Weakly Supervised Part Proposal Segmentation From Multiple Images

Fanman Meng, *Member, IEEE*, Hongliang Li, *Senior Member, IEEE*, Qingbo Wu, *Member, IEEE*, Bing Luo, and King Ngi Ngan, *Fellow, IEEE*

***Abstract*— Weakly supervised local part segmentation is challenging, due to the difficulty of modeling multiple local parts from image level prior. In this paper, we propose a new weakly supervised local part proposal segmentation method based on the observation that local parts will keep fixed along the object pose variations. Hence, the local part can be segmented by capturing object pose variations. Based on such observation, a new local part proposal segmentation model is proposed. Three aspects, such as shape similarity-based cosegmentation, shape matching-based part detection and segmentation, and graph matching-based part assignment are considered. A part segmentation energy function is first proposed. Four terms, such as MRF-based single image segmentation term, shape feature-based foreground consistency term, NCuts-based part segmentation term, and two-order graphs matching based part consistency term, are contained. Then, a three sub-minimization-based energy minimization method is proposed to accomplish approximation solution. Finally, we verify our method based on three image data sets (PASCAL VOC 2008 Part data set, UCB Bird data set, and Cat-Dog data set), and one video data set (UCF Sports) data set. The experimental results demonstrate a better segmentation performance compared with the existing object cosegmentation and part proposal generation methods.**

***Index Terms*— Part segmentation, cosegmentation, shape matching, graph matching, NCuts.**

I. Introduction

THE existing image segmentation methods paid much attention on object segmentation that segments object region from images [1]–[3], while the detailed local part regions are ignored. Note that many computer vision tasks rely on local information analysis where the part segmentation is an essential step, such as the fined bird image classification that uses the appearances of local parts to distinguish the bird subspecies [4].

Recently, a few of researchers have paid attention on local part segmentation [4]–[7], where the multiple local parts and their structures instead of object are segmented. Compared with object segmentation, it needs to handle multiple part priors and their spatial structures, and thus is a more difficult task.

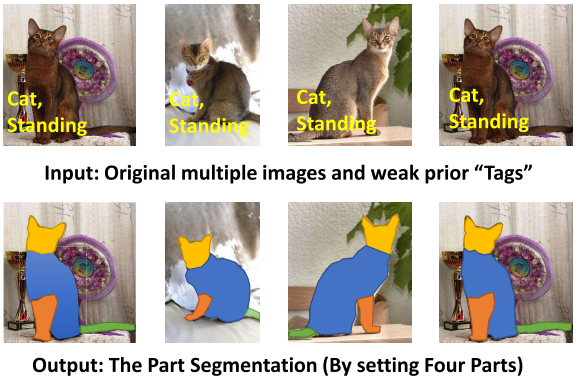


Fig. 1. An explanation of part level segmentation using weak priors, which is the input and output of the proposed method.

The existing local part segmentation methods mainly focus on supervised manner that learns each part prior from accurate training data [5]–[7]. Successful part segmentation can be achieved by the careful prior learning and segmentation model design. However, pixel-level training data is generally not available in many applications, while the rough priors such as image tags often appear. An example is shown in Fig. 1, where the image level tags can be easily provided by user. In such case, the problem changes to segment multiple part from images with tag priors, which is weakly supervised part segmentation problem.

The challenge of weakly supervised part segmentation is how to define semantic local part from weak priors. In other words, the initial priors are so rough that it is difficult to generate part priors. A feasible solution is local part proposal generation, i.e., using a set of sufficient local part proposals to provide local parts [4]. However, there still lacks a useful cue to capture part regions. Fortunately, it is observed that local parts will keep fixed among object variations. Hence, local parts can be defined as the regions that keep fixed among the variations. An example is shown in Fig. 1, where the "head" region keeps fixed among "Cat" images, and is set as a local part, while the region containing "head" and "body" varies among the images, and is treated as local region instead. By the definition, we can obtain part proposals by shape matching.

Based on such observation, we propose a weakly supervised part proposal segmentation model. Given multiple images related to an object, with the assumption that the object is contained in each image, we aim at segmenting local part

Manuscript received March 8, 2016; revised March 29, 2017; accepted May 11, 2017. Date of publication May 26, 2017; date of current version June 23, 2017. This work was supported in part by National Natural Science Foundation of China under Grant 61502084, Grant 61525102, and Grant 61601102. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Clausi. *(Corresponding author: Fanman Meng.)*

F. Meng, H. Li, Q. Wu, and B. Luo are with School of Electronic Engineering, University of Electronic Science and Technology of China, Cheng Du 611731, China (e-mail: fmmeng@uestc.edu.cn; hlli@uestc.edu.cn).

K. N. Ngan is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, and also with the School of Electronic Engineering, University of Electronic Science and Technology of China, Cheng Du 610051, China (e-mail: knngan@ee.cuhk.edu.hk).

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

Digital Object Identifier 10.1109/TIP.2017.2708839

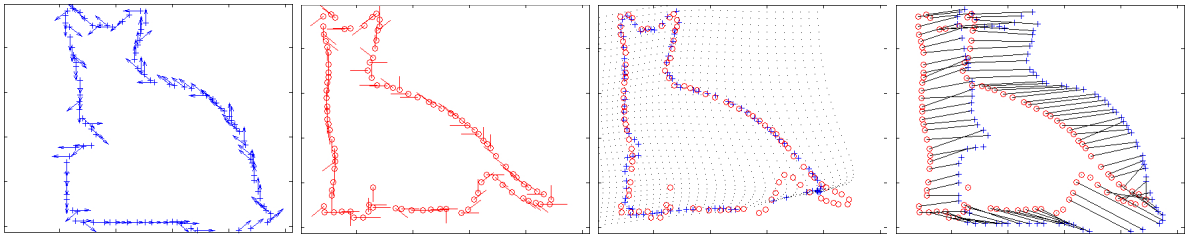


Fig. 2. The example of original region pairs and their correspondence matching. (a)(b): the original region pairs. (c): the matching results. (d): the corresponding pixels on the boundary by [8]. It is seen that local parts keep fixed along the shape variation, such as the "head" that have same pixel shifts.

proposals from the images by measuring pose variations. Our part proposal segmentation problem is formulated as a L label assignment task. An energy function is designed to measure the label assignment by four terms: the single image segmentation term, the object region consistency term, the part consistency term, and the global part structure consistency term. The first two terms are to enforce the foreground to be common objects. The third term is to constraint the consistency of parts, which is formulated by capturing shape variations, as shown in Fig. 2. The fourth term is to enforce the similarity of part structure. Based on the energy, the part segmentation is accomplished by the energy minimization, which is solved by three sub-minimization problems, such as cosegmentation problem by object proposals and shortest path searching, part generation problem by shape context matching and NCuts, and part label matching problem by two order graph matching. We verify our method on both the image and video data sets. The experimental results demonstrate that our method obtains better IOU values than several state-of-the-art object segmentation and part segmentation methods.

Our contributions are listed as follows.

- It is a weak part proposal generation method by capturing pose variation among objects, which can obtain better part segmentation.
- A new segmentation model is proposed by including cosegmentation, part segmentation, and part label matching, which can segment local part proposals from multiple images.

