Clothing Cosegmentation for Shopping Images With Cluttered Background

Bo Zhao, Xiao Wu, *Member, IEEE*, Qiang Peng, and Shuicheng Yan, *Senior Member, IEEE*

***Abstract*—In this paper, we address an important and practical problem of clothing cosegmentation (CCS): given multiple fashion model photos with natural backgrounds on e-commerce websites, to automatically and simultaneously segment all images and extract the clothing regions. However, cluttered backgrounds, variations in colors and styles, and inconsistent human poses all make it a challenging task. In this paper, a novel CCS algorithm is proposed to improve the accuracy of clothing extraction by exploiting the properties of multiple clothing images with the same apparel. First, the co-salient objects are computed by detecting the upper bodies of fashion models and transferring their locations within multiple images. Based on the coarse clothing regions determined by the upper body localization and co-salient object detection, the foreground (clothing) and background Gaussian mixture models are estimated, respectively. Finally, the clothing region in each image is extracted through energy minimization based on graph cuts iteratively. The proposed cosegmentation algorithm is mainly designed for multiple clothing images. As a byproduct, it can also be applied to single image segmentation without any modification. The experiments demonstrate that the proposed approach outperforms the state-of-the-art cosegmentation methods as well as traditional single image segmentation solution for shopping images.**

***Index Terms*—Clothing extraction, cluttered background, cosegmentation, segmentation, shopping images.**

I. Introduction

NOWADAYS, online shopping is becoming an increasingly attractive and convenient shopping way for millions of web users, especially for female customers. There are billions of diverse products available on e-commerce websites such as *eBay*, *Amazon* and *Alibaba*. In addition, the emergence of social image sharing websites, such as *Pinterest*, *Instagram* and *Flickr*, further accelerates the progress of social and personalized e-commerce. This paper focuses on the clothing images, which occupy a large portion of the online product images with great diversity.

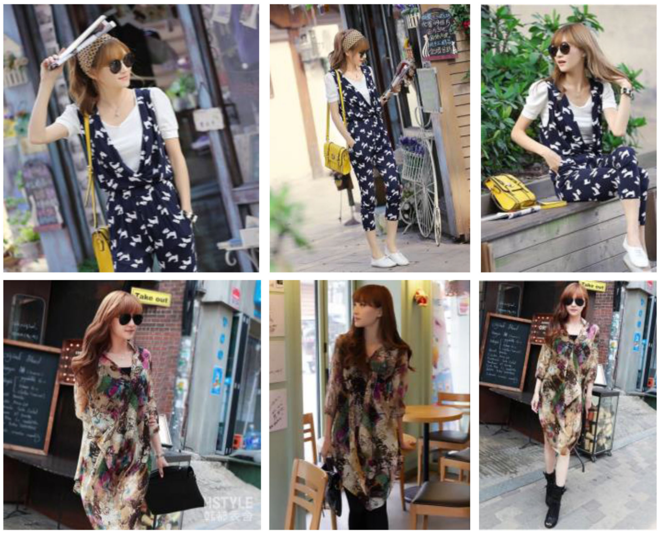


Fig. 1. Clothing images from online clothing shopping websites usually have complex backgrounds, variations in colors and styles, different lighting conditions, inconsistent human poses, viewpoint changes, and part occlusion, which make clothing extraction a challenging research task. Meanwhile, in order to better demonstrate the dressing effects of the clothes to attract consumers, the sellers usually take multiple photos of the fashion model wearing the identical clothes from different viewpoints.

To demonstrate the real dressing effects, the sellers usually upload the photos of fashion models wearing the clothes for sale. These clothing images usually contain cluttered backgrounds which can be outdoor scenes or indoor decorations. In addition, fashion models often have various poses, standing, sitting or even lying down. The pictures are captured from different angles, mainly from front and side, and in rare cases from the back. Moreover, some portions of these clothes are usually occluded by human parts (e.g., arms and hands) and personal belongings (e.g., handbags). Fig. 1 illustrates some examples from *Taobao*, raising a challenging research task, which has not been fully studied yet.

The aforementioned factors significantly affect the image search and retrieval. Current image search engines treat the images as a whole for both query images uploaded by users and product images on e-commerce websites. They have no idea of the target objects and the backgrounds. Due to complex backgrounds, the search engines usually return irrelevant results and the desired clothes are totally absent. Therefore, how to efficiently and effectively extract clothing objects and remove backgrounds becomes valuable and critical.

Image segmentation is an active research topic in computer vision and multimedia areas, which acts as an effective approach

Manuscript received April 06, 2015; revised August 14, 2015 and January 08, 2016; accepted February 25, 2016. Date of publication March 02, 2016; date of current version May 13, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61373121, Grant 61036008, and Grant 61328205, in part by the Program for Sichuan Provincial Science Fund for Distinguished Young Scholars under Grant 13QNJJ0149, in part by the Fundamental Research Funds for the Central Universities, and in part by the Sichuan Science and Technology Innovation Seedling Fund (2014-62). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marco Bertini. *(Coresponding author: Xiao Wu.)*

B. Zhao, X. Wu, and Q. Peng are with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China (e-mail: zhaobo@my.swjtu.edu.cn; wuxiaohk@home.swjtu.edu.cn; qpeng@home.swjtu.edu.cn).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: eleyans@nus.edu.sg).

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

Digital Object Identifier 10.1109/TMM.2016.2537783

to solve this problem. Its purpose is to identify and extract the object after removing the background and unrelated information. However, existing segmentation methods aim at segmenting an image into multiple regions, which mainly target natural objects, not specific for clothing. Due to the aforementioned variations, clothing extraction remains a challenging task. There are few approaches specifically proposed for clothing segmentation.

Fortunately, we have an interesting observation about the online clothing images. In order to better demonstrate the dressing effects of the clothes to attract consumers, the sellers usually take multiple photos of the fashion model wearing the identical clothes from different viewpoints, which is shown in Fig. 1. Each row of Fig. 1 demonstrates an image set of the same clothes. However, the backgrounds are different. These multiple images provide valuable clues for the clothing extraction. How to fully leverage the extra information provided by multiple images to extract the clothing regions is the main work of our clothing cosegmentation (CCS) approach.

Cosegmentation refers to the process that jointly segments the common object in two or more images, which was originally proposed by Rother *et al.* [1] in the context of simultaneously segmenting a person or object of interest from an image pair. The intuition is that the probability of regions or objects to be the interesting area in an image can be significantly boosted if they also exist in other images. Cosegmentation has been actively researched in recent years and been applied to a wide range of applications such as segmenting highly co-occurring objects in a photo collection [2], recognizing person through identical clothing in multiple images [3], etc.

Although the cosegmentation has been extensively studied in recent years, to the best of our knowledge, CCS has not been well explored. In our case, we have multiple images which contain identical clothes. These images may vary in model pose, lighting condition and background. Different from general image cosegmentation, clothing images have their characteristics. They usually contain fashion models wearing beautiful clothes to better present their actual appearance and promote the customers' purchase desire. Besides, the models usually stand in the center of the images and occupy the main area. Otherwise, the clothing on the model will be easily ignored by customers. The styles of apparel are much diverse and complicated with mixture of various patterns, as shown in Fig. 1. Despite the promising appeal of cosegmentation, very few algorithms are applicable to CCS, which motivates us to explore CCS. It requires the cosegmentation approach to not only be adaptable to heterogeneous images with high variability in content and complexity, but also integrate the characteristics of clothing.

In this paper, we propose a CCS approach to simultaneously extract clothing objects from multiple clothing images. This CCS is based on Graph Cuts [4] with the assistance of upper body detection and saliency detection. The upper body detection is first utilized to locate the potential region of clothing, which can tolerate the variations in human poses and gestures. Besides, the visually salient region in each image is also detected by transferring the multiple upper body detection results. To tolerate the detection errors, upper body detection and saliency detection are combined to coarsely locate the clothing region in each image. After modeling the Gaussian mixture models (GMM) of the clothing region and background, the common clothing regions in multiple images are extracted automatically. Experiments demonstrate that the proposed approach outperforms the state-of-the-art cosegmentation methods as well as the single image segmentation methods.

