the cosegmentation results. As with image cosegmentation methods, the labels can be either binary $l \in \{1, 0\}$ or multiple $l \in \{1, \ldots, m\}$, which are determined by problem settings. In video cosegmentation methods, initialization schemes are used as a pre-processing step to reduce the number of vertices and to assign initial labels to them, e.g., relative motion cues [36,68], saliency hints [40,41], and objectness [46,121], etc. Details of the following mentioned graph-based methods are summarized in Table 7.

As aforementioned above, MRF-based model has showed its potentials on inferring common foreground regions. The basic formulation of MRF-based video cosegmentation is akin to that of MRF-based image cosegmentation, albeit dealing with spatiotemporal correspondence, which can be briefly formulated as an energy minimization function as follows,

$$\min_{l} \overbrace{\underbrace{E^{data} + E_{spatial}^{smooth}}_{G^1} + E_{temporal}^{smooth}}^{G^2}. \tag{7}$$

E^{data} is a data term and E^{smooth} are the smoothness terms. The data term E^{data} measures the labelling costs of each vertex, which can be defined as the likelihood distributions of subregion features using logistic regressors [35], discriminative probability maps based on spatiotemporal auto-context model [46] or GMM models based on cosaliency [62]. The smoothness terms keep labels of neighboring vertices consistent. For the graph type G^1, the smoothness terms $E_{spatial}^{smooth}$ and $E_{temporal}^{smooth}$ measure respectively spatial and temporal consistency of neighboring subregions. Wang et al. [40] use spatiotemporal SIFT flow and color, while Wang et al. [46] utilize color and context features to measure temporal and spatial label consistency. For the graph type G^2, the smoothness term $E_{spatial}^{smooth}$ represents the spatial consistency of neighboring video subregions. Chen et al. [36] smooth the labels of neighboring supervoxels based on motion similarity. Guo et al. [62] use Euclidean distance and velocity estimate of trajectories for it. MRF-based model exploits temporal and spatial correspondence information in G^1 and G^2 differently. G^1 is composed of 2D subregions extracted from video frames, which does not involve temporal information. Thus, the methods based on G^1 depend on the inter-frame and across-video edges to encode the temporal correspondence. In contrast, G^2 encodes the temporal correspondence in vertices with 3D video subregions. Thus, G^2 pays more emphasis on E^{data} to explore the spatiotemporal correspondence.

Conditional random field (CRF) is another undirected probabilistic graphical model used for cosegmentation. Compared with MRF, which is a generative model, CRF is a discriminative model weakening the demand for the estimation of the prior probability distribution of the labels. CRF based cosegmentation can be formulated

Table 7
Details of graph-based video cosegmentation methods. ① and ② represent two types of graph structures: G^1 and G^2 separately. For the MRF models, E refers to E^{data} and E^{smooth}. For CRF, E refers to E^{intra}, E^{inter} and E^{across}. For maximal weight clique, E refers to the node/edge weight assignments of the graphs [46,35,40,36,62,122,123,124,41,73,125,68,69,50,49,8].

paper year	problem	graph	L	V	initialization	E: intra frame	E: inter frame	E: across video	optimization	feature	dataset
MRF											
[46], 2017	noise	complete ①	binary	super-pixel	objectness saliency	GMM	spatiotemporal auto-context feature similarity		Spatial-MILBoost	context	segTrack, MOViCS, XJTU-stevens, ViCoSS
[35], 2012	general (multiple)	complete ①	binary	region, tube	objectness saliency	logistic regression	logistic regression, SVM		graph cut	3D HOG, LBP	Chroma, segTrack
[40], 2015	noise	complete ①	binary	pixel	saliency	GMM	spatiotemporal SIFT flow, optical flow, GMM		graph cut	SIFT, LBP	ViCoSS
[36], 2012	general (pair)	complete ②	binary	super-voxel	motion	dense optical flow, relative motion similarity		GMM	graph cut	color, MRS4 [122]	-
[62], 2013	motion	complete ②	binary	trajectory	motion	2D motion GMM		cosaliency, clustering, matching	graph cut	motion boundary histogram	Chroma, UCF 50
CRF											
[123], 2015	general(multiple)	hierarchical①	binary	pixel, super-pixel,-voxel	stream GBH [124]	GMM, JointBoost, hierarchical labelling consistency			graph cut	color, texton, SIFT, LBP	MOViCS
[41], 2017	general(multiple)	complete ①	binary	proposal streams	saliency	cosaliency	color histogram similarity		graph cut	color, magnitude	self-built, Safari
[73], 2015	general(multiple)	complete ①	binary	proposal streams	objectness, motion, TCS [125], overlap	objectness, saliency, optical flow	color and shape histogram similarities		TRW-S	color, HOG	ObMiC, MOViCS
[68], 2014	MFC	complete ①	multiple	region proposal	objectness, motion, saliency, repetition	objectness, saliency, optical flow	color histogram similarity, size and location consistency	color histogram similarity	iterated conditional modes	color, motion	MOViCS, self-built
[69], 2014	MFC	multi-state selection①	multiple	region proposal	region proposals	objectness, saliency, optical flow	color histogram similarity, overlap	color histogram, shape similarity	TRW-S	color, HOG	MOViCS
[50], 2015	MFC	multi-state selection①	multiple	region proposal	region proposals	objectness, saliency, optical flow	color histogram similarity, overlap	color histogram, shape similarity	TRW-S	color, HOG	MOViCS, self-built
Maximum Weight Clique											
[49], 2014	MFC	complete ②	multiple	tracklet	region proposal	optical flow, objectness, non-maxima suppression		color histogram, shape similarity	modified Bron-Kerbosch	appearance, location, shape	MOViCS, safari
[8], 2015	MFC	bipartite ②	multiple	shot-level features	feature mapping with k-means clustering of frames			co-clustering, maximal biclique	linear programing	CENTRIST, D-SIFT, HSV	CMU-Mocap, self-built

as follow,

$$\min_{l} \underbrace{E^{\text{intra}} + E^{inter} + E^{across}}_{G^1}, \tag{8}$$

where E^{intra}, E^{inter} and E^{across} are determined respectively by intra-frame, inter-frame and across-video correspondence. As shown in Table 7, all the CRF methods use the graph type G^1, which is constructed by 2D subregions extracted from video frames. E^{intra} designs for estimating foreground probability of subregions within same frames. More specifically, E^{intra} can be defined as $E^{\text{intra}} = \sum_i \sum_j \sum_r \psi^{\text{intra}}(l^i_{j,r})$, in which $\psi^{intra}(\cdot)$ evaluates the foreground probability of subregions. Fu et al. [50,69] combine objectness, optical flow, and saliency detection results to determine the foreground likelihood of subregions. Wang et al. [41] generate cosaliency region maps and Guo et al. [123] use dual probabilistic models: i.e., JointBoost and GMMs for it. E^{inter} represents the label consistency between adjacent frames $\mathscr{F}^i_j$ and $\mathscr{F}'^i_j$ in a same video $\mathscr{V}_i$, which can be defined as $E^{inter} = \sum_i \sum_{(j,j')} \sum_{(r,r')} \psi^{\text{inter}}(l^i_{j,r}, l^i_{j',r'})$, where $\psi^{inter}(\cdot)$ measures the correspondence among subregions in the adjacent frames. For example, ψ^{inter} is defined as l_2-norm distance between color features of subregions in [41], $\mathcal{X}^2$ distance based color histogram similarity in [69], and boundary strength plus optical flow in [68]. E^{across} exploits the correspondence among videos $\mathscr{V}_i$ and $\mathscr{V}'_i$. More clearly, it can be defined as $E^{across} = \sum_{(i,i')} \sum_{(j,m)} \sum_{(r,k)} \psi^{across}(l^i_{j,r}, l^{i'}_{m,k})$, where ψ^{across} measures subregion similarity among different videos. For instance, Fu et al. [50,69] utilize color histogram and shape similarity to compute ψ^{across}. Compared with MRF, the CRF model exploits the temporal correspondence through the inter-frame and across-video terms. Meanwhile, the CRF model is more flexible for imposing varied correspondence constraints. In addition to E^{intra}, E^{inter} and E^{across}, some other coherent information among videos can also be added to the formulation of CRF. For example, a diversity term is defined in [50] to encode the independence of the multiple-state selection graph, and a smoothness term is given in [41] to encourage label consistency.