The paper is organized as follows. We present related work in Section II. The proposed method, including energy design, energy minimization, and detailed algorithm is introduced in Section III. Section IV displays the results of our method and the comparison methods. We finally draw the conclusion in Section V.

II. Related Work

Image segmentation is a clustering process that clusters pixels into semantic regions. The existing segmentation methods can be classified into superpixel level segmentation, object level segmentation, and part level segmentation according to the semantic level of segmentation targets.

A. Superpixel Level Segmentation

Superpixel level segmentation is unsupervised manner that automatically clusters pixels into a set of local smooth regions. The similarities among adjacent pixels are employed to guide clustering. The typical methods are spectral clustering based NCuts segmentation [9], Mean-shift clustering based segmentation [10], edge clustering based UCM superpixel [11], and K-means clustering based SLIC method [12]. Because the number of superpixels are much smaller than pixels, superpixel is mainly used to take place of the pixels in practice applications in order to release the computational burden. However, since the similarities among neighboring pixels are not sufficient to provide semantic priors, the superpixels are not semantic regions.

B. Object Level Segmentation

Another important research branch is the object level segmentation, which aims at extracting semantic object regions. The object prior is needed in such segmentation, which is generally learned from two types of training data: the pixel-level training data, and the image-level training data. The first type of data is accurate, and results in the good segmentation, such as the recent CNN learning based specific object segmentation and objectness evaluation based general object proposal generation [13]–[16]. Meanwhile, it is hard to provide this type of data, since the training data is obtained by either manually drawing or using interactive image segmentation [17], [18], which are very time-consuming for large scale of images.

Compared with pixel-level training data, the image-level training data is easier to obtain, which is called weakly supervised segmentation [19], [20]. The main idea is to learn the prior from the similar regions among multiple relevant images. There are usually two steps: similar region matching, and object prior learning, which are iteratively performed until the convergence. In general, the two steps are formulated in a CRF segmentation framework that is usually minimized by EM algorithm with α-expansion algorithm. Note that object level segmentation focus on the whole object region segmentation, which ignores the semantic local part segmentation.

By seeing the discrimination ability of local region and their spatial structure in object region representation, Zhang *et al.* in [21] propose an excellent weakly-supervised semantic segmentation framework by exploiting spatial structure cue from image-level labels. In the framework, the local regions are first represented by graphlets structure [22]. Then, three cues such as image level label, global spatial layout and geometric context are combined in the manifold embedding to discover the discriminative spatial structure. Based on the results of manifold embedding, normalized cut based

segmentation is finally employed to obtain the semantic regions. Zhang *et al.* further extend the framework by considering feature contributions and object region relationships in [23]. Patch alignment and Bayesian network are used, and better results are obtained. The framework by Zhang *et al.* shows the usefulness of spatial structure in capturing local region relationships. Meanwhile, our method is different from the method by Zhang *et al.* The main difference is that our targets are the local part proposal regions, while the targets of the methods [21], [23] are the semantic object regions. Therefore, they are still object-level segmentation rather than part-level segmentation.

C. Part Level Segmentation

Local part and their structure have been widely used in many high-level object detection, recognition and understanding tasks. In the part level segmentation, two aspects need to be considered: learning the multiple part prior models, and measuring the part relationships, which makes the part level segmentation more difficult than object level segmentation. The existing part segmentation methods focus on supervised manner, i.e., learning the part prior models and part structure from accurate training data, and then applying them to accomplish part segmentation. For example, Luo *et al.* [5] propose a Deep Decompositional Network to segment local parts of pedestrian. Three layers such as occlusion estimation layers, completion layers, and decomposition layers are carefully designed to directly obtain the label map. Wang and Yuille [6] propose a semantic part segmentation model by using compositional model to describe the relationships among parts. The latent SVM is employed to learn the parameters of the model and is used to accomplish the model inference with the dynamic programming. The results show an improvement compared with object level segmentation methods. Wang *et al*. [7] intend to obtain both the object and part segmentation with the concept of semantic compositional parts (SCP). The segmentation is performed in a novel fully connected conditional random field model with the SCP potentials learned from FCN network. Note that these methods rely on the accurate learning of the part prior models and part structures, which are fully supervised methods.

Recently, Krause *et al.* [4] try to automatically generate part annotations by only given the bounding box of the object. Compared with the above fully supervised methods, it is weakly supervised manner. The method first performs cosegmentation to obtain the object regions, and then align them to generate the parts. However, since it is difficult to determine the part regions without any part annotations, the method instead generates diverse set of part candidate by random sampling, which expects to be remedied by learning the discriminative parts in the following object recognition tasks. Here, we try to generate the part proposals by capturing the pose variations.

D. Cosegmentation

Another related work is cosegmentation, which assumes that a common object is contained in each image, and segments object by extracting similar regions among images. Although many cosegmentation methods have been proposed, such as single class cosegmentation [24], [25], multiple class cosegmentation [26], [27], multiple group cosegmentation [28], noise image based cosegmentation, and RGBD cosegmentation [29], these methods focus on object region segmentation, which are not part level segmentation. Compared with cosegmentation, our method tries to obtain local part regions, which is a more detailed and difficult task.

III. The Proposed Method

In this section, we introduce our method by first illustrating the label assignment based problem formulation. Then, our energy function consisting of four terms are introduced. Finally, the model minimization is presented based on three sub-minimization problems: cosegmentation, part generation, and part assignment.

A. The Problem Formulation

Given multiple images $I = \{I_1, \cdots, I_n\}$ with number n, a common object is contained in each image I_i. We denote the object regions as $\mathcal{S} = \{S_1, \cdots, S_n\}$. Each object region S_i consists of N local parts, i.e., $S_i = \{P_{i1}, \cdots, P_{iN}\}$. Based on the assumption that the object and their parts are similar among images, our task is to obtain the object part set $\mathcal{S}$ from the multiple images according to their similarity consistency and shape variations. For simplicity, we assume the images have the same size with the same number of pixels m.

We formulate the part-level segmentation problem as label assignment problems. Specifically, denoting $\mathcal{C} = \{C_0, C_1, \cdots, C_N\}$ as the background label and the part labels, every pixel p_{ik} in each image I_i is assigned a label $l_{ik} \in \mathcal{C}$ that represents its part classes. By denoting $L_i = \{l_{i1}, \cdots, l_{im}\}$ as the label set for I_i, and $L = \{L_1, \cdots, L_n\}$ as the label set of all images, we formulate the part segmentation problem by searching $L^* \in \Omega_L$ that best fits the semantic part segmentation, which can be represented as

$$L^* = \arg\min_{L \in \Omega_L} E(L) \tag{1}$$

where Ω_L is the domain of L, $E(L)$ is the energy function measuring the fitness between L and image regions. A small energy function indicates a good label assignment.

There are two challenges in modeling (1): the design of energy $E(L)$, and the energy minimization. One the one hand, the energy E designed from weak image level tags should efficiently evaluate the fitness between L and semantic parts. On the other hand, easy minimization can be deduced from the energy to obtain global or approximation solution. Here, our energy is designed from shape matching by four terms: single image segmentation, multiple image cosegmentation, part consistency and part structure consistency. The energy minimization is accomplished by three sub-optimization problem: cosegmentation minimization, NCuts minimization and graph matching minimization. We next detail our energy design and the corresponding energy minimization, respectively.