In this work, we assume that a set of images containing the same clothes is given in advance, because our algorithm targets e-commerce websites, such as Taobao, eBay, Amazon, Mecy's, etc. On these websites, we can directly obtain a group of images containing the same clothing from the product webpage. However, the websites and SNSs such as Fashionista, The Locals and Chictopia, mainly provide fashion news or portray fashion on the streets of major cities all over the world. For this type of websites, multiple images of the same clothing are not necessarily provided, but our cosegmentation algorithm can also segment these images since it can also be well adapted to single image segmentation.

The main contributions of this paper are as follows.

1) A novel CCS framework is proposed to simultaneously extract the clothing regions in multiple shopping images that frequently contain a live fashion model with a visible upper body.
2) A co-saliency detection method leveraging the upper body detection for multiple clothing images is proposed. The upper body detection and saliency detection are integrated into the CCS algorithm to determine the clothing region coarsely.
3) A new Gibbs energy function is defined to extend the traditional graph cut-based segmentation to multiple images and can be optimized efficiently.
4) The proposed cosegmentation algorithm can also be used for single image segmentation with competitive results.

The rest of this paper is organized as follows. A brief overview of related work is given in Section II. The framework of the proposed approach is introduced in Section III. The clothing and background localization and CCS are elaborated in Sections IV and V, respectively. Experiments and results are presented in Section VI. Finally, the paper is concluded with a summary.

II. Related Work

A. Clothing Study

A great deal of work related to clothing has been conducted in recent years. Pose estimation and supervised region labeling are used to parse clothing in fashion photographs in [5]. Later, a retrieval-based approach [6] is incorporated to improve the results of clothes parsing. By modeling the appearance of human clothing and surrounding context, the occupation of the person can be predicted in [7]. In mobile product image search [8], a location and an outline shape for the query object will be predicted for each image, which improves the segmentation accuracy. Both works of Liu *et al.* [9] and Kalantidis *et al.* [10] allow a user to upload a daily human photo captured in the general environment and find similar clothes in online shops. *Magic closet* [11] focuses more on the recommendation of most suitable clothes according to the specified occasion such as wedding or dating. It can also automatically pair reference clothing with the most suitable one from online shops. A fully automatic system [12] is proposed to describe the semantic attributes for clothing on the human upper body. Pose estimation is used to help feature extraction and style rules are modeled by observing co-occurrences of the attributes. In addition, clothing is used for person identification in [13]. The clothing region is first segmented based on a clothing model. The facial features and clothing features are then used to recognize the person in different images. In our previous exploration, an interactive product image search [14] provides a different manner to search for similar products. Although many research works on clothing have been conducted recently, most of the aforementioned works mainly consider the images with clean backgrounds, which is not fully adaptive to the real world. There are a great number of clothing images in online shops which are captured in a daily environment with cluttered backgrounds, which motivates this work.

B. Image Segmentation

Image segmentation is a fundamental but challenging problem in many computer vision applications. In the past few decades, numerous image segmentation approaches based on different techniques have been proposed. Most image segmentation approaches solve the problem by assigning a label to each pixel in a globalization framework. The problem of image segmentation can be reduced to contour detection [15], which combines multiple local cues such as brightness, color and texture gradients into a globalization machinery based on spectral clustering. The appearance-based local methods are integrated into global approaches in [16] and [17], and better segmentation results are achieved. Energy-based segmentation approaches [18]–[21] are very popular in recent years, in which an object function with a minimal value corresponds to the optimal segmentation. The famous *Normalized Cuts* [18] computes eigenvectors of an image and uses a clustering algorithm to partition an image. However, the eigenvector computation of Normalized Cut is expensive, which is not suitable for practical use. *Efficient graph cut* [19] overcomes the efficiency problem and has become one of the most popular approaches. As a semi-supervised approach, *Grabcuts* [20] needs a manually specified bounding box around the object to estimate a GMM for the foreground and background, respectively. After the iterative energy minimization and border matting, the extracted object has very impressive precision. Based on normalized cuts, the approach in [21] utilizes a multiscale graph bias to solve the multiscale normalized cut on a single image. Hierarchal segmentation is used in [22] to generate some object proposals and top-ranked regions are likely to be good segmentations of different objects.

In terms of clothing segmentation, the clothing is extracted by exploiting the human body detection in most existing literature. The context sensitive grammar in an AND–OR graph representation is first adopted in [23] to detect the human torso and then the body is segmented. Human body in static images is segmented based on independent component analysis at two-scale superpixel [24]. The characteristics of clothing such as location and structure are taken into consideration in clothing segmentation approaches to establish the clothing model. In [25], the clothing region is extracted by coarse region localization and fine foreground/background estimation. Certain areas below the face are treated as the potential clothing regions according to the proportion of human body and head. This idea is also used by O'Hare *et al.* [26]–[28]. However, these methods are very sensitive to the face detection and cannot deal with the variety of poses properly. Therefore, in our paper, we consider the upper body detection of models to avoid this problem. Besides, the salient region is also computed to improve the accuracy of clothing localization, since the model wearing the clothes is usually at the salient region in an image.

C. Cosegmentation

Cosegmentation is a hot but not fully explored research topic in computer vision. Current cosegmentation methods can be roughly categorized into three groups according to the number of images and common object classes they deal with. The first group consists of the methods that can only deal with two or a small number of images, such as TRGC model [1], L2 norm model [29], reward model [30] and Boykov-Jolly model [31]. Different from Markov random field (MRF) based models, a discriminative clustering framework using spectral clustering technique and positive definitive kernel is proposed in [2] to separate foreground and background with the largest margin.

The approaches of the second group (e.g., [32]–[36]) extend the cosegmentation to a large number of images. A scale invariant method [32] is proposed to segment the common object with arbitrary size in different images. The method in [33] generates a mask by retrieving visually similar objects to cosegment the common object. Some other methods are also proposed for the cosegmentation task, such as bi-level co-segmentation (BiCoS) [34], shortest path-based cosegmentation [35], and CoSand [36]. A BiCoS method [34] is proposed for image classification. At the bottom level, each image is segmented individually using the GrabCut algorithm [20] and at the top level a discriminative classification is performed to jointly segment multiple images. In [35], a digraph is first constructed by combining local region similarity and co-saliency values. And then a shortest path problem is formulated to solve the cosegmentation problem. In CoSand [36], each image is cosegmented into multiple regions based on anisotropic heat diffusion, which is submodular and can be solved efficiently. It is also adaptive to web-scale images, but cannot model a heterogeneous object that consists of multiple distinctive regions (e.g., a person) as a single foreground.

Recently, the cosegmentation approach is extended to the multi-class problem, which means a finite number of foregrounds repeatedly occur in multiple images, but only an unknown subset of them is presented in each image. In [37], a spectral-clustering term and a discriminative term are used to cosegment multiple class objects. A multiple foreground cosegmentation method is proposed in [38], which does not need all objects to recur in each image. In [39], the unknown object-like proposals are first discovered to form the energy potentials

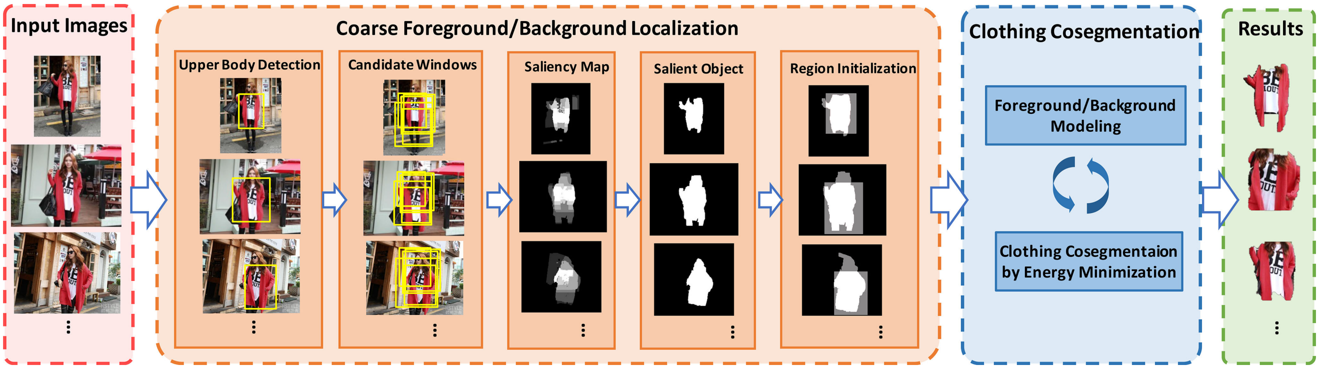


Fig. 2. Framework of the proposed CCS algorithm. It mainly consists of two phases, coarse foreground and background localization and CCS.

across all the images and α-expansion is used to minimize the energy term. The method in [40] induces affinities among image parts across images to compute the cosegments and improve the accuracy by propagating the cosegmentation likelihood maps among different images.