The maximum-weight-clique model can also be used for cosegmentation [8,49], which pays attention on the weights of graph vertices/edges. A maximum clique represents the largest subgroup of vertices fully connected with each other. Each vertex in the maximum clique is directly connected with all other vertices. Under this approach, the cosegmentation problem is solved by determining the maximum weighted subgraph, which represents the potential common foregrounds. Zhang et al. [49] model the video cosegmentation problem as a maximum edge-weight clique problem based on the graph of video tracklets. Chu et al. [8] utilize maximal biclique based video cosegmentation to enhance the performance of video co-summarization.

In addition to segment single common foreground, some researchers consider MFC video cosegmentation. There are two strategies for MFC problem in videos. 1) To segment multiple common foregrounds, some methods firstly find out vertices with the highest common foreground probability and then remove all these vertices and repeat the cosegmentation process [49,68]. 2) Some other methods consider multiple labelling models and directly divide graph structures into multiple subgraphs [69].

Table 8
Details of clustering-based video cosegmentation methods. K represents the number of clusters. Intra-frame, inter-frame, and across-video correspondence are listed for each method [45,70,126].

paper year	problem	K	sub-region	clustering	intra-frame	inter-frame	across-video	optimization	feature	dataset
[45], 2016	general (multiple)	binary	region proposal	affinity propagation	optical flow, boundary occlusion		cosaliency, color histogram similarity	loopy belief prorogation	color	MOViCS, new video
[70] ,2017	MFC	multi	super-voxel	hierarchical	objectness, color gradient		dissimilarity measure	observation scale based hierarchical segmentation		ObMiC, MOViCS
[126] ,2014			super-pixel	iterative constrained	insided-outside map, motion saliency	optical flow, pairwise constraints	SVM, color and motion similarity	graph cut		CFViCS, MOViCS

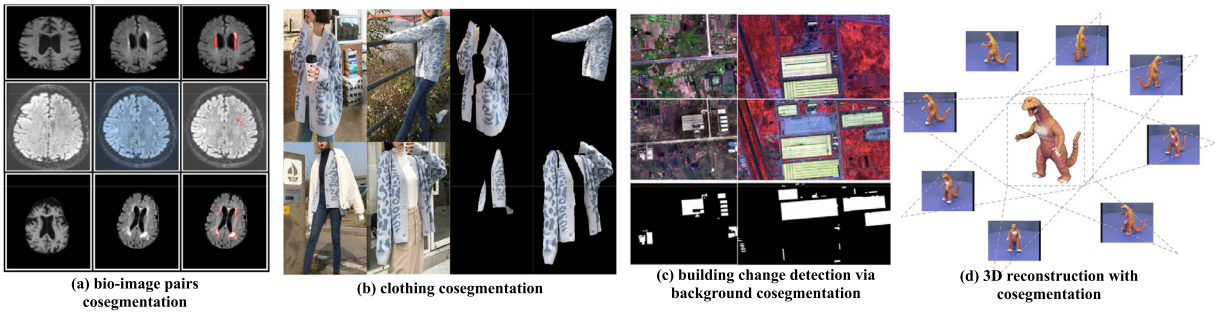


Fig. 6. Examples for the application of image cosegmentation. (a) biomedical image cosegmentation [75]. The first two images in each row represent the input images and the last image in each row represent the cosegmentation result; (b) clothing cosegmentation for shopping image retrieval (images downloaded from an online shopping website); (c) building change detection with the cosegmentation of common background[134]; (d) cosegmentation with images taken from different perspectives of objects for 3D reconstruction.

3.2.2. Clustering-based video cosegmentation

As with image cosegmentation (Section 3.1.2), clustering-based video cosegmentation divides subregions of videos into groups. The subregions can be 2D frame subregions or 3D video subregions discussed before. Different from image cosegmentation, temporal correspondence is taken into consideration when performing clustering. For example, Zhang et al. [45] propose affinity propagation clustering based on χ^2 distance between optical flow histograms and a cluster selection method to prune irrelevant frames. Rodrigues et al. [70] utilize graph-based hierarchical clustering with supervoxel correspondence for video cosegmentation. Guo et al. [126] employ iterative constrained clustering as a filter to remove incorrect correspondence and estimate the number of foreground categories. More details are summarized in Table 8. Although clustering-based cosegmentation is still under-developed, it has shown its potentials on MFC problems, which contain multiple common foregrounds in videos. In addition to the algorithmic categories mentioned before, there are some other well-designed video cosegmentation methods which are hard to be classified into any of the categories but worth to be studied, including the integer programming model [127], distant-dependent Chinese Restaurant Processes [71], absorbing Markov chain [9], and particle swan optimization [128].

4. Applications

The main applications of image cosegmentation are listed as follows. **Biomedical Imaging** usually processes images with the same types of foreground objects, e.g., cancer cells. Manually labelling all the data requires precise and careful work. Therefore, image cosegmentation becomes a possible solution (Fig. 6(a)). For instance, Mukherjee et al. [30] and Hochbaum et al. [75] use cosegmentation to detect small pathologies in brain image volumes. Wang et al. [94] implement image cosegmentation on the endocardium extraction from an echo cardiac image sequence. **Image Retrieval, Retargeting, and Summarization** are also potential applications of image cosegmentation, which have rough class labels but lack fine pixel-level annotations. For example, Rother et al. [28] study the possibility of improving image retrieval performance with image cosegmentation. Lin et al. [129] utilize image cosegmentation to assist retargeting of stereoscopic images. Batra et al. [6] combine image summarization with image cosegmentation. **Clothing Cosegmentation** aims at extracting consistent clothing from photos (Fig. 6(b)) for online shopping. Co-delineating the clothing regions in images can improve retrieval performances and provide more satisfying searching results for customers. For instance, the work in [130] proposes a cosegmentation model based on automatically learned global clothing mask. Liang et al. [131] model clothing cosegmentation as a Gaussian-mixture-model. A co-parsing model is proposed in [132] combining cosegmentation and SVM-based image co-labelling. **Remote Sensing** images are usually lack of manual annotations, which can be processed with cosegmentation. For instances, Champion et al. [133] and Xiao et al. [134] use image cosegmentation to detect building change(Fig. 6(c)). A multi-temporal cosegmentation strategy is proposed based on the spatial correspondence of changed objects. Chen et al. [135] introduce a clustering-based cosegmentation method to detect individual buildings from blocks. **3D Model Reconstruction** reconstructs objects' 3D structures through images captured from different perspectives of objects, which can be assisted by image cosegmentation (Fig. 6(d)). For examples, Kowdle et al. [136] integrate interactive image cosegmentation [6] with a shape-from-approach algorithm to extract objects for 3D reconstruction. Mustafa et al. [137] combine semantically coherent cosegmentation and reconstruction of dynamic scenes into a joint formulation.