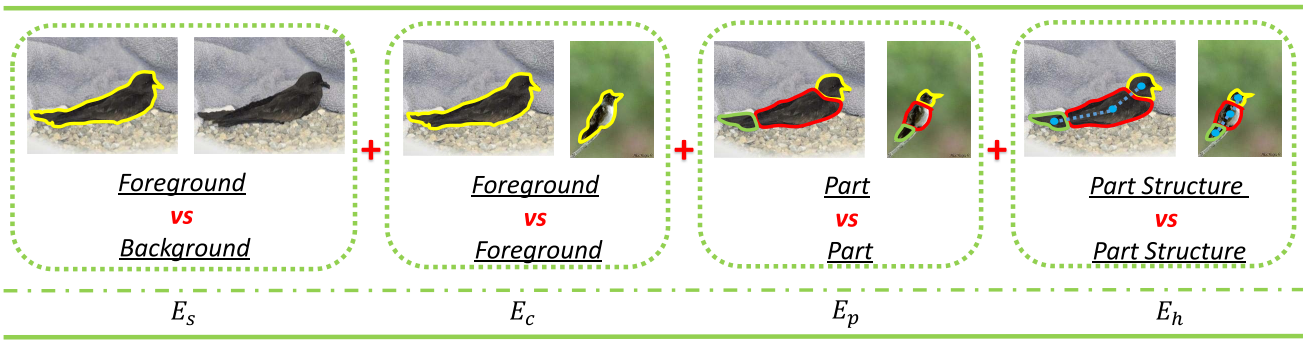


Fig. 3. An illustration of our energy design. There are four terms, i.e., single image segmentation E_s, multiple image cosegmentation E_c, part consistency E_p and part structure consistency E_h, respectively.

B. The Energy Function Design

Our energy function is designed by four terms, which can be represented as:

$$E = E_s + E_c + E_p + E_h \tag{2}$$

where E_s is the segmentation evaluation in each image, which makes the foreground to be different from the background. E_c is the cosegmentation evaluation, which measures the similarity among foregrounds, and constraints the object region to be similar. The first two terms can be concluded as cosegmentation terms. E_p is the part consistency evaluation among images, which enforces the similarity of parts. E_h is the part structure consistency, which measures the consistency of part structure. Fig. 3 displays an example to illustrate our energy design. Note that the last two terms are related to part, and we name them as part terms.

1) E_s: We design E_s by pixel consistency of the foreground and background, which is the classical energy of single image segmentation model. Here, Markov Random Field segmentation model is employed for E_s. Given an image I_k and its label set L_k, the background pixels and foreground pixels are denoted as $B_k = \{p_{ik} | l_{ki} = C_0\}$ and $F_k = \{p_{ki} | l_{ki} \neq C_0\}$, where l_{ki} is the label of pixel p_{ki}. Then, we evaluate the label L_k by the data term and pairwise term, which is represented by

$$\begin{aligned} E_k^s = \sum_{p_{ki} \in \Omega_k} \Big[& P(p_{ki}|\theta_{kF}) \delta(l_{ki} \neq C_0) \\ & + P(p_{ki}|\theta_{kB}) \delta(l_{ki} = C_0) \Big] \\ + \sum_{(p_{ki}, p_{kj}) \in \mathcal{N}} & w_{sk}(p_{ki}, p_{kj}) \delta(l_{ki} \neq l_{kj}) \end{aligned} \tag{3}$$

where E_k^s is the energy for image I_k, Ω_k is the pixel domain of image I_k, θ_{kF} and θ_{kB} are parameters of the foreground and background models, which is learned from F_k and B_k by Mixture Gaussian Model. $P(p_{ki}|\theta_{kF})$ is the probability of pixel p_{ki} under foreground model. A large value indicates a good consistency of the pixel and the region. $\delta(\cdot) = 1$ if $\cdot$ is true. Otherwise, it is zero. The first term is also known as data term. w_{ks} is the similarity matrix of pixels in each single image I_i. $\mathcal{N}$ is the 3×3 neighboring relationship. The second term (pairwise term) punishes the label changes among neighboring pixel unless there are very large color variations.

Based on (3), we define E_s as the sum of E_{sk} for all images, which can be represented as

$$E_s = \sum_{k=1}^{n} \exp(-E_k^s) \tag{4}$$

2) E_c: We use E_c to enforce the foreground similar among images, which is the global term in cosegmentation. Given a pair of images I_k and I_r, and their labels are L_k and L_r, we first obtain the foreground regions F_k and F_r from the images by the labels. Then, we define the foreground similarity of I_k and I_r by

$$E_{kr}^c = d(f(F_k), f(F_r)) \tag{5}$$

where f is the feature extraction function of region, d is the Euclidean distance between the features. Similar features have small value of E_{kr}^c. Here, we use shape context feature [8] as f in order to capture the mid-level features. It is seen that E_{kr}^c forces the foreground to be similar in shape.

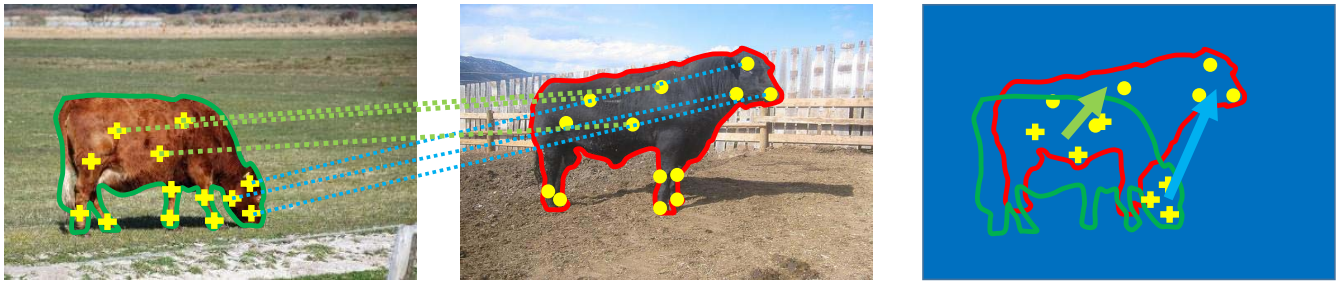
By considering the multiple images, we define E^c by summing up E_{kr}^c of all image pairs, which is represented as

$$E_c = \sum_{k=1}^{n} \sum_{r=1}^{n} E_{kr}^c = \sum_{k=1}^{n} \sum_{r=1}^{n} d(f(F_k), f(F_r)) \tag{6}$$

3) E_p: E_p is to evaluate the part consistency, which is defined by the assumption that there are pixel level matching between image pairs (I_k, I_r). A matching example is shown in Fig. 4, where the match is performed based on the shape variation matching, and each pixel in I_k (the first image) has a match pixel in I_r (the second image). Based on the match, the label energy E_{kr}^p for I_k and I_r is defined by:

$$\begin{aligned} E_{kr}^p = & - \sum_{p_i \in \Omega_k} \delta(l_i = \tilde{l}_i) \\ & + \sum_{(p_i, p_j) \in \mathcal{N}_k} Ncut(d(\Delta(p_i), \Delta(p_j))) \cdot \delta(l_i \neq l_j) \end{aligned} \tag{7}$$

where, $l_i \in L_k$ is the label of $p_i \in I_k$, $\tilde{p}_i \in I_r$ is the match pixel of p_i and $\tilde{l}_i \in L_r$ is its label, $\Delta(p_i)$ is the shift vector of p_i defined as $v(p_i) - v(\tilde{p}_i)$, where $v(p) = (x_p, y_p)$ is the



(a) (b) (c)

Fig. 4. An examples of the region matching. (a)(b): Two images containing the region of object "Cow". Each pixel in the first object region has a matched pixel in the second object region, as the lines in the images. Based on the matching, the pixels within a part region have the same matching shifts, while the pixels of different part regions have very large shift, as shown in (c).