However, most of aforementioned cosegmentation approaches are designed for general objects, and cannot be well adapted to clothing images since they do not consider the characteristics of clothing. Besides, there has been little research exploring the cosegmentation for clothing images with cluttered backgrounds.

In the experiment section, we will compare the performance of the proposed cosegmentation algorithm CCS with DClust [2], CoSand [36] and CoComp [40]. These baseline algorithms come from the aforementioned three groups of cosegmentation methods, respectively. They are widely used for experimental comparison in many works (e.g., [37], [38], [41]) and have been proved effective for cosegmentation.

III. Framework

The framework of the proposed CCS is illustrated in Fig. 2. The whole cosegmentation process mainly consists of two phases, coarse foreground and background localization and simultaneous clothing cosegmentation. To coarsely locate the potential clothing region, the upper body detection is firstly conducted. The bounding box enclosing the human upper body but excluding the head will be treated as the candidate clothing region. Unfortunately, upper body detection is not sufficient for locating the clothing region when the background is cluttered. Motivated by the observations that the clothes are always the major object of a shopping image and are commonly placed in the central position to attract users' attention, the co-saliency map of each image is also considered. The results of upper body detection and co-salient objects are combined to improve the accuracy of clothing localization, in which each region in the image set is initially classified as three labels: clothing, background, or uncertain.

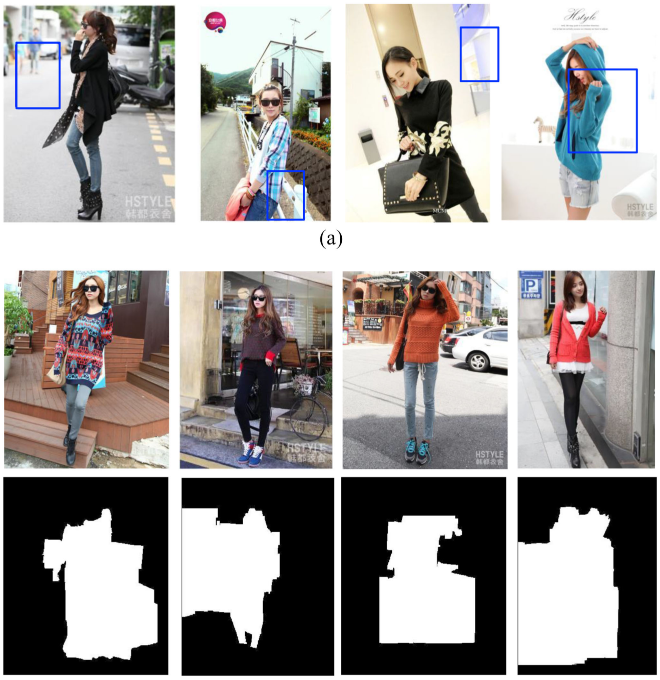
The following phase for CCS is clothing/background modeling and clothing region cosegmentation. The clothing and background modeling is first conducted, in which *GMM* is adopted to learn the appearance of clothing and background in each image. In the next step, each image in the group will use the same foreground and background GMMs to cosegment the common clothing region. The results obtained from the cosegmentation step are further used to update the GMMs of clothing and background. The step of clothing and background modeling and the phase of cosegmentation are solved iteratively in such a way until convergence. Finally, we will get the common clothing region of each image.

IV. Clothing and Background Localization

Suppose we are given M images $I = \{I_1, \ldots, I_M\}$ which contain the same clothing but have similar or different backgrounds. The coarse clothing and background region localization is first performed, in which the upper body detection and saliency detection are combined to achieve this goal.

A. Upper Body Detection

The sellers make a fashion model wear the clothes to demonstrate the real dressing effect. Therefore, if the human body is detected, the clothing region can be roughly located as well. Based on this observation, a human upper body detector is trained using the code of Felzenszwalb *et al.* [42] to locate the upper part of human body excluding the head. The upper body model is defined by a coarse root filter, several high resolution part filters and a spatial model for the location of each part relative to the root. The result of upper body detection is a bounding box which encloses all the upper body excluding the head of the model, as shown in the second row of Fig. 4. We manually annotate upper body regions of additional 1000 clothing images and test it on a validation set of 500 images. The detection will be treated as correct if the intersection area between the detected region and the annotated region is greater than 70%. The average accuracy of our detector is 80.4%. So, in most cases, the clothing region is within this bounding box. We set the region within the bounding box as the clothing region and use F_u^i to denote it for an image I_i. However, when the model stands in one side or takes some items, such as a bag, in his/her hands, it adds difficulty to the upper body detection, since the upper body model captures frontal and near frontal views. The upper body filter may have higher response at some regions in the cluttered background compared to the actual body region, which leads to a wrong bounding box. Fig. 3(a) shows some error examples of upper body detection.



(a)

(b)

Fig. 3. Error examples for upper body detection and saliency detection. (a) Examples of upper body detection error. (b) Examples of saliency detection error.

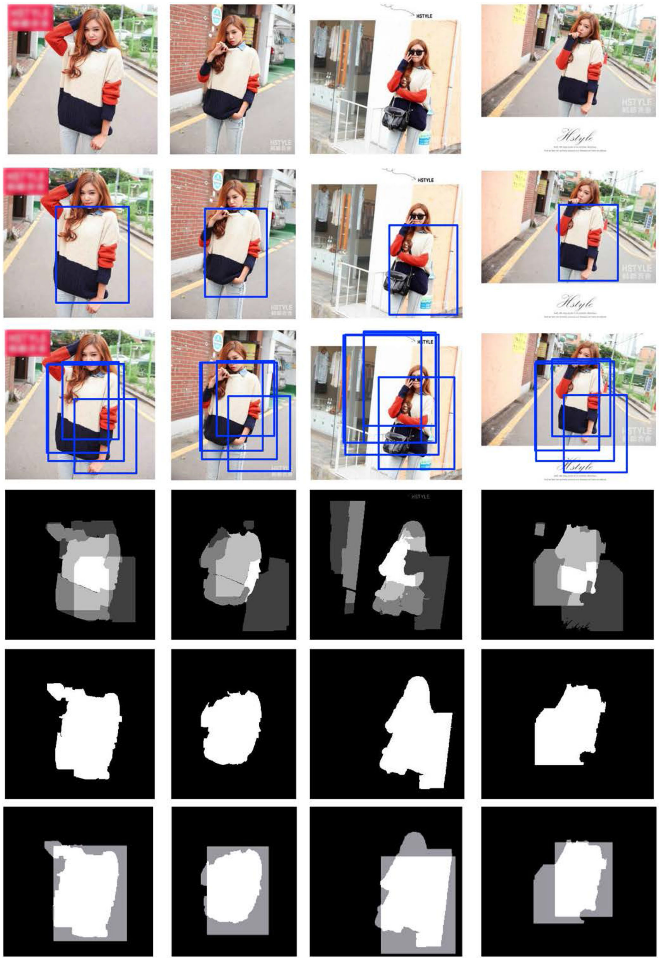


Fig. 4. Examples of region assignment. The first row shows the original images. The second row shows the results of upper body detection. The region within the blue bounding box is the clothing region. The third row is the candidate windows by mapping the bounding boxes in different images. The fourth row is the results of saliency detection. By setting a threshold, a salient object mask of each image is obtained as shown in the fifth row. The last row illustrates three different regions after upper body detection and saliency fusion.