Video cosegmentation also has a wide range of applications. **Video Summarization** aims at automatically selecting key frames

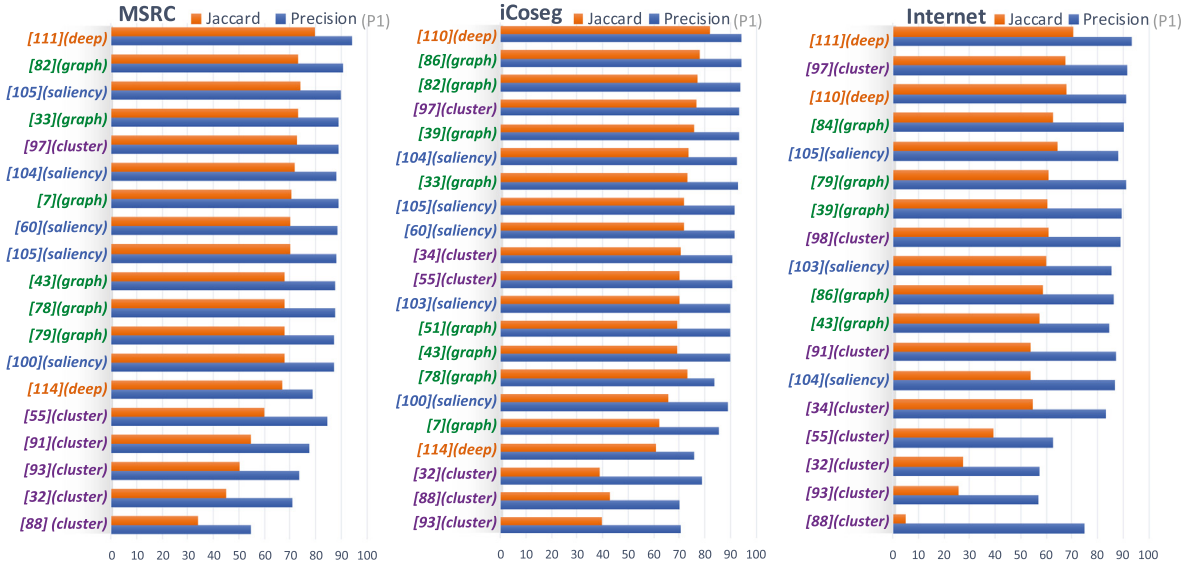


Fig. 7. Evaluation results of the selected image cosegmentation methods on MSRC, iCoseg and Internet datasets. Two kinds of metrics are considered: Jaccard similarity and Precision. The results have been sorted by the sum of the 2 metric values from high (top) to low (bottom).

from a set of videos and summarizing the main topics [138]. To avoid annotation bias and reduce manual work, video cosegmentation can be used to extract common visual concepts for video summarization [8]. **Content-based Video Retrieval** [139–141] aims at indexing and searching videos from a large database. Compared with content-based image retrieval, video retrieval needs to handle diverse video contents and more changeable visual themes. Video cosegmentation can extract common shots from videos, which can help to improve the performance of video retrieval and lead to a better understanding of videos. **Action Recognition** aims at identifying actions in videos [142]. One of the recent trends is utilizing common motion detection [9,128] to assist action recognition. More clearly, cosegmentation is used to simultaneously find out common activities among videos under unsupervised conditions. It can reduce manpower cost for constructing action recognition databases.

5. Benchmark datasets

As for general image cosegmentation datasets, the **iCoseg dataset**[1] [6] (634 images, 38 classes) is collected from Flickr online photo collection with pixel-level ground truth (Fig. 2(a)). With distinct foregrounds and simple backgrounds, this dataset is suitable for measuring the basic functionality of image cosegmentation algorithms. The **MSRC dataset**[2] [143] (591 images, 23 classes) and its selected version MSRC-v2 (420 images,14 classes) are also classic benchmark dataset for image cosegmentation methods (Fig. 2(b)). Compared with iCoseg, MSRC is more complicated with considering different individuals of a same species into a same class. For example, the bird class in MSRC contains images of swans, pigeons and ducks, etc.

As for shape similarity image datasets, the **Coseg-rep dataset**[3] [51] (572 images, 23 classes) is collected (Fig. 2(c)), which includes a special category with images of repetitive patterns. The **Coseg-INCT** [58] (291 images, 12 categories) is an expanded version of Coseg-rep with more repetitive instances contained in each image. The **ETHZ Shape Classes** [144,145] (255 images, 5 classes) have objects with various scales and considerable variation of intra-class shapes. The **CoShape dataset** [2] (703 images, 15 classes) is a relatively large dataset constructed through collecting images from other classic datasets. More recently, the **CO-SKEL dataset** [52] (353 images, 26 categories) is developed for co-skeletonization with skeleton ground truth masks.

As for large image datasets, the **internet dataset**[4] [43] (4,337 car images, 6831 horse images, 4542 airplane images) is automatically downloaded through querying the Wikipedia (Fig. 2(c)). A portion of the images have been labelled manually with the LabelMe annotation toolbox [146]. This dataset contains noisy images in each visual class, therefore, it can be used to evaluate image cosegmentation methods with noisy-screening functions. The **Caltech 101 dataset**[5] [147] (101 classes, 40 to 800 images per class) and the **Caltech UCSD Bird dataset**[6] [148] (11,788 images, 200 classes) contain varying objects and clutter backgrounds. The **ImageNet dataset**[7] [149] (21841 non-empty synsets, around 500 to1000 images per synset) is also a large image dataset including diverse appearances, positions, and viewpoints. Existing image cosegmentation methods usually select a small subgroup of ImageNet for evaluation [88,91].

As for multiple-foreground image datasets, the **FlickrMFC dataset**[8] [85](14 groups, 12 to 20 images in each group) is constructed (Fig. 2), which contains two or more subjects in each class. The **PASCAL-VOC 2010 dataset**[9] [150] (23,374 images, 20 classes) includes large intra-class variability and complicated backgrounds, which is also suitable for evaluating methods designed for MFC problems. In addition, there exist some other cosegmentation datasets, which are also useful to evaluate the image cosegmentation problems: the **Oxford flowers dataset**[10], the **Weizman horses dataset**[11] [151], and the **RGB-D cosegmentation dataset**[12] [44].

[1] available at chenlab.ece.cornell.edu/projects/touch-coseg/.
[2] available at www.microsoft.com/en-us/research.
[3] available at www.stat.ucla.edujifeng.dai/research/CosegmentationCosketch.html.
[4] available at people.csail.mit.edu/mrub/ObjectDiscovery.
[5] available at www.vision.caltech.edu/Image_Datasets/Caltech101/.
[6] available at www.vision.caltech.edu/visipedia/CUB-200-2011.html.
[7] available at www.image-net.org/about-stats.
[8] available at vision.snu.ac.kr/ ~ gunhee/r_mfc.html.
[9] available at host.robots.ox.ac.uk/pascal/VOC/voc2010/.
[10] available at www.robots.ox.ac.ukvgg/data/flowers/17/.
[11] available at www.msri.org/people/members/eranb/.
[12] available at sites.google.com/site/huazhufu/home/rgbdseg.

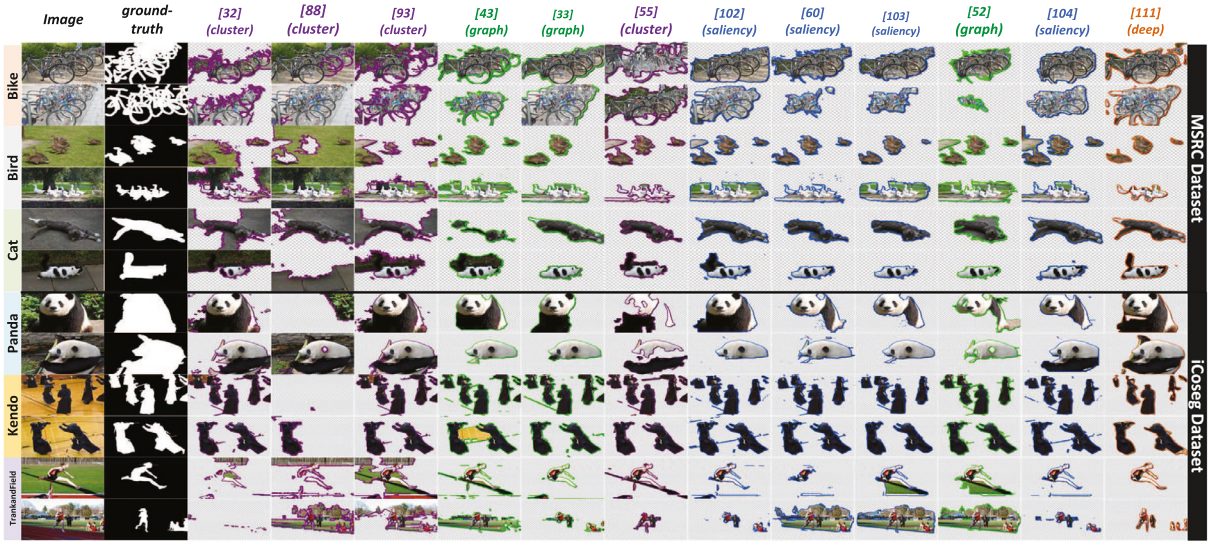


Fig. 8. Image cosegmentation results of the selected image cosegmentation methods on MSRC and iCoseg datasets. The results have been ranked by the year of publication from past (left) to recent (right).