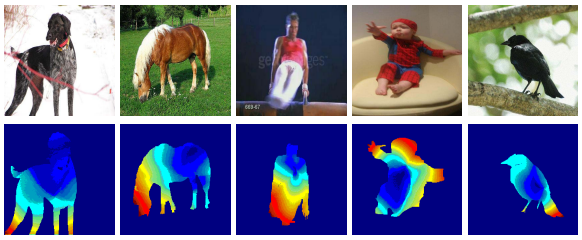


Fig. 5. Some examples of the shift maps obtained by our method. The shift value is represented by the color. It is seen that the color of the parts are different, which demonstrates the effectiveness of our method.

vector of p based on its location. It describes the shift of pixel p_i as shown in Fig. 4(c). By considering the shift vectors of all pixels in I_k, we can obtain a shift map M_s, where the value of each pixel is the shift vector, i.e., $M_s(p_i) = \Delta(p_i)$. Fig. 5 displays some shift maps, where five original images and their shift maps are shown. It is seen that shift map distinguishes the local parts in these images, which guarantees the following part proposal generation.

In the last term of (7), $d(\Delta(p_i), \Delta(p_j))$ is the distance between the shift vectors, which describes the matching shift consistency between image pairs. Large value indicates a large change of shift, and corresponds to the border of part regions. $Ncut(d(\Delta(p_i), \Delta(p_j)))$ is the cutting evaluation among different label regions $\delta(l_i \neq l_j)$, which prefers to segment local parts along the large variations of $d(\Delta(p_i), \Delta(p_j))$. We formulate it as the Normalized cut defined in [9]. It is seen that because the part always keeps fixed among images, pixels within one part have similar Δ. Meanwhile, the pixels with different part have large differences of Δ due to the pose variation. Hence, the second term aims at dividing object along with pixels with large changes of Δ.

In (7), the first term is the number of the matched pixels with the same labels. Larger value of this term indicates a good part consistency. The second term is the cost by labelling neighboring pixels. It enforces the label to be consistency among neighboring pixels. Once there is label change, it hopes to have large matching shifts, i.e., they come from different parts. It is seen that the two terms in (7) are similar to the data term and pairwise term in MRF segmentation. However, the first term is based on the part matching, which is a global cue. Furthermore, the second term is based on the matching shift, which is different from the pixel color variations. Hence, the regions can be segmented into several segments even if the pixels have same colors.

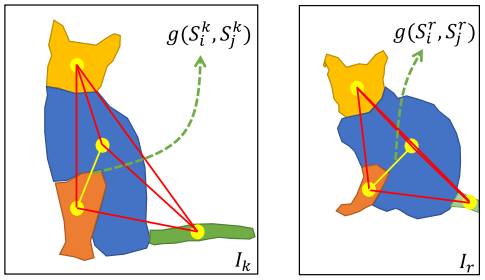


Fig. 6. An example to illustrate the fourth term E_h of part structure consistency. g is the part relationship by the spatial distance, and d is the part structure consistency evaluation, which is based on the relationships of all part pairs.

Based on (7), we set the energy function E_p by considering all image pairs, i.e.,

$$E_p = \sum_{k\in\Omega}\sum_{r\in\Omega}\Big[- \sum_{p_i\in\Omega_k} \delta(l_i = \tilde{l}_i) + \sum_{(p_i,p_j)\in\mathcal{N}} Ncut(d(\Delta^r(p_i), \Delta^r(p_j)) \cdot \delta(l_i \neq l_j)\Big] \tag{8}$$

where $\Delta^r(p_i)$ is the matching shift vector of $p_i \in I_k$ based on image I_r.

4) E_h: E_h evaluates the part structure consistency. We consider two aspects: part spatial relationship, and the relationship consistency. Given an image pair (I_k, I_r) with labels (L_k, L_r), E_{kr}^h is defined as,

$$E_{kr}^h = \sum_{i=1}^{n}\sum_{j=1}^{n} d(g(S_i^k, S_j^k), g(S_i^r, S_j^r)) \tag{9}$$

where S_i^k is the region of label L_i in image I_k, $g(S_i^k, S_j^k)$ is the spatial relationship between the ith and jth regions in image I_k, which is represented by function g. $d(g(S_i^k, S_j^k), g(S_i^r, S_j^r))$ is the distance between a pair of relationships, which measure the consistency of part pair matching. An example is shown in Fig. 6. The final structure consistency is the sum of all part pairs.

By considering all image pairs, E_h is defined as

$$E_h = \sum_{k=1}^{n}\sum_{r=1}^{n}\left[\sum_{i=1}^{m}\sum_{j=1}^{m} d(g(S_i^k, S_j^k), g(S_i^r, S_j^r))\right] \quad (10)$$

It is seen that formula (9) is the consistency of part spatial relationships ("Head"-"Body" to "Head"-"Body"), which is the second order matching. Note that the first term in (7) is the part similarity ("Head" to "Head"), which is the first order matching. Hence, by combing the two terms, it is a two-order graph matching problem. Only the labels with similar features among the same label region and the same spatial structure in terms of high-order graph matching will lead to small values.

By introducing the terms (4), (6), (8), and (10) into (2), we obtain the final energy function. We next introduce the energy minimization.

C. The Energy Function Minimization

Because our energy contains nonlinear terms such as E_c, and is also formed by many sums of multiple images, it is difficult to globally minimize the energy. Instead, we pursue approximate solution by diving original problem into three sub-minimization problems, i.e., cosegmentation problem, part generation problem and region matching problems.

1) Cosegmentation Problem: We combine the first two terms E_s and E_c to form the cosegmentation problem, which can be represented as:

$$\begin{aligned} E_C &= E_s + E_c \\ &= \sum_{i\in\Omega} E_{si} + \sum_{(i,j)\in\Omega\times\Omega} d(f(F_i), f(F_j)) \end{aligned} \quad (11)$$

This is classical cosegmentation problem, but with difficult shape similarity constraints. Here, we use the strategy in [30], [31] for the minimization. The main idea is to first consider all the segments that satisfying the first term E_s by object proposals, and then select regions that best satisfying E_s to be the final results. The minimizing of the second term can be solved by using a fully connected graph to represent the relationships of the proposals, and then performing the belief propagation on the graph to score the common regions, as used in [31]. By considering the computational cost of the fully connected graph, we only construct the graph based on neighboring images, and achieve the common object segmentation by dynamic programming [30]. In our method, the object proposals are generated by [32]. Since it obtains the bounding boxes rather than regions, we perform Grabcut on the bounding boxes to obtain region proposals. In the graph generation, shape feature in [8] is used for the edge weight calculation.