B. Saliency Detection

To attract the attention of customers, the fashion model is usually the salient region in each image. Besides, the size of this region should not be too small. Otherwise, it will be easily ignored by customers. Therefore, it is reasonable to locate the model by detecting the salient region of the image. In this paper, a simple but effective way is proposed to detect the salient object in multiple clothing images, which consists of three steps. The first step is to generate the candidate salient object windows from multiple clothing images. In [43], many predefined salient object windows are used to compute the intra-image saliency map. However, these windows do not consider the positions of the objects in the image. Different from their work, we propose another method to automatically generate the candidate salient object windows. For a group of clothing images, we find that the salient object (i.e., the clothing region) in one image is more likely to be at the same or nearby positions as in other images. Therefore, we can use the bounding boxes in different images to estimate the possible clothing region in each image, together with its own bounding box. The bounding boxes in different images of that group will be mapped to one image. If the images have different sizes, we first scale the source image and then map its bounding box to the target image. The results of this procedure can be seen in the third row of Fig. 4. Therefore, for an M-image group, we can have M bounding boxes for each image. They can be seen as the candidate salient object windows. One window is from the upper body detector and the remaining is from the other $M-1$ images in that group.

In the second step, Grabcut [20] is performed for each image using the M candidate salient object windows. We will obtain M segments, and the clothing regions or their parts may appear many times due to the overlapped candidate windows. A pixel is salient if it is segmented as the object pixel by Grabcut for multiple times. In the last step, we combine all the segmented results to generate the saliency map. $S(p)$ denotes the clothing image saliency value at pixel p as [43]. It is defined as

$$S(p) = \frac{1}{Z}\Sigma_{s=1}^{M}\delta(L_s(p)) \tag{1}$$

where Z is a normalized constant to ensure $S(p)$ in the range of [0,1]. M is the number of bounding boxes in each image. $L_s(p)$ indicates whether the pixel p is the foreground or the background by Grabcut. $\delta(\cdot)$ will be 1 if it is foreground, otherwise it will be 0. A pixel will have the maximum salient value if it is always segmented as the foreground by different bounding boxes.

The saliency maps are shown in the fourth row of Fig. 4. Most clothing regions are highlighted with large salient values. The fifth row is the salient objects by setting a threshold on its salient map. The threshold is computed adaptively as in [44], by two times the mean saliency of a given image. The light area is the foreground while the dark area is treated as the background. From the images shown in Fig. 3(b), we can find that the salient area is not always the clothing region. This is because the wrong transferred bounding box will lead to mistakes. In some cases, the GrabCut will falsely incorporate the background into the salient region mistakenly.

C. Coarse Foreground and Background Localization

Although both upper body detection and saliency detection may have mistakes, they complement each other from different viewpoints. To get a robust estimation of the potential clothing region, they are combined to better locate the clothing region. Each image is initially classified into three categories: foreground, background and uncertain regions. Let F_u^i and F_s^i be the regions of upper body and salient object, respectively. The intersection of these two regions $F_u^i \cap F_s^i$ is treated as the initial clothing region T_F^i of the image I_i, while the intersection of background regions $\overline{F_u^i} \cap \overline{F_s^i}$ is regarded as the initial background T_B^i. The remaining regions of the image $I_i - F_u^i \cap F_s^i - \overline{F_u^i} \cap \overline{F_s^i}$ is the unknown region T_U^i, which needs to be determined later. The last row of Fig. 4 illustrates three different regions after upper body detection and saliency fusion. The white area is the clothing region, and the grey area is the uncertain region. The remaining black area is the background.

V. Clothing Cosegmentation

Once the coarse clothing region is located, the foreground and background GMM models will be computed, respectively. An iterative scheme is used to cosegment multiple clothing images.

A. Foreground and Background Modeling

After coarse foreground and background localization, the initial clothing regions $T_F = \{T_F^1, \ldots, T_F^M\}$ and background regions $T_B = \{T_B^1, \ldots, T_B^M\}$ are obtained. Let $\alpha = \{\alpha^1, \ldots, \alpha^M\}$, where $\alpha^i = (\alpha_1^i, \ldots, \alpha_n^i, \ldots, \alpha_N^i)$ indicates the pixel x_n^i in the image I_i belonging to the foreground or the background. Specifically, N refers to the total number of pixels in the image I_i and $\alpha_n^i \in \{0, 1\}$, with 0 for background and 1 for foreground. The remaining pixels are unknown which need to be decided later. So all pixels $x_n^i \in T_F$ are set to 1 and pixels $x_n^i \in T_B$ are set to 0.

Since the initial foreground T_F and background T_B contain several major colors, two GMMs are used to model the color distributions of foreground T_F and background T_B, respectively. In this work, the RGB color space is deployed. The clothing and background GMMs are estimated using the pixels $x_n^i \in T_F$ and $x_n^i \in T_B$, respectively. The distribution of the GMM model is defined as

$$p(x|c) = \sum_{k=1}^{K} \frac{\pi_k^c}{(2\pi)^{\frac{d}{2}} |\Sigma_k^c|^{\frac{d}{2}}} \cdot \exp\left[-\frac{1}{2}(x - \mu_k^c)^{\mathrm{T}} (\Sigma_k^c)^{-1} (x - \mu_k^c)\right] \tag{2}$$

where c indicates the background GMM ($c = 0$) or clothing GMM ($c = 1$). x is a three-dimensional vector standing for the RGB values of the pixel x. d represents the dimension of x, which is equal to 3. μ_k^c and Σ_k^c are the mean value and covariance matrix of the kth Gaussian of the clothing/background GMM, respectively. π_k^c is the weighting factors of the kth Gaussian of clothing/background, respectively. All parameters are determined by the EM algorithm. K is the number of Gaussian distributions, which is set as 4 in our experiments.

So, the parameters of the GMM models are

$$\underline{\theta} = \{\pi(c, k), \mu(c, k), \Sigma(c, k), c = 0, 1 \text{ and } k = 1, \ldots, K\} \tag{3}$$

where π is the weight. μ is the mean and Σ is the covariance of the $2K$ Gaussian components for the clothing and background distributions.

B. Clothing Region Cosegmentation

After modeling the initial foreground and background GMMs, these two models are used to cosegment the foreground (clothing regions) T_F from each image I_i. The foreground GMM and background GMM are two full-covariance Gaussian mixtures with K components, respectively. Each image I_i has an additional vector $k^i = \{k_1^i, \ldots, k_n^i, \ldots, k_N^i\}$, with $k_n^i \in \{1, \ldots, K\}$, which indicates each pixel belonging to the clothing or background GMM model, according to $\alpha_n^i = 0$ or 1.

The energy minimization approach is adopted to solve the cosegmentation task based on *Graph Cut* [4], which is a very popular and widely used approach in computer vision problems, such as image segmentation and image matting, to solve the energy minimization problem. A foreground/background segmentation corresponds to an array $\underline{\alpha}^i = (\alpha_1^i, \ldots, \alpha_n^i)$ for each pixel. $a_n^i \in \{0, 1\}$, with 0 for background and 1 for foreground. Each segmentation result also has an energy E, in which a good segmentation should have a small value.

To extend the traditional Graph Cut-based segmentation to multiple images, we redefine the *Gibbs* energy for M images as

$$E = \sum_{i=1}^{M} e(\underline{\alpha}^i, k^i, \underline{\theta}, x^i) \tag{4}$$

where the *Gibbs* energy for image I_i is

$$e(\underline{\alpha}^i, k^i, \underline{\theta}, x^i) = U(\underline{\alpha}^i, k^i, \underline{\theta}, x^i) + V(\underline{\alpha}^i, x^i). \tag{5}$$

U is the data term which evaluates the fitness of the distribution $\underline{\alpha}^i$ to the data x^i in image I_i. $\underline{\theta}$ represents the GMM model.

The data term U is defined by taking account of the color GMM models as

$$U(\underline{\alpha}^i, k^i, \underline{\theta}, x^i) = \sum_n D(\alpha_n^i, k_n^i, \underline{\theta}, x_n^i) \tag{6}$$

where

$$D(\alpha_n^i, k_n^i, \underline{\theta}, x_n^i) = -\log p(x_n^i | \alpha_n^i, k_n^i, \underline{\theta}) - \log \pi(\alpha_n^i, k_n^i) \tag{7}$$

$p(\cdot)$ is the Gaussian probability distribution, and $\pi(\cdot)$ is the mixture weighting coefficient.