As for general video cosegmentation datasets, the **segTrack dataset**[13] [152] (6 videos, 21 to 70 frames in each video) contains both single-object and multiple-object videos, in which all frames are fully annotated with pixel-wise labels. The **segTrack v2 dataset** [27] is the expanded version of SegTrack, which provides additional annotations and 8 new videos. All the videos in these datasets are carefully selected to include various challenges, such as appearance deformation, random occlusion, and slow-motion, etc. The **MOViCS dataset**[14] [71] (4 video sets, 11 videos, 514 frames), as an abbreviation of Multi-Object Video Cosegmentation dataset, is constructed for multiple foreground video cosegmentation research. The **Safari dataset**[15] [49] is a combination of some videos from MOViCS and some others, containing 5 categories and 9 animal videos. Similarly, the **New Sports dataset** [68] only focusses on the cosegmentation of sports videos, which is collected from Youtube with manually labeled ground truth.

As for large video cosegmentation datasets, the **ViCoSS dataset**[16] [40] (12 categories, 30 videos, 20 to 125 frames in each video) is designed for studying single-foreground video cosegmentation with irrelevant frames. This dataset is suitable for both general (multiple) and noise screening problems of video cosegmentation. The **XJTU-Stevens video cosegmentation and classification dataset** [46] (10 categories, 101 videos, 13,398 frames) is a new dataset established with a large number of videos, more irrelevant frames, and more challenging coherent objects. Pixel-level manual annotations are provided for all relevant frames.

As for motion-based video cosegmentation datasets, the **80-pair dataset** [62] is composed of 50 pairs of video sequences from the UCF50 dataset [153] of human actions and 30 pairs from BBC animal documentaries of animal actions. The **YouTube co-activity dataset**[17] [9] contains 11 activities with at least 10 Youtube videos each and manual annotations for individual frames.

6. Evaluation metrics

Jaccard similarity (J), also known as **intersection over union (IOU)**,is the most widely used evaluation metrics for image and video cosegmentation. The Jaccard similarity is defined as $J = \frac{|A^p \bigcap A^g|}{|A^p \bigcup A^g|}$, where A^p is a resultant common foreground in an image or a video frame, A^g is the corresponding ground truth foreground and $|.|$ counts the pixel numbers.

Precision (P2), recall (R) and F-measure (F)**,** which are classical metrics for image recognition and computer vision [154], are also employed for evaluating image and video cosegmentation methods. Rather than the final cosegmentation results, these metrics are usually based on the final heat-map results, which represent the common foreground probabilities. Given a normalized heat-map $H^p \in [0, 1]$, $H^p(t)$ represents the common foreground regions at a certain threshold t. The threshold changes continuously from 0 to 1 to obtain a series of $H^p(t)$. Given A^g as the ground truth, $P2(t)$ is defined as the ratio of the correctly segmented results to the output foreground results at the threshold t: $P2(t) = \frac{|H^p(t) \bigcap A^g|}{|H^p(t)|}$, where A^g represents the ground truth. Meanwhile, the Recall (R) represents the ratio of precisely segmented common foreground regions to the ground truth: $R(t) = \frac{|H^p(t) \bigcap A^g|}{|A^g|}$. The F-measure ($F$) is the weighted harmonic mean of precision and recall: $F = 2 \times \frac{P2 \times R}{P2 + R}$. Higher value of F-measure represents better cosegmentation results.

Precision (P1) is also a popular measurement for image cosegmentation, which indicates the percentage of correctly labeled pixels in both foregrounds and backgrounds. The precision can be calculated through $P1 = \frac{|A^p \bigcap A^g| + |B^p \bigcap B^g|}{|A^g \bigcup B^g|}$, where B^p is the resultant background and the B^g is the corresponding ground-truth background in images. Meanwhile, the **average per-frame pixel error (error)** is another widely used measurement for video cosegmentation, which is defined as the ratio of falsely labelled pixels in both foreground and backgrounds: $error = \frac{|XOR(Seg, GT)|}{N}$, where Seg is the set of the frame-by-frame common foreground results of all videos, GT is the ground truth, and N is the total number of frames in videos. Smaller error refers to better performance.

To compare the performances of the methods from different methodology categories, some cosegmentation methods are selected for evaluation based on two main reasons: (1) The selected methods are either classic or state-of-the-art. (2) Quantitative results for the classic evaluation metrics are available, which can be obtained from the papers, calculated by given codes or validated by some other methods. For image cosegmentation, 25 methods are evaluated on 3 benchmark datasets based on Jaccard simi-

[13] available at cpl.cc.gatech.edu/projects/SegTrack/.
[14] available at sites.google.com/site/walonchiu/projects/cosegmentation.
[15] available at crcv.ucf.edu/projects/video_object_cosegmentation/.
[16] available at github.com/shenjianbing/Robust-video-object-co-segmentation.
[17] available at cvlab.postech.ac.kr/coactivity/.

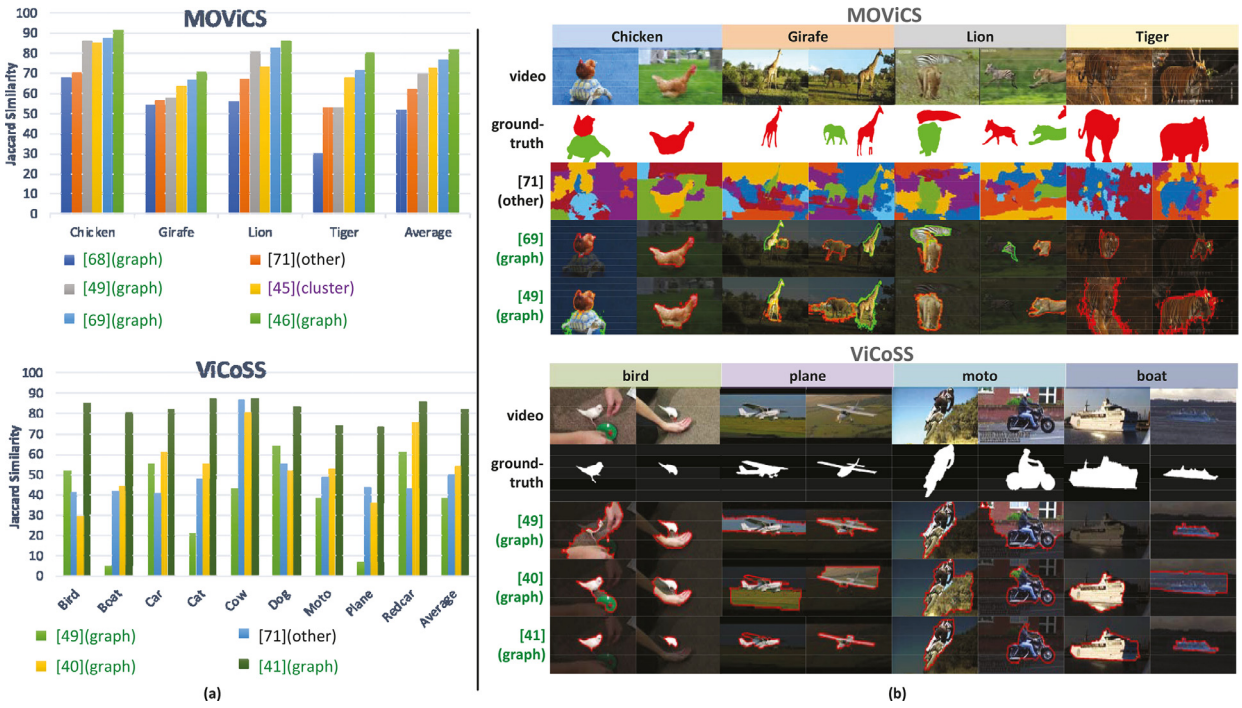


Fig. 9. (a) Evaluation results of the selected video cosegmentation methods on MOViCS and ViCoSS datasets based on Jaccard similarity. (b) Qualitative results of the selected video cosegmentation methods on MOViCS and ViCoSS datasets. For example, for the Chicken class in MOViCS, the illustrated images represent video frames extracted from two videos, which both contain the Chicken objects.