2) Part Generation Problem: We treat the second term in E_p as the part generation sub problem, which is represented as

$$\begin{aligned} E_P = \sum_{k\in\Omega}\sum_{r\in\Omega}\Bigg[\sum_{(p_i,p_j)\in\mathcal{N}_k} & Ncut(d(\Delta^r(p_i), \Delta^r(p_j)) \\ & \cdot\delta(l_i \neq l_j)\Bigg] \end{aligned} \quad (12)$$

Algorithm 1 Weakly Supervised Local Part Segmentation

Input: Multiple Images $I = \{I_1, \cdots, I_n\}$.
Initialize: The number of parts N
Output: Local Part $L = \{L_1, \cdots, L_n\}$

---

*% **Object Proposal Generation***
For $i = 1$ to n **do**
1) Generation windows of proposals for I_i by [32].
2) Generating object regions from windows by GrabCut.

End For

---

*% **Object Cosegmentation***
1) Constructing graph to representing the similarity relationships among proposals by [30] and shape context feature [8].
2) Obtaining common object regions based on the graph by dynamic programming.

---

*% **Shape Matching***
For All (i, j),$i, j \in [1, n]$, $i \neq j$ **do**
1) Performing shape matching on the foregrounds F_i, and F_j of (I_i, I_j) by [8].
2) Generating shift map based on the pixel matching by Section III-C2.

End For

---

*% **Part Generation***
For $i = 1$ to n **do**
- Generating part regions for I_i by (14) using the method in [9].

End For

---

*% **Part Label Matching***
- Obtaining the matched part labels by solving the problem in (16).

$$\begin{aligned} = \sum_{k\in\Omega}\Bigg[\sum_{(p_i,p_j)\in\mathcal{N}_k}\sum_{r\in\Omega} & Ncut(d(\Delta^r(p_i), \Delta^r(p_j)) \\ & \cdot\delta(l_i \neq l_j)\Bigg] \end{aligned} \quad (13)$$

To solve this problem, we divide the problem into many sub-problems based on each image, and forms the sub-problem as,

$$l^* = \arg\min_l \left[\sum_{(p_i,p_j)\in\mathcal{N}_k} Ncut(d(\Delta(p_i), \Delta(p_j)) \cdot \delta(l_i \neq l_j)\right] \quad (14)$$

where $\Delta(p_i) = \frac{1}{n}\sum_{r=1}^{n}\Delta^r(p_i)$ is the average shift of all images. In other words, because there are multiple images, and each image will result in a shift map, we average these shift maps to form the final one. It is seen that the sub-problem in (14) is the classical normalized cut segmentation problem with the part number n, which can be solved by spectral techniques. Based on (14), we minimize (12) by solving these sub-problems in (14) one by one.

Fig. 7. Our part proposal segmentation results on PASCAL 2008 part datasets by setting $N_s = 4$ and various number of local part N. The original image is shown in the first row. The segmentation results with $N = 3$ to 8 are displayed from the second row to the seventh row, respectively. The bottom row shows the results by $N = 1$, which are the cosegmentation results.

3) Local Part Matching: The third step is to combine the first term in (8) and (10), and forms a two order graph matching problem, which is represented as

$$l^* = \arg\min_{l} \sum_{k\in\Omega}\sum_{r\in\Omega}\Bigg[-\sum_{p_i\in\Omega_k}\delta(l_i=\tilde{l}_i) + \sum_{k=1}^{n}\sum_{r=1}^{n}\Bigg[\sum_{i=1}^{m}\sum_{j=1}^{m} d(g(S_i^k,S_j^k),g(S_i^r,S_j^r))\Bigg]\Bigg] \quad (15)$$

and is equal to the problem as

$$l = \arg\max_{l^*} \sum_{(k,r)} \frac{\sum_{(i,j)\in m} M_1^{kr}(i,j)}{\sum_{(i,j)} |M_2^k(i,j) - M_2^r(i,j)|} \quad (16)$$

where $M_1^{kr}(i,j)$ is the matching scores between the labels l_i and l_j, which is based on the matching of image pairs (I_k, I_r), and is defined as

$$M_1^{kr}(i,j) = \frac{\sum_{p\in l_i^k}\delta(p'\in l_j^r)}{N_{lp}} \quad (17)$$

where l_i^k is the pixels of I_k in the region l_i, $p' \in I_r$ is the matched pixel of $p \in l_i^k$, N_{lp} is the number of pixels in l_i^k.

Meanwhile, $M_2^k(i,j)$ is the spatial distance between region pair (S_i, S_j) in image I_k, and $|M_2^k(i,j) - M_2^r(i,j)|$ is the spatial relationship consistency evaluation. Small values indicate a good consistency. It is seen from (16) that the region pairs with good label match will have many matching pixels, and lead to large value of M_1^{kr}. In addition, the consistent part region labels will have similar M_2, and result in small value of denominator. Hence, the best region label has the largest value of the fraction. In this paper, we use gradient descent to minimize the problem in (16) with grid based initial value setting. Note that when the number of part is small, traversal method can also be used to search global solution with fast speed.

After continuously performing the three sub-minimization problems, we finally obtain the approximate solution of our model. Algorithm 1 shows the process of our model.

IV. Experimental Results

In this section, our method is verified by subjective and objective results. A part dataset constructed from three image datasets such as PASCAL 2010 part datasets, Caltech-UCSD Birds dataset, Cat-Dog dataset, and one video dataset such as UCF Sports Actions dataset, is used for the verification.