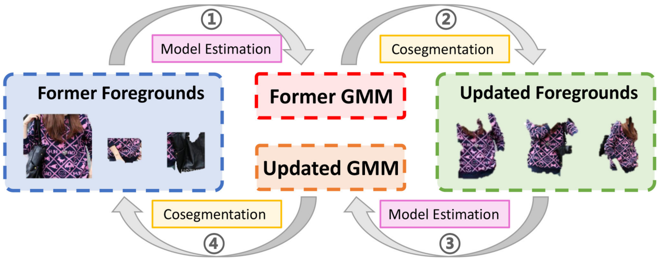


Fig. 5. Illustration of the clothing model. At the beginning, the clothing regions are coarsely assigned by upper body detection and saliency detection. By using this initial clothing model, our cosegmentation algorithm can get a refined result. Then, these refined results are used to update the clothing model, which will be used for next round cosegmentation.

The smoothness term V is computed as

$$V(\underline{\alpha}^i, x^i) = \gamma \sum_{(m,n)\in\mathcal{N}} \delta(\alpha_n^i - \alpha_m^i) e^{-\beta\|x_m^i - x_n^i\|^2} \tag{8}$$

where $\delta(\cdot)$ is the indicator function which means that this smoothness term only exists when $\alpha_n^i \neq \alpha_m^i$. $\|\cdot\|$ is the *Euclidean* distance of neighboring pixels x_m^i and x_n^i in neighborhood $\mathcal{N}$. γ is a constant to decide whether to encourage smoothness or not. β is another parameter to adjust the similarity value as in [4].

Minimum cut is adopted to solve the energy minimization problem. We use the foreground/background GMM models to segment each image I_i and get its minimum energy $e(\underline{\alpha}^i, k^i, \underline{\theta}, x^i)$. The whole energy E of multiple images is the sum of these individual energies as defined in (4). A minimum energy and its corresponding segment result will be obtained after cosegmentation. All the segmented foreground and background will be used to update the GMM models, since the initial foreground GMM and background GMM are not accurate due to the coarse clothing region localization. Inspired by the work of GrabCut [20], we use an iterative scheme to solve the cosegmentation problem. During the foreground and background modeling, the cosegmentation results T'_F and T'_B are used to update the previous foreground and background GMMs, respectively, as shown in the left side of Fig. 5. During the clothing region cosegmentation, the updated models will be used to cosegment the images as shown in the right side of Fig. 5. Then, the energy minimization problem is optimized again based on the new foreground/background GMM. Such an iteration will be repeated until the energy E converges. So, the parameter α is automatically refined in each iteration and the newly formed foreground T'_F and background T'_B refine the color GMM parameters θ. The algorithm is guaranteed to converge at least to a local minimum according to Rother *et al.* [20].

The cosegmentation algorithm is summarized in Algorithm 1. We start with the initial foreground T_F, background T_B and undetermined region T_U. Then the foreground and background GMM models are learned from these regions, respectively. After that, the clothing regions are iteratively cosegmented and the foreground and background are updated until convergence.

Algorithm 1: Clothing Cosegmentation.

```
Input:
    (1) $T_B$: the initial background regions in multiple
     images; $T_F$: the initial clothing regions in multiple
     images; $T_U$: the remaining unknown regions;
    (2) Initialize $\alpha_n = 0$ for $n \in T_B$ and $\alpha_n = 1$ for
     $n \in T_F$;
    (3) Background and clothing GMMs are initialized
     from sets $\alpha_n = 0$ and $\alpha_n = 1$, respectively;
Output:
     Clothing regions $T'_F$;
while not converged do
    Assign GMM components to pixels: for each $n$ in $T_U$
    $k_n^i := \arg\min_{k_n^i} D(\alpha_n^i, k_n^i, \theta, x_n^i)$
    Learn GMM parameters from data $x$:
    $\underline{\theta} := \arg\min_{\underline{\theta}} U(\underline{\alpha}, k, \underline{\theta}, x)$
    Estimate segmentation: use min cut to solve:
    $\min_{\{a_n : n \in T_U\}} \min_k \sum_{i=1}^{M} e(\underline{\alpha}^i, k^i, \underline{\theta}, x^i)$
end while
return $T'_F$;
```

TABLE I
Category Distribution of the Cosegmentation Dataset

Cotton Clothes	23.1%	Jacket	3.3%	Dress	14.1%
Wind Coat	5.6%	Suit	4.9%	Sweater	20.1%
Cotton Coat	16.4%	Shirt	9.9%	T-shirt	2.6%

VI. Experiments

To verify the performance of the proposed approach, we conduct the experiments to quantitatively compare our method with other state-of-the-art cosegmentation approaches as well as traditional single image segmentation algorithms.

A. Dataset and Performance Metric

Although there are a limited number of datasets involving clothing images (e.g., [6], [45], [46]), they are not suitable for the CCS task, because they only have single image. Since there is no publicly available dataset for CCS, a new dataset specific for CCS has to be built. To evaluate the cosegmentation performance, we have to manually segment every image and mark the clothing region as the ground truth, which is a laborious task. It is impossible to construct a large scale dataset for evaluation. Therefore, we collect 1 000 images downloaded from *Taobao*, one of the largest online shopping websites in the world, as the dataset. This dataset consists of 304 groups of clothing images (with resolution of 420×420 or 420×350), in which each group contains 2∼7 images with fashion models wearing the same clothes as shown in Fig. 1. These images are taken in the natural scene with cluttered backgrounds and from different viewpoints. The distribution of the clothing categories of the dataset is listed in Table I.

TABLE II
PERFORMANCE COMPARISON ON DIFFERENT METHODS OF CLOTHING AND BACKGROUND LOCALIZATION

Algorithm	Accuracy (%)
CCS with fixed bounding box	40.1 ± 21.7
CCS with saliency detection only	44.0 ± 21.7
CCS with upper body detection only	56.9 ± 20.8
CCS combining upper body and saliency detection	**62.7** ± 17.8

TABLE III
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART COSEGMENTATION APPROACHES

Algorithm	Accuracy (%)
CoSand [36]	17.7 ± 17.7
DClust [2]	26.1 ± 14.8
CoComp [40]	42.9 ± 21.1
CCS	**62.7** ± 17.8

As the widely used measurement for *PASCAL* challenges and cosegmentation evaluation (e.g., [34], [36]–[38]), *accuracy* is adopted as the performance metric, which is defined as the intersection between the cosegmentation result and the ground truth $\frac{|GT_i \cap T_i|}{|GT_i \cup T_i|}$, where GT_i and T_i are the ground truth and the segmentation result of the ith image, respectively.

Three sets of experiments are conducted to verify the effectiveness of the proposed approach. The performance of clothing region localization is first conducted. And then the performance of the proposed algorithm is compared with the state-of-the-art cosegmentation algorithms and several single image segmentation methods, respectively. The speed efficiency is finally analyzed.

B. Comparison of Clothing Region Localization Methods

Since the initial clothing region localization will affect the result of co-segmentation, we will first evaluate the performance of the initial region assignment to verify the improvement of the proposed method. It is compared with three different region assignment methods, i.e., simple fixed bounding box, saliency detection only, and upper body detection only. The following steps of CCS remain the same for these methods. Since the human being or clothing is usually placed in the middle area of a shopping image, we use a simple region initialization as the baseline, in which the central area in each image is regarded as the initial clothing area. The central area is set to 60% of the width and height of the original image size. For the saliency detection only baseline, the salient object mask is used to determine the coarse clothing region. The white region shown in the last row of Fig. 4 is the initial clothing region and the black region is the initial background region. For the baseline of upper body detection only, the bounding box of upper body is used to locate the initial clothing region. That is, the area inside the bounding box is treated as the initial foreground region, as shown in the second row of Fig. 4, while the remaining is regarded as the initial background area.

The performance comparison is listed in Table II. From this table, we can see that different strategies for clothing and background localization affect the performance significantly. The proposed algorithm CCS combining upper body detection and saliency detection outperforms other region localization methods. CCS with the fixed bounding box method has the worst result. The accuracy is only 40.1%. A fixed bounding box does not consider the varied poses and the proportion of the human body in different images as shown in Fig. 1. The accuracy of CCS with saliency detection only is 44.0%, around 10% improvement compared to the fixed bounding box method. Since the detected saliency region is not always good, especially in the scenario of cluttered background, causing the relatively poor cosegmentation performance. With the assistance of human parts and spatial relationships among them, the clothing region localization based on upper body detection can better locate the clothing, so that cosegmentation performance is improved. Upper body detection and saliency detection complement each other. Their combination can more accurately locate the potential regions of clothing and background, which further improves the performance. It reaches 62.7%.