larity and Precision metrics (Fig. 7). These methods cover all the 4 main methodology categories. The deep-model-based methods perform the best on all the 3 datasets, which benefit from the high-level features. The graph-based category is the second best, which exploits diverse geometrical information for cosegmentation. The saliency-based category gets the third place, which integrates cosegmentation with saliency detection. Although the average performances of the clustering-based methods are not the best, some of its specific methods still can perform very well [97]. For video cosegmentation, 8 methods are evaluated on 2 benchmark datasets based on Jaccard similarity (Fig. 9(a)). The graph-based category, which is the most popular model for video cosegmentation, outperform the others for video cosegmentation. More illustrations of the image and video cosegmentation results can be found in Figs. 8 and 9(b).

7. Conclusion and promising research directions

This paper provides a systematic survey on image and video cosegmentation. The problems and challenges of cosegmentation are described; methodology categories are given and potential applications and benchmark datasets are summarized. Through a common taxonomy for both image and video cosegmentation, the similarities and differences between the two related research problems are pointed out. The state-of-the-art methods are evaluated and discussed. Some existing challenges and future research directions are listed below.

Cosegmentation with Deep Models: As mentioned in Section 3.1.4, some researchers have considered to use deep learning for image cosegmentation. However, the research direction that exploits the representation power of deep learning for video cosegmentation has not been fully explored. There are some related works in other areas. For example, Yoon et al. [155] construct a pixel-level matching network for video object segmentation. These works can serve as references for developing video cosegmentation methods based on deep learning.

Cosegmentation on More Challenging Datasets: Large-scale datasets with diverse image/video quality and various common foregrounds are essential for developing and evaluating reliable cosegmentation methods. These datasets are even more critical for the development of cosegmentation methods based on deep learning. Although researchers have constructed more and more image datasets for image cosegmentation, their size is still significantly smaller than well-known datasets for other computer vision problems, e.g., image classification and object detection. More seriously, the diversity of common foregrounds is limited by the number of classes in the existing datasets. As for video cosegmentation, the number of large-scale benchmark datasets and the number of different types of common foregrounds are both limited. Fortunately, a new benchmark dataset named DAVIS is collected for video object segmentation [156], which have diverse challenges like occlusions, motion blur, and appearance changes and it can be further developed for establishing a large video cosegmentation dataset. Moreover, the RGB-D based action recognition datasets [157] can also be considered for video cosegmentation.

Cosegmentation under Unconstrained Conditions: Most of the existing cosegmentation methods are weakly supervised, which rely on the existence of common foregrounds in images and videos. To remove this requirement and apply cosegmentation under more unconstrained conditions, a new cosegmentation approach, which is named unconstrained cosegmentation is needed. More clearly, given a large set of diverse images and videos, unconstrained cosegmentation is first to group images or videos with common foregrounds together and then delineate pixel-level outlines of common foreground belonging to the same group. Inspirations can be found in [158], which utilizes a knowledge-based topic model for object discovery and localization from images. As for video cosegmentation, common video event saliency discovery might be a thought-provoking work [159], which aims at finding out common events rather than objects from videos. As an analogy, video event cosegmentation is another potential direction, which deals with common events and common foregrounds simultaneously.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The paper is supported by the State Key Program and the Joint Funds of the National Natural Science Foundation of China (Grant No. 61836009 and No. U1701267). This work done by A.W.K. Kong is partially supported by the Ministry of Education, Singapore through Academic Research Fund Tier 2, MOE2016-T2-1-042(S).