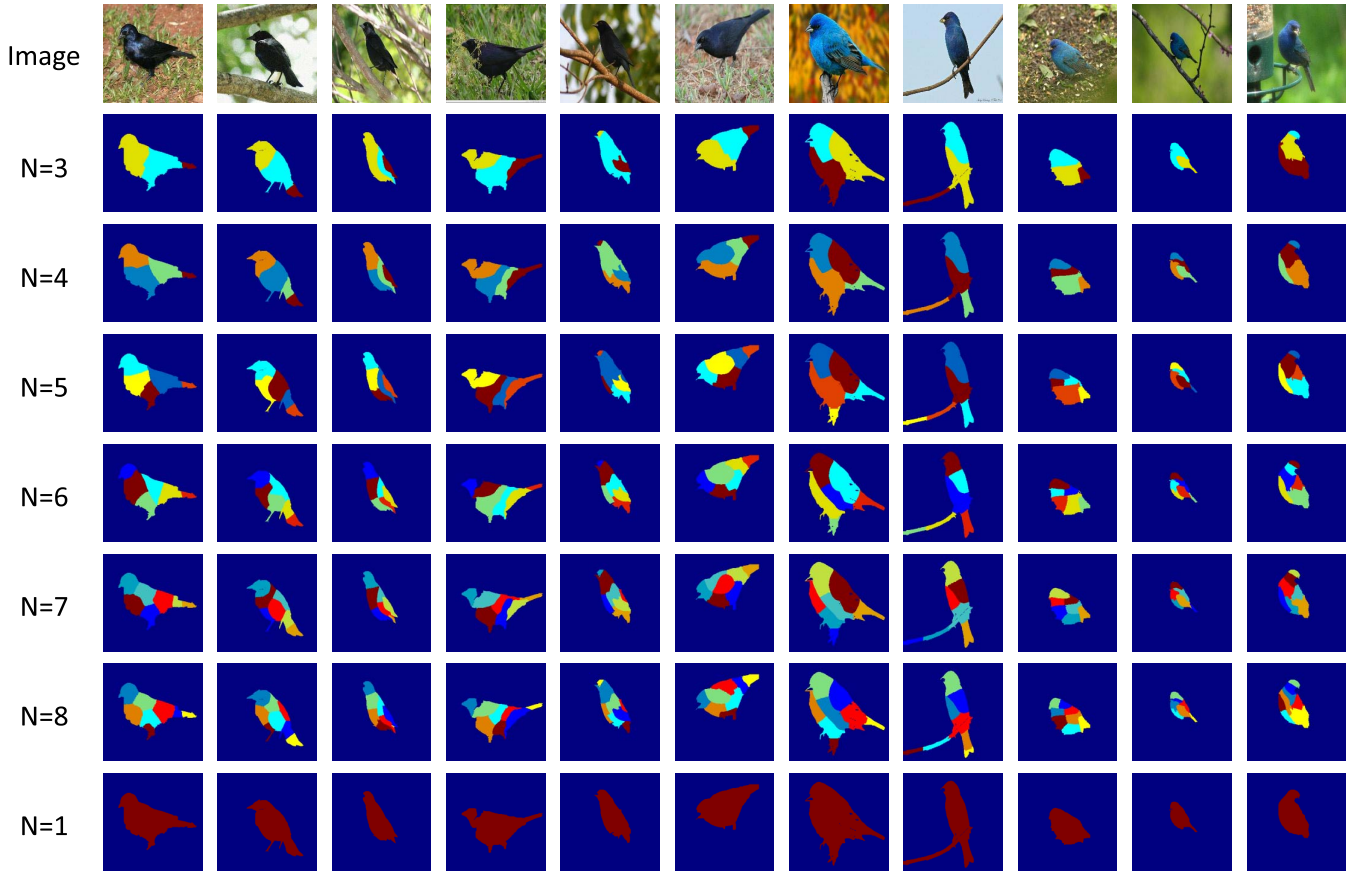


Fig. 8. Our part proposal segmentation results on Caltech-UCSD Birds dataset. Each rows are the same to Fig. 7.

A. Implementation Details

The shift map M_s is comprised of vectors, which is represented by two channels: length and angle. Because the shift vectors gradually change based on the shape context matching, the value distances of neighbor pixels are small, which leads to unsuccessful segmentation. Hence, we refine the two channels by replacing the value into the class centers, which are obtained by k-means algorithm. Denoting the number of classes in k-means as N_s, we adjust N_s to obtain multiple layer of proposals. Meanwhile, the number of part N is also adjusted to obtain the proposals. Hence, our model adjusts two parameters N and N_s for the proposal generation. In this paper, we empirically set $N_s = 4$ and $N_s = N$, and $N \in [3, 8]$, with the consideration of computational cost.

B. Subjective Results

Some part proposal generation results are shown in Fig. 7, 8, 9 and 10 for the four datasets. These results are obtained based on $N_s = 4$. The original image is shown in the first row. The segmentation results with $N = 3$ to 8 are displayed from the second row to the seventh row, respectively. The bottom row shows the results by $N = 1$, which are the cosegmentation results. We can see from the results of $N = 1$ that the common objects with similar shapes can be segmented from these images, such as "Cow" in Fig. 7, and "Girl" in Fig. 10. This indicates that our cosegmentation guarantees the following part segmentation.

Furthermore, by seeing each row, it is seen that part proposals have been matched well, although there may have noise regions caused by cosegmentation. For example, in Fig. 7, each part of "Cow" has been matched in the row of $N = 6$, and the matches keep spatial consistency, such as the relationship of "Head" and "Leg". These results indicate the effectiveness of our spatial consistency constraints in E_h.

The results also show that the local part can be represented by one of segmentation layers. For example, the "Head" regions are extracted in layers of $N = 6$ and $N = 5$ in the Fig. 8 and Fig. 9, respectively. Meanwhile, the "Tails" are segmented in both the layers of $N = 8$ for the two set of images. This indicates the fact that shape variation can provide part regions.

It is also seen that there are failure cases in these segmentation results, such as the sixth image of "Cat" in Fig. 9, and the last segmentation results in Fig.10. We also display other more failed segmentation results in Fig. 11, where three images and their segmentation are displayed. The unsuccessful segments are mainly caused by the initial object segmentation. When the object segmentation is inaccurate, wrong matching will be performed, which will lead to failed segmentation. Note that when most of the object regions are successfully extracted, good segmentation can also be obtained, since the failed shape matching can be corrected by these success segmentation.

C. Objective Results

We next verify our method based on the objective value. The IOU value is used in our experiment, which is defined as $\frac{F \cap G}{F \cup G}$, where F and G are the regions of the segment and