C. Comparison of State-of-the-Art Cosegmentation Methods

In this section, we will compare the performance of the proposed cosegmentation algorithm CCS with three cosegmentation algorithms, DClust [2], CoSand [36] and CoComp [40]. Their publicly available codes are run on our CCS dataset. The parameters are set as the default values. DClust [2] simultaneously assigns foreground and background labels to all images by combining the bottom-up image segmentation and kernel method in a discriminative clustering framework. CoSand [36] models the segmentation task as the temperature maximization problem on anisotropic heat diffusion. The finite heat sources which maximize the temperature correspond to segments after the diffusion theoretic optimization. For CoComp [40], visually matching techniques across images are introduced to compose the common object. The performance comparison with these cosegmentation methods is listed in Table III.

From Table III, we can see that the cosegmentation performance of DClust and CoSand is quite poor, only 26.1% and 17.7%, respectively. The result of CoComp is relatively better than the former two cosegmentaion methods, reaching 42.9%. These three cosegmentation algorithms are general object oriented, not special for clothing images, which do not consider the characteristics of clothing images. For the complicated problem of clothing image cosegmentation, general-purpose cosegmentation approaches totally fail. These images captured in natural scenes have too many common objects compared to those in traditional applications. The fashion models, the clothes and even the background objects are very similar among multiple images, although there exist differences in some parts, which make classic approaches incompetent to cope with these complicated scenarios. On the contrary, the proposed CCS method makes good use of the properties of clothing images, which achieves good performance. With the help of clothing and background localization, the clothing region is roughly located. Then

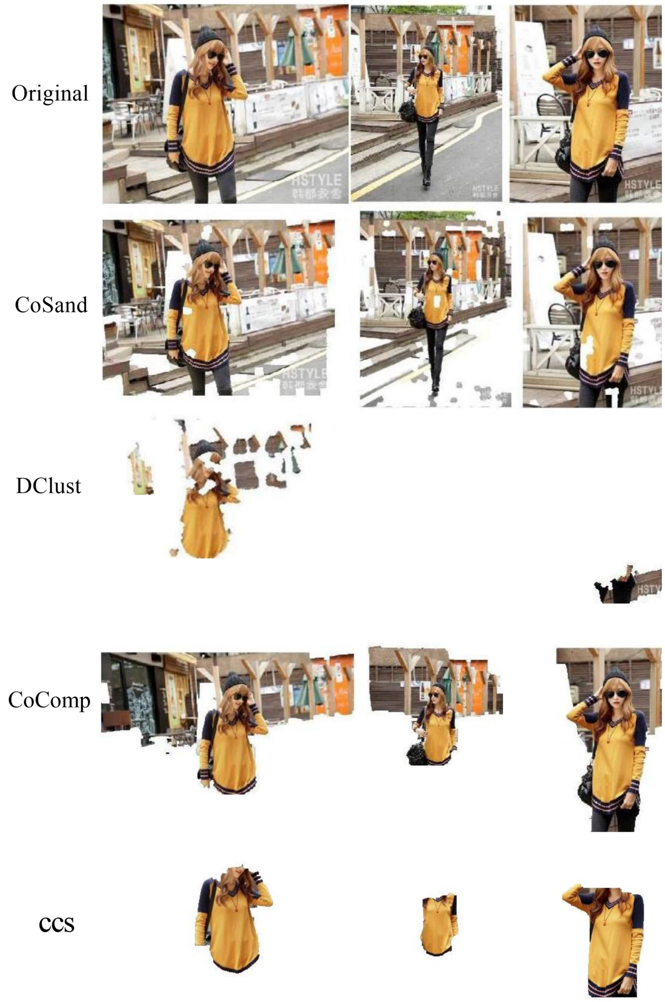


Fig. 6. Cosegmentation result comparison with state-of-the-art cosegmentation approaches. The first row shows the original images. The second to fourth rows are the results using DClust [2], CoSand [36], and CoComp [40], respectively. The last row shows the results of the proposed cosegmentation algorithm CCS.

the foreground and background GMM models from multiple images are jointly modeled and an iterative scheme is used to cosegment these images. The clothes in multiple images can be correctly captured.

Fig. 6 demonstrates the results of CCS using different cosegmentation approaches. The second to fourth rows correspond to the results of CoSand, DClust and CoComp, respectively. The last row shows the cosegmentation results of CCS. From this figure, we can see that the performance of CCS outperforms classic cosegmentation algorithms. Because the clothing images usually have cluttered backgrounds, they cannot distinguish the clothing regions from the backgrounds. Due to the complex background, the performance of CoSand is quite bad. Most of the cosegmentation results contain too much background. For some images in the second row of Fig. 6, the results after cosegmentation are almost the same as the original images, in which a large portion of background is falsely treated as foreground. The clothing cannot be correctly extracted. The performance of DClust is better than CoSand. Unfortunately, some clothing regions are falsely treated as background and are removed. Even for some cases, no clothing regions are kept, for example the second and third images shown in the third row of Fig. 6. The results generated by CoComp are relatively better than the former two algorithms. We can see that the results mistakenly incorporate some background regions. Since the CoComp will first compute the affinities between images parts, it will find the shared regions from background. Therefore, when the backgrounds of multiple images are similar, it tends to treat some background regions as the shared foregrounds. On the contrary, although the results of CCS are not perfect, the performance is much better than other cosegmentation approaches, which are very close to the real clothing regions. The clothing and background models can be more accurately estimated. By using these models, the cosegmentation accuracy is much higher.

TABLE IV
PERFORMANCE COMPARISON WITH CLASSIC SINGLE IMAGE SEGMENTATION ALGORITHMS

Algorithm	Accuracy(%)
Efficient Graph Cut [19]	37.6 ± 17.8
MNcut [21]	18.1 ± 9.5
GrabCut [20]	54.2 ± 18.3
Category Independent Object Proposals [22]	35.1 ± 26.2
Paper Doll Parsing [6]	43.5 ± 18.3
CCS on single image	**55.8** ± 19.9

D. Comparison of Single Image Segmentation Algorithms

In addition to the comparison with the state-of-the-art cosegmentation approaches, we also compare our algorithm with the classic single image segmentation approaches. Five algorithms, Efficient Graph Cuts [19], Multiscale Normalized Cut (MNcut) [21], GrabCut [20], Category Independent Object Proposals [22] and Paper Doll Paring [6] are chosen as the baselines, which are widely used for single image segmentation or clothing parsing. Efficient Graph Cuts [19] is based on graph theory and pairwise region comparison, which is very fast in practice. The default parameters are used to segment the images and the best clothing region is manually selected. MNcut [21] is based on the Normalized Cut graph partitioning framework. It compresses a large fully connected graph into a multiscale graph capturing the image structure at an increasingly large neighborhood, and segments the image into k parts. The publicly available tool is used in this experiment. We set $k = 2$, and manually select the segmented region which produces a higher accuracy. GrabCut [20] provides an interactive way for image segmentation. With the assistance of a user specified bounding box around the object, it estimates the color distributions of the target object and background, and the energy minimization is conducted alternatively until convergence. In our experiment, we initially set the center region (60% of the width or height) to be the potential clothing region instead of manually specifying. Although Category Independent Object Proposals [22] is not designed for image segmentation, it can generate some object