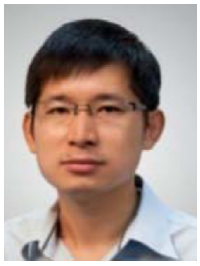
References

[1] K.Y. Chang, T.L. Liu, S.H. Lai, From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model, in: CVPR, 42, 2011, pp. 2129–2136.
[2] W. Tao, K. Li, K. Sun, Sacoseg: object cosegmentation by shape conformability, IEEE TIP 24 (3) (2015) 943–955.
[3] Y. Chai, E. Rahtu, V. Lempitsky, L.V. Gool, Zisserman, Tricos: a tri-level class-discriminative co-segmentation method for image classification, in: ECCV, 2012, pp. 794–807.
[4] Y. Chai, V. Lempitsky, A. Zisserman, BiCos: A Bi-level co-segmentation method for image classification, in: ICCV, 2011, pp. 2579–2586.
[5] S.D. Jain, K. Grauman, Which image pairs will cosegment well? predicting partners for cosegmentation, in: ACCV, 9005, 2014, pp. 175–190.
[6] D. Batra, A. Kowdle, D. Parikh, J. Luo, T. Chen, iCoseg: Interactive co-segmentation with intelligent scribble guidance, in: CVPR, 2010, pp. 3169–3176.
[7] S. Vicente, C. Rother, V. Kolmogorov, Object cosegmentation, in: CVPR, 42, 2011, pp. 2217–2224.
[8] W.S. Chu, Y. Song, A. Jaimes, Video co-summarization: Video summarization by visual co-occurrence, in: CVPR, 2015, pp. 3584–3592.
[9] D. Yeo, B. Han, J.H. Han, Unsupervised co-activity detection from multiple videos using absorbing Markov chain, in: AAAI, 2016, pp. 3662–3668.
[10] H. Zhu, F. Meng, J. Cai, S. Lu, Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation, J. Vis. Commun. Image Represent. 34 (2) (2015) 12–27.
[11] V.B. Biradar, D. GV, V. RH, Review of image co segmentation techniques, Int. J. Sci. Res.Dev. 3 (5) (2015) 981–984.
[12] D. Tri, A. Sheeraz, Y. Shu, Survey: recent trends and techniques in image co-segmentation challenges, issues and its applications, Int. J. Comput. Sci.Softw. Eng. 6 (5) (2017) 99–114.
[13] Y.-J. Zhang, An Overview of Image and Video Segmentation in the Last 40 Years, in: Advances in Image and Video Segmentation, IGI Global, 2006, pp. 1–16.
[14] B. Peng, L. Zhang, D. Zhang, A survey of graph theoretical approaches to image segmentation, Pattern Recognit. 46 (3) (2013) 1020–1038.
[15] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: a high--definition ground truth database, Pattern Recognit. Lett. 30 (2) (2009) 88–97.
[16] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, S. Yan, Learning to segment with image-level annotations, Pattern Recognit. 59 (2016) 234–244.
[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE TPAMI 40 (4) (2017) 834–848.
[18] Y. Wang, J. Liu, Y. Li, J. Fu, M. Xu, H. Lu, Hierarchically supervised deconvolutional network for semantic video segmentation, Pattern Recognit. 64 (2017) 437–445.
[19] M. Ren, R.S. Zemel, End-to-end instance segmentation with recurrent attention, in: CVPR, 2017, pp. 6656–6664.
[20] A. Sasithradevi, S.M.M. Roomi, M. Mareeswari, Video object segmentation: a survey, in: International Conference on Communication and Electronics Systems, 2017, pp. 1–5.
[21] D. Zhang, H. Fu, J. Han, A. Borji, X. Li, A review of co-saliency detection algorithms: fundamentals, applications, and challenges, ACM TIST 9 (4) (2018) 38.
[22] H. Song, Z. Liu, Y. Xie, L. Wu, M. Huang, RGBD Co-saliency detection via bagging-based clustering, Signal Process. Lett. 23 (12) (2016) 1722–1726.
[23] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, C. Hou, An iterative co-saliency framework for RGBD images, IEEE TCYB 49 (1) (2017) 233–246.
[24] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, C. Hou, Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation, IEEE TIP 27 (2) (2017) 568–579.
[25] D. Zhang, J. Han, J. Han, L. Shao, Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining, IEEE Trans. Neural Netw. 27 (6) (2016) 1163–1176.
[26] I. Endres, D. Hoiem, Category-independent object proposals with diverse ranking, IEEE TPAMI 36 (2) (2014) 222–234.
[27] F. Li, T. Kim, A. Humayun, D. Tsai, J.M. Rehg, Video segmentation by tracking many figure-ground segments, in: ICCV, 2014, pp. 2192–2199.
[28] C. Rother, T. Minka, A. Blake, V. Kolmogorov, Cosegmentation of image pairs by histogram matching - incorporating a global constraint into MRFs, in: CVPR, 2006, pp. 993–1000.
[29] T. Riklin-Raviv, N. Sochen, N. Kiryati, Shape-based mutual segmentation, IJCV 79 (3) (2008) 231–245.
[30] L. Mukherjee, V. Singh, C.R. Dyer, Half-integrality based algorithms for cosegmentation of images, in: CVPR, 2009, pp. 2028–2035.
[31] L. Mukherjee, V. Singh, J. Peng, Scale invariant cosegmentation for image groups, in: CVPR, 2011, pp. 1881–1888.
[32] A. Joulin, F. Bach, J. Ponce, Discriminative clustering for image co-segmentation, in: CVPR, 2010, pp. 1943–1950.
[33] A. Faktor, M. Irani, Co-segmentation by composition, in: ICCV, 2013, pp. 1297–1304.
[34] Z. Tao, H. Liu, H. Fu, Y. Fu, Image cosegmentation via saliency-guided constrained clustering with cosine similarity, AAAI, 2017.
[35] J.C. Rubio, J. Serrat, Video co-segmentation, in: ACCV, 2012, pp. 13–24.
[36] D.J. Chen, H.T. Chen, L.W. Chang, Video object cosegmentation, in: ACM Int. Conf. Multimed., 2012, pp. 805–808.
[37] M.P. Kumar, P.H.S. Ton, A. Zisserman, Obj cut, in: CVPR, 1, 2005, pp. 18–25.
[38] F. Meng, H. Li, K.N. Ngan, L. Zeng, Q. Wu, Feature adaptive co-segmentation by complexity awareness, IEEE TIP 22 (12) (2013) 4809–4824.
[39] R. Quan, J. Han, D. Zhang, F. Nie, Object co-segmentation via graph optimized-flexible manifold ranking, in: CVPR, 2016, pp. 687–695.
[40] W. Wang, J. Shen, X. Li, F. Porikli, Robust video object cosegmentation, IEEE TIP 24 (10) (2015) 3137–3148.
[41] W. Wang, J. Shen, H. Sun, L. Shao, Video co-saliency guided co-segmentation, IEEE TCSVT 28 (8) (2018) 1727–1736.
[42] D. Batra, A. Kowdle, D. Parikh, J. Luo, T. Chen, Interactively co-segmentating topically related images with intelligent scribble guidance, IJCV 93 (3) (2011) 273–292.
[43] M. Rubinstein, A. Joulin, J. Kopf, C. Liu, Unsupervised joint object discovery and segmentation in internet images, in: CVPR, 2013, pp. 1939–1946.
[44] H. Fu, D. Xu, S. Lin, J. Liu, Object-based RGBD image co-segmentation with mutex constraint, in: CVPR, 2015, pp. 4428–4436.
[45] J. Zhang, K. Li, W. Tao, Multivideo object cosegmentation for irrelevant frames involved videos, Signal Process. Lett. 23 (6) (2016) 785–789.
[46] L. Wang, G. Hua, R. Sukthankar, J. Xue, Z. Niu, N. Zheng, Video object discovery and co-segmentation with extremely weak supervision, IEEE TPAMI 39 (10) (2016) 2074–2088.
[47] F. Meng, H. Li, S. Zhu, B. Luo, C. Huang, B. Zeng, M. Gabbouj, Constrained directed graph clustering and segmentation propagation for multiple foregrounds cosegmentation, IEEE TCSVT 25 (11) (2015) 1735–1748.
[48] W. Yang, Z. Sun, B. Li, J. Hu, K. Yang, Unsupervised multiple object cosegmentation via ensemble miml learning, in: International Conference on MultiMedia Modeling, 2017, pp. 393–404.
[49] D. Zhang, O. Javed, M. Shah, Video object co-segmentation by regulated maximum weight cliques, in: ECCV, 2014, pp. 551–566.
[50] H. Fu, D. Xu, B. Zhang, S. Lin, R.K. Ward, Object-based multiple foreground video co-segmentation via multi-state selection graph, IEEE TIP 24 (11) (2015) 3415–3424.
[51] J. Dai, Y.N. Wu, J. Zhou, S.C. Zhu, Cosegmentation and cosketch by unsupervised learning, in: ICCV, 2013, pp. 1305–1312.
[52] K.R. Jerripothula, J. Cai, J. Lu, J. Yuan, Object co-skeletonization with co-segmentation, in: CVPR, 2017, pp. 3881–3889.
[53] F. Meng, H. Li, Q. Wu, B. Luo, K.N. Ngan, Weakly supervised part proposal segmentation from multiple images, IEEE TIP 26 (8) (2017) 4019–4031.
[54] F. Meng, H. Li, Q. Wu, K.N. Ngan, J. Cai, Seeds-based part segmentation by seeds propagation and region convexity decomposition, IEEE TMM 20 (2) (2018) 310–322.
[55] C. Lee, W.D. Jang, J.Y. Sim, C.S. Kim, Multiple random walkers and their application to image cosegmentation, in: CVPR, 2015, pp. 3837–3845.
[56] T. Taniai, S.N. Sinha, Y. Sato, Joint recovery of dense correspondence and cosegmentation in two images, in: CVPR, 2016, pp. 4246–4255.
[57] L. Mukherjee, V. Singh, J. Xu, M.D. Collins, Analyzing the subspace structure of related images: concurrent segmentation of image sets, in: ECCV, 2012, pp. 128–142.
[58] K. Li, J. Zhang, W. Tao, Unsupervised co-segmentation for indefinite number of common foreground objects, IEEE TIP 25 (4) (2016) 1898–1909.
[59] F. Wang, Q. Huang, M. Ovsjanikov, L.J. Guibas, Unsupervised multi-class joint image segmentation, in: CVPR, 2014, pp. 3142–3149.
[60] K.R. Jerripothula, J. Cai, F. Meng, J. Yuan, Automatic image co-segmentation using geometric mean saliency, in: ICIP, 2015, pp. 3277–3281.
[61] L. Li, Z. Liu, J. Zhang, Unsupervised image co-segmentation via guidance of simple images, Neurocomputing 275 (2018) 1650–1661.
[62] J. Guo, Z. Li, L.F. Cheong, S.Z. Zhou, Video co-segmentation for meaningful action extraction, in: ICCV, 2013, pp. 2232–2239.
[63] W.S. Chu, C.P. Chen, C.S. Chen, MOMI-cosegmentation: Simultaneous segmentation of multiple objects among multiple images, in: ACCV, 2010, pp. 355–368.
[64] M.D. Collins, J. Xu, L. Grady, V. Singh, Random walks based multi-image segmentation: quasiconvexity results and GPU-based solutions, in: CVPR, 2012, pp. 1656–1663.
[65] L. Liu, W. Tao, H. Liu, Complementary saliency driven co-segmentation with region searching and hierarchical constraint, Inf. Sci. 372 (2016) 72–83.
[66] F. Meng, J. Cai, H. Li, Cosegmentation of multiple image groups, in: CVIU, 146, 2016, pp. 67–76.