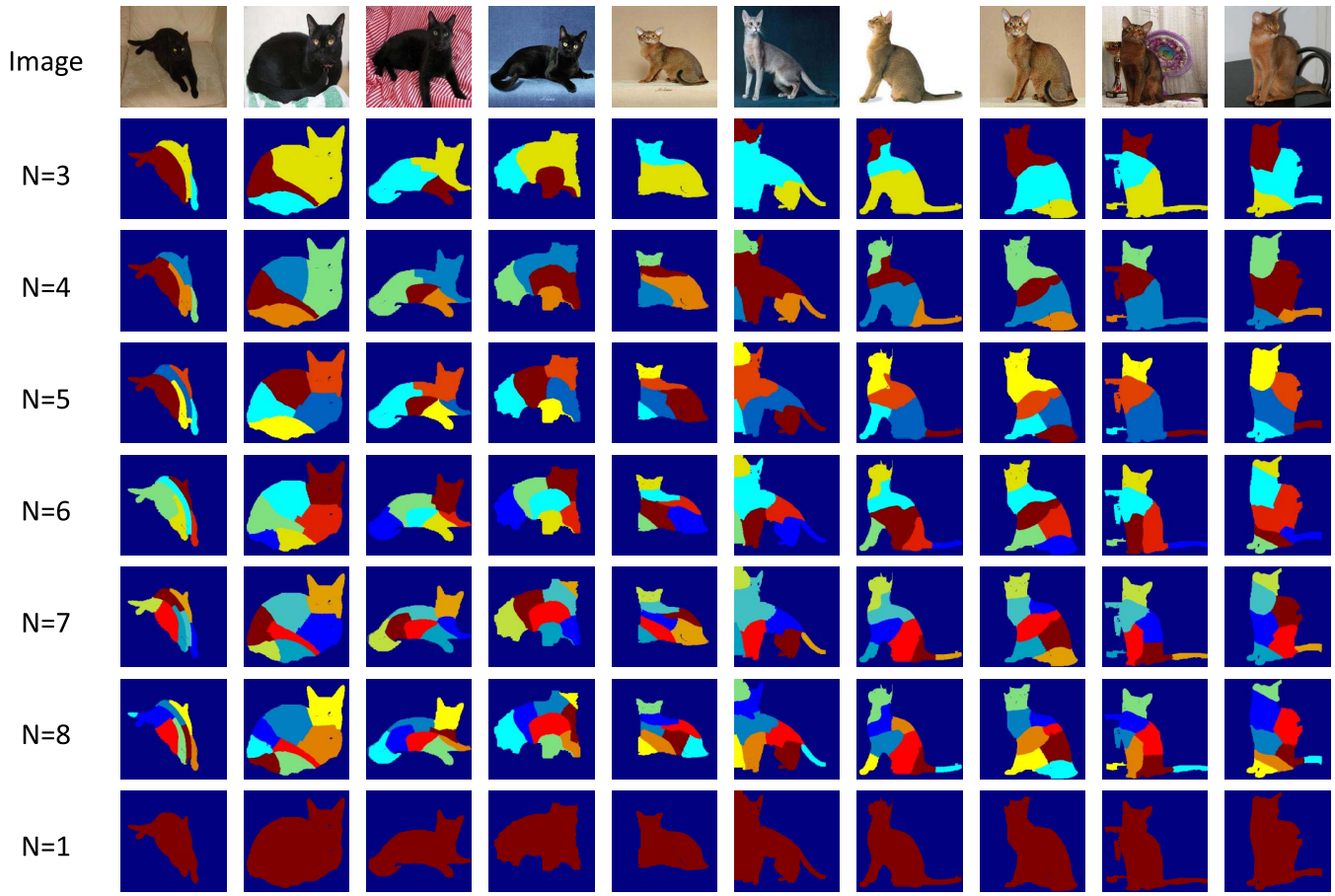


Fig. 9. Our part proposal segmentation results on Cat-Dog datasets. Each rows are the same to Fig. 7.

the groundtruth, respectively. A large IOU value indicates a good segmentation. Since there are multiple images and parts, we evaluate the results based on the regions of each part. Given a part, we have n groundtruth regions. Then we select one label regions, and calculate their average IOU values compared with the groundtruth regions, and use the largest value as the IOU value of this part. The final object value is the average of all parts. We can see although the segmentation results are only one label regions, the IOU value can still be obtained for the evaluation. Hence, we can compare our method with both part level and region level methods.

We show our objective results in the last row of Table I, where the IOU values of the classes are shown. The average IOU values are displayed in the last column. It is seen from the results that the IOU value is 0.186, which is low. This is caused by the fact that part segmentation is a very challenging task with the needs to obtain consistent part regions among images. It is also seen that the value of Pascal 2010 dataset is lower than the other datasets, which is caused by the large object variations among images and the complicated backgrounds.

The results with setting $N_s = 4$ and $N_s = N$ are also displayed for comparison. It is seen that the proposal results are affected by the setting of N_s. In addition, the average IOU value of $N_s = N$ is 0.179, which is better than $N_s = 4$ of 0.168. But their combination obtains the average IOU value of 0.186, which is the best one among the results.

We next display the average IOU values along with N in Fig. 12, where x-axis is the part number, and y-axis is the average IOU values. It is seen from the curve that the average IOU values become larger along with N, which is caused by the fact that there are some small part regions, such as "Ear" and "Eye", and the set of large part number can obtain more detailed regions that covers these small part, and leads to larger average IOU values.

We also compare our method with four existing methods, including weak object level segmentation [30], weakly supervised part segmentation [4], interactive cosegmentation [33] and interactive binary region segmentation [34]. The codes of these methods publicly released by the authors are used. The method in [30] is a cosegmentation method that extracts similar shape common regions from multiple images. The method in [4] is a recent part segmentation method that extracts part regions based on average sampling. In the implementation of [4], we replace the CNN feature to the shape feature for simplification, since authors indicate the robustness of the method to feature selection. Furthermore, since the results of [4] are windows instead of regions, we use the windows of segment and groundtruth for calculating the IOU value. The method in [33] designs a cosegmentation model by three cues such as user interaction, local smooth and

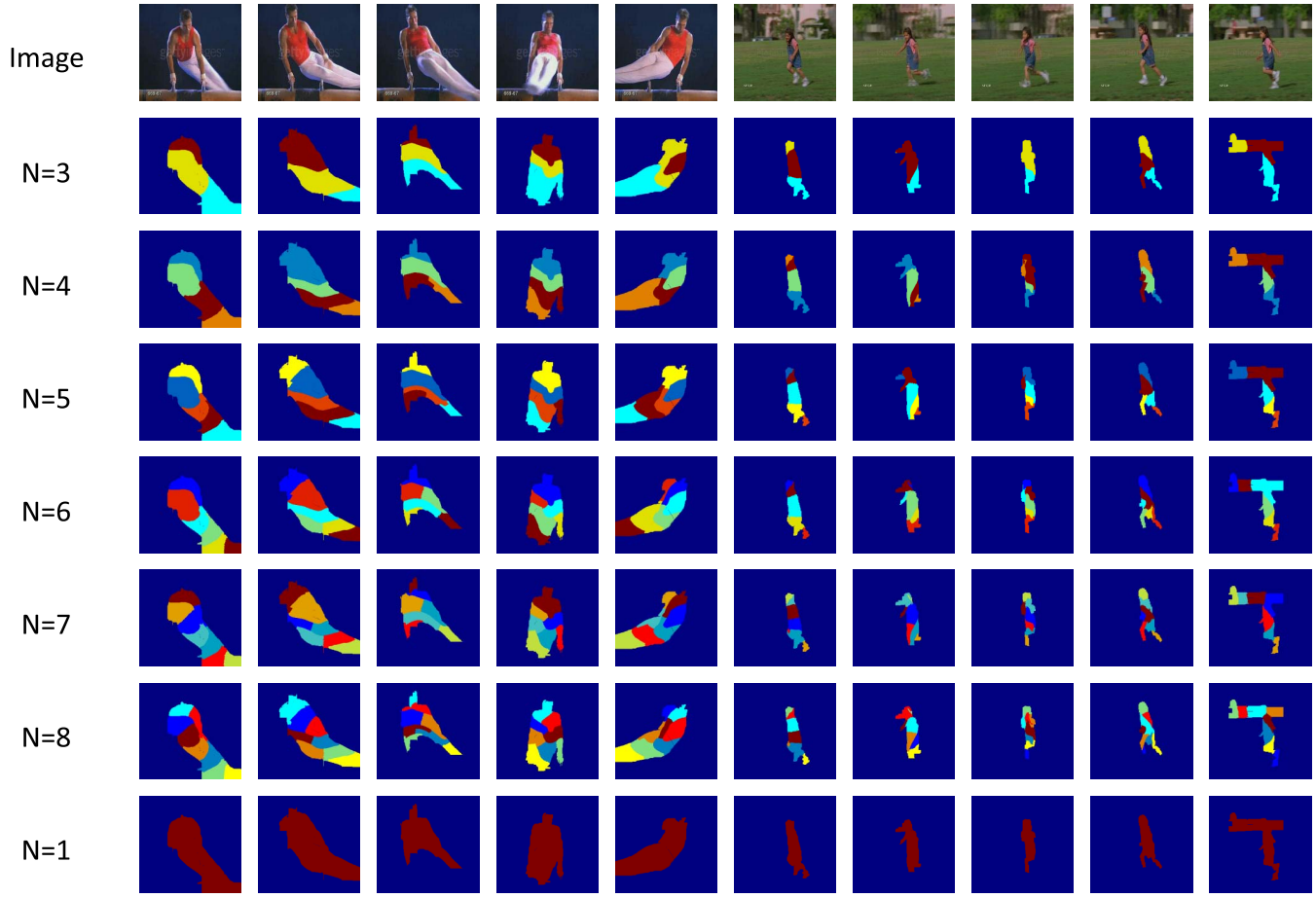


Fig. 10. Our part proposal segmentation results on UCF sports action dataset. Each row is the same to Fig. 7.