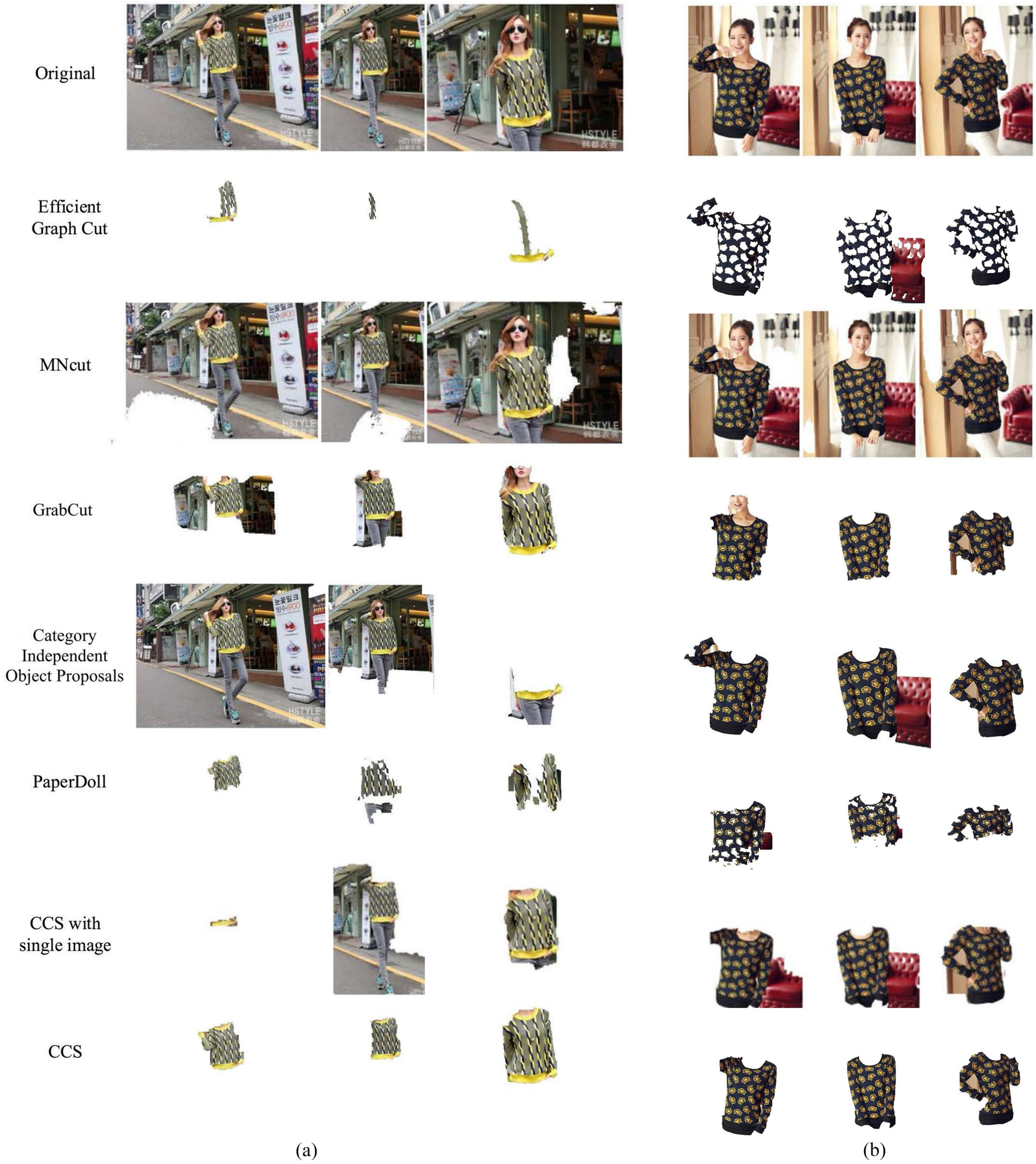


Fig. 7. Performance comparison with classic single image segmentation algorithms. The first row is the original images. The results of Efficient Graph Cut [19], MNcut [21], GrabCut [20], Category Independent Object Proposals [22], and Paper Doll Parsing [6] are show in the second to sixth rows, respectively. The results of CCS using only single images are shown in the fourth row. The last row is the cosegmentation results of CCS using multiple clothing images.

proposals and the clothing can be treated as an object. So we also compare with it. The best proposal is chosen as the clothing region. The clothing can be parsed with Paper Doll Parsing [6], by retrieving similar outfits from the parsed collection, building local models from retrieved clothing items, and transferring inferred clothing items from retrieved samples to the query image. The results usually contain multiple clothing items as well as some human parts. A best parsing area is manually selected as the clothing region.

Table IV lists the performance comparison with five segmentation or parsing approaches. The accuracy of Efficient Graph Cut is 37.6%. It can segment the clothing with similar color well, since it mainly uses the color feature to do the pairwise region comparison. Diverse colors or textures will make the result incomplete. The accuracy of MNcut is very poor, only 18.1%. The edges with high contrast existing in both clothing and background make it difficult to segment the clothing properly. Besides, MNcut tends to segment an image into homogeneous regions with similar color or texture as Efficient Graph Cut. However, the clothing is usually heterogeneous (i.e., the clothing may consist of different colors and textures), so MNcut may not segment the whole clothing properly. Moreover, the cluttered background further aggravates this situation. For GrabCut, the accuracy reaches 54.2%, which is much better than MNcut. With the assistance of clothing localization, the performance is significantly improved. As an object detection algorithm, the best object proposal generated by Endres and Hoiem [22] is not always the clothing region, since it lacks the prior information about the clothing category. Its accuracy is 35%, slightly worse than Efficient Graph Cut. The accuracy of Paper Doll Parsing is 43.5%, much better than MNcut but relatively worse than GrabCut. It can accurately locate the human body and remove the cluttered backgrounds. However, the clothing region will be incomplete when the clothes are striped or plaid. The result of our single image implementation of CCS is slightly better than GrabCut, reaching 55.8%. When only one image is available, the segmentation of CCS is quite similar to GrabCut, both using the GMM and graph cut. The difference mainly lies in the clothing localization.

Fig. 7 illustrates two sets of clothing segmentation results using different single image segmentation approaches. We can see that the performance of Efficient Graph Cut is poor when the clothing contains multiple colors or has strips such as Fig. 7(a). It will segment it into several regions, leading to an incomplete result. The results of MNcut are also very poor. Since the MNcut is set to bi-segment the image, it tends to classify a heterogeneous region apart and leave the remaining cluttered region as a different class. Usually the clothing is in the cluttered region which includes the majority of the image. The result of GrabCut is much better. Most clothing can be segmented properly. However, the fixed bounding box is not competent for all cases of fashion models. If the fashion model is not in the center or the model only takes a small portion of the whole image, such as the first image in Fig. 7(a), the segmentation result becomes poor. The extracted clothing is incomplete or a large portion of background is falsely treated as the clothing region. Category Independent Object Proposals of [22] is bad when the image has cluttered background, such as Fig. 7(a), which contains the majority of the background or only a small part of the clothing region. In general, the results generated by Paper Doll Parsing are not too bad. Most of the clothing region is correctly extracted. It can parse the clothing well when the clothing has pure color or simple texture. Unfortunately, when the clothing is striped or plaid, the clothing region will be incomplete or disconnected. CCS on single images works generally better than GrabCut, although some background may be falsely treated as clothing, e.g., the red sofa in the second set. The segmentation performance of CCS on multiple images can be further improved because of the extra information provided by other images. Even if the clothing location is not properly initialized, the global clothing/background GMM models can partially correct the mistakes.

E. Efficiency Analysis

In addition to the performance on effectiveness, we also evaluate the speed efficiency. The experiments are conducted on a 3.1 GHz processor with 4 GB memory. The code is implemented with MATLAB. It takes 3 min on average, and 5$\sim$15 min to segment a pair of images using CoComp and DClust, respectively, which are very slow. For the image group containing more than three clothing images, DClust takes even more than 30 min, which is intolerable. The optimization method of DClust has an overall complexity of $O(s^2)$, where s is the number of superpixels. CoSand is the most efficient one, which usually takes 7$\sim$10 s to cosegment a group of images.

The following three factors mainly affect the computation cost of the proposed method: upper body detection, saliency detection, and cosegmentation. The iterative CCS process is very efficient. Segmenting a pair of images only takes 0.3 s on average. For the image group with the largest number in our experiment (7 images), the time cost is 1.2 s. The common clothing regions will be extracted efficiently by the proposed approach. Typically, the time cost is mainly spent on the step of coarse clothing region localization. It takes on average 3.2 s to detect the upper body in each image. To accelerate the speed of saliency detection, we run GrabCut using different candidate windows in parallel for each image and it takes about 3.1 s to compute the saliency map. Overall, it takes less than 8 s to cosegment each group of images. Moreover, since the proposed algorithm is based on the graph cuts and MRFs, the Gibbs energy can be optimized by minimum cut efficiently. Given the bounding box of a clothing area, it will take less than 0.5 s to finish the segmentation, including GMM modeling and energy optimization. The GrabCut will take about 0.7 s to segment one image. It needs more time to do the border matting operation. From this experiment, we can see that the proposed approach is more efficient compared to the state-of-the-art methods.