[67] L. Huang, R. Gan, G. Zeng, Object cosegmentation by similarity propagation with saliency information and objectness frequency map, in: International Conference on Systems and Informatics, 2017, pp. 906–911.
[68] Z. Lou, T. Gevers, Extracting primary objects by video co-segmentation, IEEE TMM 16 (8) (2014) 2110–2117.
[69] H. Fu, D. Xu, B. Zhang, S. Lin, Object-based multiple foreground video co-segmentation, in: CVPR, 2014, pp. 3166–3173.
[70] F. Rodrigues, P. Leal, Y. Kenmochi, J. Cousty, L. Najman, S. Guimaraes, Z. Patrocınio, Graph-based hierarchical video cosegmentation, in: International Conference on Image Analysis and Processing, 2017, pp. 15–26.
[71] W.C. Chiu, M. Fritz, Multi-class video co-segmentation with a generative multi-video model, in: CVPR, 2013, pp. 321–328.
[72] N. Otsu, A threshold selection method from gray-level histograms, IEEE TCYB 9 (1) (2007) 62–66.
[73] M.Y. Yang, M. Reso, J. Tang, W. Liao, B. Rosenhahn, Temporally object-based video co-segmentation, in: International Symposium on Visual Computing, 2015, pp. 198–209.
[74] S. Vicente, V. Kolmogorov, C. Rother, Cosegmentation revisited: models and optimization, in: ECCV, 2010, pp. 465–479.
[75] D.S. Hochbaum, V. Singh, An efficient algorithm for co-segmentation, in: ICCV, 2010, pp. 269–276.
[76] W. Wang, J. Shen, Higher-order image co-segmentation, IEEE TMM 18 (6) (2016) 1011–1021.
[77] X. Dong, J. Shen, L. Shao, M.H. Yang, Interactive cosegmentation using global and local energy optimization., IEEE TIP 24 (11) (2015) 3966–3977.
[78] J.C. Rubio, Unsupervised co-segmentation through region matching, in: CVPR, 2012, pp. 749–756.
[79] Q. Ning, Z. Liu, J. Zhu, M. Song, J. Bu, C. Chen, Noise-aware co-segmentation with local and global priors, Neurocomputing 287 (2018).
[80] F. Meng, H. Li, Q. Wu, K.N. Ngan, J. Cai, Seeds-based part segmentation by seeds propagation and region convexity decomposition, IEEE TMM 20 (2) (2017) 310–322.
[81] F. Meng, H. Li, G. Liu, K.N. Ngan, Object co-segmentation based on shortest path algorithm and saliency model, IEEE TMM 14 (5) (2012) 1429–1441.
[82] C. Wang, H. Zhang, L. Yang, X. Cao, H. Xiong, Multiple semantic matching on augmented n-partite graph for object co-segmentation, IEEE TIP 26 (12) (2017) 5825–5839.
[83] C. Liu, J. Yuen, A. Torralba, Sift flow: dense correspondence across scenes and its applications, IEEE TPAMI 33 (5) (2010) 978–994.
[84] J. Han, R. Quan, D. Zhang, F. Nie, Robust object co-segmentation using background prior., IEEE TIP 27 (4) (2018) 1639–1651.
[85] G. Kim, E.P. Xing, On multiple foreground cosegmentation, in: CVPR, 2012, pp. 837–844.
[86] Z. Liu, J. Zhu, J. Bu, C. Chen, Object cosegmentation by nonrigid mapping, Neurocomputing 135 (2014) 107–116.
[87] F. Meng, H. Li, Q. Wu, B. Luo, J. Cai, C. Huang, Part propagation for local part segmentation, Visual Communications and Image Processing, 2016.
[88] G. Kim, E.P. Xing, F.F. Li, T. Kanade, Distributed cosegmentation via submodular optimization on anisotropic diffusion, in: ICCV, 2011, pp. 169–176.
[89] L. Liu, K. Li, X. Liao, Image co-segmentation by co-diffusion, in: Circuits Systems and Signal Processing, 36, 2017, pp. 4423–4440.
[90] H. Li, F. Meng, Q. Wu, B. Luo, Unsupervised multiclass region cosegmentation via ensemble clustering and energy minimization, IEEE TCSVT 24 (5) (2014) 789–801.
[91] J. Sun, J. Ponce, Learning dictionary of discriminative part detectors for image categorization and cosegmentation, IJCV 120 (2) (2016) 111–133.
[92] A. Sharma, One shot joint colocalization and cosegmentation, 2017 arXiv:1705.06000.
[93] A. Joulin, F. Bach, J. Ponce, Multi-class cosegmentation, in: CVPR, 2012, pp. 542–549.
[94] Y. Wang, B.J. Yoon, X. Qian, Co-segmentation of multiple images through random walk on graphs, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 1811–1815.
[95] X. Huang, H. Li, E. Kim, A hierarchical image clustering cosegmentation framework, in: CVPR, 2012, pp. 686–693.
[96] L. Lattari, A. Montenegro, C. Vasconcelos, Unsupervised cosegmentation based on global clustering and saliency, in: ICIP, 2015, pp. 2890–2894.
[97] Z. Tao, H. Liu, H. Fu, Y. Fua, Multi-view saliency-guided clustering for image cosegmentation, IEEE TIP (2019).
[98] X. Chen, A. Shrivastava, A. Gupta, Enriching visual knowledge bases via object discovery and segmentation, in: CVPR, 2014, pp. 2035–2042.
[99] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE TPAMI 22 (8) (2000) 888–905.
[100] K.R. Jerripothula, J. Cai, J. Yuan, Group saliency propagation for large scale and quick image co-segmentation, in: ICIP, 2015, pp. 4639–4643.
[101] Y. Li, J. Liu, Z. Li, Y. Liu, H. Lu, Object co-segmentation via discriminative low rank matrix recovery, in: ACM Int. Conf. Multimed., 2013, pp. 749–752.
[102] Y. Li, J. Liu, Z. Li, H. Lu, S. Ma, Object co-segmentation via salient and common regions discovery, Neurocomputing 172 (2016) 225–234.
[103] K.R. Jerripothula, J. Cai, J. Yuan, Image co-segmentation via saliency co-fusion, IEEE TMM 18 (9) (2016) 1896–1909.
[104] Y. Ren, L. Jiao, S. Yang, S. Wang, Mutual learning between saliency and similarity: image cosegmentation via tree structured sparsity and tree graph matching, IEEE TIP 27 (9) (2018) 4690–4704.
[105] K.R. Jerripothula, J. Cai, J. Yuan, Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization, IEEE TMM 20 (9) (2018) 2466–2477.
[106] A. Borji, D.N. Sihite, L. Itti, Salient object detection: a benchmark, IEEE TIP 24 (12) (2015) 5706–5722.
[107] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.
[108] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. Lecun, Overfeat: Integrated recognition, localization and detection using convolutional networks, in: International Conference on Learning Representations, 2014.
[109] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: CVPR, 2015, pp. 3431–3440.
[110] Z. Yuan, T. Lu, Y. Wu, Z. Yuan, T. Lu, Y. Wu, Z. Yuan, T. Lu, Y. Wu, Deep–dense conditional random fields for object co-segmentation, in: IJCAI, 2017, pp. 3371–3377.
[111] W. Li, O.H. Jafari, C. Rother, Deep object co-segmentation, 2018 arXiv:1804.06423v1.
[112] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014 arXiv:1409.1556.
[113] S. Choi, S. Kim, K. Park, K. Sohn, Multispectral human co-segmentation via joint convolutional neural networks, in: ICIP, 2017, pp. 3115–3119.
[114] P. Mukherjee, B. Lall, S. Lattupally, Object cosegmentation using deep siamese network, 2018 arXiv:1803.02555v2.
[115] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C.Berg, F.-F. Li, Imagenet large scale visual recognition challenge, IJCV 115 (3) (2015) 211–252.
[116] H. Chen, Y. Huang, H. Nakayama, Semantic aware attention based deep object co-segmentation, 2018 arXiv:1810.06859v1.
[117] S.D. Jain, B. Xiong, K. Grauman, Pixel objectness, 2017 arXiv:1701.05349.
[118] F. Wang, Q. Huang, L.J. Guibas, Image co-segmentation via consistent functional maps, in: ICCV, 2013, pp. 849–856.
[119] F. Meng, H. Li, G. Liu, K.N. Ngan, Image cosegmentation by incorporating color reward strategy and active contour model, IEEE TCYB 43 (2) (2013) 725–737.
[120] C. Xu, J.J. Corso, Evaluation of super-voxel methods for early video processing, in: CVPR, IEEE, 2012, pp. 1202–1209.
[121] A. Joulin, K. Tang, F.F. Li, Efficient image and video co-localization with Frank–Wolfe algorithm, in: ECCV, 8694, 2014, pp. 253–268.
[122] J.-M. Geusebroek, A.W. Smeulders, J. Van DeWeijer, Fast anisotropic gauss filtering, IEEE TIP 12 (8) (2003) 938–943.
[123] L. Guo, T. Cheng, Y. Huang, J. Zhao, R. Zhang, Unsupervised video object segmentation by spatiotemporal graphical model, in: Multimedia Tools and Applications, 76, 2015, pp. 1–17.
[124] C. Xu, C. Xiong, J.J. Corso, Streaming hierarchical video segmentation, in: ECCV, 2012, pp. 626–639.
[125] M. Reso, J. Jachalsky, B. Rosenhahn, J. Ostermann, Temporally consistent superpixels, in: ICCV, 2013, pp. 385–392.
[126] J. Guo, L.F. Cheong, R.T. Tan, S.Z. Zhou, Consistent foreground co-segmentation, in: ACCV, 2014, pp. 241–257.
[127] W.S. Chu, F. Zhou, F.D.L. Torre, Unsupervised temporal commonality discovery, in: ECCV, 2012, pp. 373–387.
[128] K. Papoutsakis, C. Panagiotakis, A.A. Argyros, Temporal action co-segmentation in 3D motion capture data and videos, in: CVPR, 2017, pp. 6827–6836.
[129] S.S. Lin, C.H. Lin, S.H. Chang, T.Y. Lee, Object-coherence warping for stereoscopic image retargeting, IEEE TCSVT 24 (5) (2014) 759–768.
[130] A.C. Gallagher, T. Chen, Clothing cosegmentation for recognizing people, in: CVPR, 2008, pp. 1–8.
[131] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, S. Yan, Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval, IEEE TMM 18 (6) (2016) 1175–1186.
[132] B. Zhao, X. Wu, Q. Peng, S. Yan, Clothing cosegmentation for shopping images with cluttered background, IEEE TMM 18 (6) (2016) 1111–1123.
[133] N. Champion, D. Boldo, M. Pierrot-Deseilligny, G. Stamon, 2D Building change detection from high resolution satelliteimagery: a two-step hierarchical method based on 3D invariant primitives, Pattern Recognit. Lett. 31 (10) (2010) 1138–1147.
[134] P. Xiao, M. Yuan, X. Zhang, X. Feng, Y. Guo, Cosegmentation for object-based building change detection from high-resolution remotely sensed images, IEEE TGRS 55 (3) (2017) 1587–1603.
[135] J. Chen, H. Liu, J. Hou, M. Yang, M. Deng, Improving building change detection in vhr remote sensing imagery by combining coarse location and co-segmentation, ISPRS Int J Geoinf 7 (6) (2018) 213–233.
[136] A. Kowdle, D. Batra, W.C. Chen, T. Chen, iModel: interactive co-segmentation for object of interest 3d modeling, in: European Conference on Trends and Topics in Computer Vision, 2010, pp. 211–224.
[137] A. Mustafa, A. Hilton, Semantically coherent co-segmentation and reconstruction of dynamic scenes, CVPR, 2017.
[138] A.G. Money, H. Agius, Video summarisation: a conceptual framework and survey of the state of the art, J Vis Commun Image Represent 19 (2) (2008) 121–143.
[139] H.J. Zhang, J. Wu, D. Zhong, S.W. Smoliar, An integrated system for content-based video retrieval and browsing, in: Pattern Recognit., 30, 1997, pp. 643–658.
[140] P. Geetha, V. Narayanan, A survey of content-based video retrieval, J. Comput. Sci. 4 (6) (2008) 734.

[141] A. Podlesnaya, S. Podlesnyy, Deep learning based semantic video indexing and retrieval, in: Sai Intelligent Systems Conference, 2016, pp. 359–372.
[142] J.M. Chaquet, E.J. Carmona, A. Fernndez-Caballero, A survey of video datasets for human action and activity recognition, in: CVIU, 117, 2013, pp. 633–659.
[143] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: ECCV, 2006, pp. 1–15.
[144] V. Ferrari, T. Tuytelaars, L. VanGool, Object detection by contour segment networks, in: ECCV, Springer, 2006, pp. 14–28.
[145] V. Ferrari, F. Jurie, C. Schmid, From images to shape models for object detection, IJCV 87 (3) (2010) 284–303.
[146] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: A database and web-based tool for image annotation, in: IJCV, 77, 2008, pp. 157–173.
[147] F.F. Li, R. Fergus, P. Perona, One-shot learning of object categories, IEEE TPAMI 28 (4) (2006) 594–611.
[148] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, Technical Report, California Institute of Technology, 2011.
[149] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, F.F. Li, Imagenet: a large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255.
[150] M. Everingham, L. VanGool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL VOC 2010 results, 2010, (http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html).
[151] E. Borenstein, S. Ullman, Class-specific, top-down segmentation, in: ECCV, 2002, pp. 109–124.
[152] D. Tsai, M. Flagg, J.M. Rehg, Motion coherent tracking with multi-label MRF optimization., IJCV 100 (2) (2012) 190–202.
[153] K.K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, in: Machine Vision and Applications, 24, 2013, pp. 971–981.
[154] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[155] J.S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, I.S. Kweon, Pixel-level matching for video object segmentation using convolutional neural networks, in: ICCV, 2017, pp. 2186–2195.
[156] F. Perazzi, J. Pont-Tuset, B. Mcwilliams, L.V. Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: CVPR, 2016, pp. 724–732.
[157] J. Zhang, W. Li, P.O. Ogunbona, P. Wang, C. Tang, RGB-D-Based action recognition datasets: a survey, Pattern Recognit. 60 (2016) 86–105.
[158] Z. Niu, G. Hua, L. Wang, X. Gao, Knowledge-based topic model for unsupervised object discovery and localization, IEEE TIP 27 (1) (2017) 1746–1758.
[159] D. Zhang, J. Han, J. Lu, S. Ye, X. Chang, Revealing event saliency in unconstrained video collection, IEEE TIP 26 (4) (2017) 1746–1758.

Yan Ren received her B.Eng. degree and Ph.D. degree from Xidian University, Xi'an, China. She is now a research fellow in Nanyang Technological University, Singapore. Her research interests include image and video object detection, localization and recognition.



Adams Wai-Kin Kong received his PhD from the University of Waterloo, Canada. Currently, he is an associate professor at the Nanyang Technological University, Singapore. His research interests include biometrics, forensics, image processing, and pattern recognition.

Licheng Jiao received his Ph.D. degrees from Xian Jiaotong University, Xian, China. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, Xian. His research interests include image processing, natural computation, machine learning, and intelligent information processing.