VII. Conclusion

In this paper, we propose a novel clothing image cosegmentation algorithm which mainly consists of two phases, coarse clothing region localization and clothing cosegmentation. Upper body detection and saliency detection are combined to coarsely locate the clothing region. Based on these initial clothing/background areas, the clothing and background GMMs are learnt and updated to jointly segment and extract the clothing regions in those images. The experimental results demonstrate that the effectiveness of the proposed method both in single image segmentation and multiple image cosegmentation.

Our cosegmentation algorithm is mainly designed for the clothing images with fashion models, because these images are usually taken in the natural scenes with cluttered backgrounds. The clothing images without human usually have clean backgrounds and are easy to extract the clothing region.

In our future work, we plan to integrate the characteristics of clothing, such as symmetry, style consistency and location constraint into our cosegmentation framework to further improve the cosegmentation performance. For clothing images with cluttered background, the search results of current commercial search engines are very bad. In the future, the clothing extraction will be integrated into the clothing visual search systems to improve the retrieval accuracy, which can deal with street photos with complex backgrounds.

REFERENCES

[1] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, pp. 993–1000.

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

[3] A. C. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[4] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary $\&amp;$ region segmentation of objects in n-d images," in *Proc. Int. Conf. Comput. Vis.*, Jul. 2001, vol. 1, pp. 105–112.

[5] K. Yamaguchi, M. Kiapour, L. Ortiz, and T. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3570–3577.

[6] K. Yamaguchi, M. Kiapour, and T. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3519–3526.

[7] Z. Song, M. Wang, X. Hua, and S. Yan, "Predicting occupation via human clothing and contexts," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1084–1091.

[8] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Mobile product image search by automatic query object extraction," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 114–127.

[9] S. Liu *et al.*, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3330–3337.

[10] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proc. ACM Conf. Multimedia Retrieval*, Apr. 2013, pp. 105–112.

[11] S. Liu *et al.*, "Hi, magic closet, tell me what to wear!" in *Proc. ACM Int. Conf. Multimedia*, Oct. 2012, pp. 619–628.

[12] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 609–623.

[13] A. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[14] X. Wu, X.-P. Deng, L.-L. Liang, and Q. Peng, "Interactive product image search with complex scenes," in *Proc. Int. Conf. Internet Multimedia Comput. Serv.*, May 2012, pp. 136–139.

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

[16] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[17] T. H. Kim, K.-M. Lee, and S.-U. Lee, "Nonparametric higher-order learning for interactive segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3201–3208.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[19] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[20] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[21] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 2, pp. 1124–1131.

[22] I. Endres and D. Hoiem, "Category independent object proposals," in *Proc. Euro. Conf. Comput. Vis.*, Sep. 2010, pp. 575–588.

[23] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu, "Composite templates for cloth modeling and sketching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 1, pp. 943–950.

[24] S. Li, H.-C. Lu, X. Ruan, and Y.-W. Chen, "Human body segmentation based on deformable models and two-scale superpixel," *Pattern Anal. Appl.*, vol. 15, no. 4, pp. 399–413, Nov. 2012.

[25] X. Wu, B. Zhao, L.-L. Liang, and Q. Peng, "Clothing extraction by coarse region localization and fine foreground/background estimation," in *Proc. Int. Conf. Multimedia Model.*, 2013, pp. 316–326.

[26] N. O'Hare and A. F. Smeaton, "Context-aware person identification in personal photo collections," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 220–228, Feb. 2009.

[27] W. Zhang, T. Zhang, and D. Tretter, "Clothing-based person clustering in family photos," in *Proc. Int. Conf. Image Process.*, Sep. 2010, pp. 4593–4596.

[28] Z. Hu, H. Yan, and X. Lin, "Clothing segmentation using foreground and background estimation based on the constrained delaunay triangulation," *Pattern Recog.*, vol. 41, no. 5, pp. 1581–1592, May 2008.

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

[30] D. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. Int. Conf. Comput. Vis.*, Sep. 2009, pp. 269–276.

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

[32] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1881–1888.

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

[34] Y. Chai, V. Lempitsky, and A. Zisserman, "Bicos: A bi-level co-segmentation method for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2579–2586.

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

[36] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 169–176.

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

[38] G. Kim and E. P. Xing, "On multiple foreground cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 837–844.

[39] H. Li, F. Meng, Q. Wu, and B. Luo, "Unsupervised multiclass region cosegmentation via ensemble clustering and energy minimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 789–801, May 2014.

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

[41] E. Kim, H. Li, and X. Huang, "A hierarchical image clustering cosegmentation framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 686–693.

[42] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[43] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1896–1909, Dec. 2013.

[44] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1597–1604.

[45] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1543–1550.

[46] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *Proc. Int. Conf. Image Process.*, Sep. 2011, pp. 2937–2940.

Bo Zhao received the B.Sc. degree in networking engineering from Southwest Jiaotong University, Chengdu, China, in 2010, where he is currently working toward the Ph.D. degree in information science and technology.

He is currently a Visiting Scholar with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include multimedia, computer vision, and machine learning.

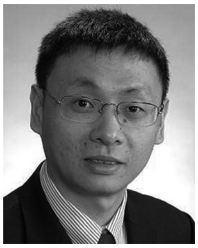
Xiao Wu (S'05–M'08) received the B.Eng. and M.S. degrees in computer science from Yunnan University, Yunnan, China, in 1999 and 2002, respectively, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, China, in 2008.

He is currently an Associate Professor with Southwest Jiaotong University, Chengdu, China. He is the Assistant Dean of the School of Information Science and Technology and the Head of the Department of Computer Science and Technology with Southwest Jiaotong University. He is currently with the School of Information and Computer Science, University of California, Irvine, CA, USA, as a Visiting Associate Professor. He was a Research Assistant and a Senior Research Associate with the City University of Hong Kong from 2003 to 2004, and 2007 to 2009, respectively. From 2006 to 2007, he was with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, as a Visiting Scholar. He was with the Institute of Software, Chinese Academy of Sciences, Beijing, China, from 2001 to 2002. His research interests include multimedia information retrieval, image/video computing, and data mining.

Prof. Wu was the recipient of the Second Prize of Natural Science Award of the Ministry of Education, China, in 2015.

Qiang Peng received the B.E. degree in automation control from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.Eng. degree in computer application and technology and the Ph.D. degree in traffic information and control engineering from Southwest Jiaotong University, Chengdu, China, in 1987, and 2004, respectively.

He is currently a Professor with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China. He has been in charge of more than ten national scientific projects, authored or coauthored more than 80 papers, and holds 10 Chinese patents. His research interests include digital video compression and transmission, image/graphics processing, traffic information detection and simulation, virtual reality technology, and multimedia system and application.



Shuicheng Yan (M'06–SM'09) is currently an Associate Professor with the National University of Singapore, Singapore, and the founding Lead of the Learning and Vision Research Group. He has authored or coauthored nearly 400 technical papers over a wide range of research topics, with over 12 000 Google Scholar citations. His research interests include machine learning, computer vision, and multimedia.

Prof. Yan is an ISI highly cited Researcher 2014, and an IAPR Fellow 2014. He has been serving as an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Computer Vision and Image Understanding*, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He was the recipient of the Best Paper Award from ACM Multimedia Conference (ACM MM) 2013 (Best Paper and Best Student Paper), ACM MM 2012 (Best Demo), The Pacific-Rim Conference on Multimedia (PCM) 2011, ACM MM 2010, IEEE International Conference on Multimedia and Expo (ICME) 2010, and International Conference on Internet Multimedia Computing and Service (ICIMCS) 2009, the Runner-Up Prize of ILSVRC 2013, the Winner Prize of the classification task in PASCAL VOC 2010–2012, the Winner Prize of the Segmentation Task in PASCAL VOC 2012, the Honorable Mention Prize of the Detection Task in PASCAL VOC 2010, the 2010 TCSVT Best Associate Editor Award, the 2010 Young Faculty Research Award, the 2011 Singapore Young Scientist Award, and the 2012 NUS Young Researcher Award.