TABLE I

THE OBJECTIVE VALUES OF OUR METHODS, AND THE COMPARISON RESULTS

	PASCAL 2010 Part Dataset							Caltech-UCSD Birds		
Method	Bird	Cat	Cow	Dog	Horse	Person	Sheep	Albatross	Cowbird	Bunting
$N_s = 4$	0.089	0.097	0.112	0.126	0.066	0.112	0.091	0.406	0.259	0.215
$N_s = N$	0.109	0.112	0.117	0.137	0.073	0.123	0.101	0.365	0.256	0.234
[30]	0.064	0.053	0.049	0.047	0.026	0.032	0.043	0.143	0.144	0.128
[33]	0.118	0.120	0.097	0.069	0.089	0.039	0.093	0.039	0.119	0.141
[34]+Part	0.530	0.704	0.658	0.641	0.651	0.748	0.699	0.745	0.625	0.686
[4]	0.099	0.135	0.115	0.141	0.067	0.106	0.105	0.272	0.267	0.233
ours+Combination	0.111	0.113	0.124	0.142	0.075	0.128	0.106	0.407	0.273	0.238

	Cat-Dog Dataset			UCF Sports Actions		
Method	Bombay	Abyssinian	bulldog	Run-side11	Swing04	Average
$N_s = 4$	0.165	0.197	0.198	0.208	0.180	0.168
$N_s = N$	0.182	0.231	0.222	0.235	0.195	0.179
[30]	0.080	0.080	0.080	0.076	0.098	0.076
[33]	0.086	0.100	0.029	0.021	0.040	0.080
[34]+Part	0.513	0.731	0.324	0.438	0.796	0.633
[4]	0.172	0.222	0.170	0.287	0.241	0.176
Ours+Combination	0.185	0.233	0.223	0.236	0.200	0.186

foreground consistency. The model is converted to constrained quadratic programming problem, with a simple iterative solution. By a few of user interactions, better segmentation results are obtained. The method in [34] is a user-interaction based binary segmentation model. Three constraints such as shape convexity, user-defined hard constraint and other standard constraints are considered. The model is efficiently minimized by trust region approach. Meanwhile, we observe that such model can segment multiple part regions by separately scrawling foreground and background seeds for each part due to the shape convexity constraint. Hence, we simply change the interaction-based binary segmentation [34] to interaction-based part segmentation by implementing the method in [34] for each part separately, and name it as [34]+Part.

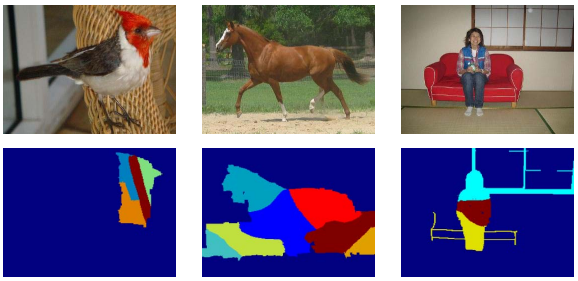


Fig. 11. The original images and the failed segmentation results, which is caused by the unsuccessful object region extraction.

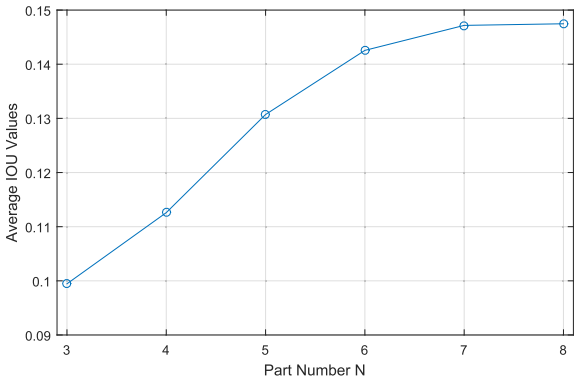


Fig. 12. The average IOU values by varying the part number N. It is seen that the average IOU values become larger along with N.

The objective results by the comparison methods are shown in Table I. It is seen that the average IOU value by [34]+Part (0.633) is obviously larger than our method (0.186) and the comparison methods due to the consideration of part-level segmentation and user interaction. Meanwhile, our method outperforms object-level segmentation methods [33] (0.080) and [30] (0.076) due to the fact that cosegmentation obtains object-level regions rather than part-level regions. Furthermore, the results of [4] is 0.176, which is better than our results with $N_s = 4$. Meanwhile, our final result (0.186) is larger than the method in [4], which demonstrates the usefulness of shape variation in part proposal segmentation.

D. Discussions

Our method is a weakly supervised part proposal generation method, which aims at segmenting part regions from a set of images with image-level labels. This is a new research topic in weakly supervised segmentation. Although a little bit of weakly supervised part segmentation methods [4] have been proposed recently, the basic problem on how to efficiently define a part region is still underdeveloped. Compared with the existing weakly supervised part segmentation methods, we design the part-level segmentation model in a new view of pose variation that is a different and efficient cue for discovering part regions. Furthermore, our method is different from our previous cosegmentation work [30] that is used in our model. The difference is that the work in [30] aims at extracting common objects from multiple images, which first describes the relationships of the regions by digraph, and then formulates cosegmentation as shortest path problem. But our method aims at segmenting more detailed part regions, which is a more difficult problem. In addition, our previous work [30] is used here to simplify the minimization of the cosegmentation term of our model. Note that other cosegmentation minimizations such as belief-propagation-based method [31] can also be used to replace our previous work [30].

In our model, as the usual assumption in common object segmentation [1], [29], [35], we assume that similar objects share similar feature f, such as the mid-level shape feature used in our method. The successful segmentation of similar objects can be guaranteed when the feature f describes the object similarity well. However, when the objects are different greatly by f, the foreground consistency term E_c in (5) cannot correctly evaluate the similarities of regions. Wrong segmentation will be obtained. Note that these wrong segmentation results can be avoided by using more effective feature. We will study more robust and adaptive feature description such as deep learning based feature extraction in the future to further improve our model.

V. Conclusions

In this paper, a weakly supervised part region segmentation method is proposed. Object pose variations are captured to obtain the fixed regions, which is then used to generate the local parts. Four aspects, such as shape feature based cosegmentation, shape matching based variation capture, NCuts based part proposal generation, and second order graph matching based part label matching, are considered. The four aspects are combined to form our energy function, which is minimized by three sub-minimization steps, including cosegmentation, NCuts segmentation and label matching. Our method is verified on three image datasets and one video dataset. The experimental results demonstrate the effectiveness of the proposed method.

References

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

[2] 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.

[3] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

[4] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5546–5555.

[5] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2648–2655.

[6] J. Wang and A. L. Yuille, "Semantic part segmentation using compositional model combining shape and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1788–1795.

[7] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Joint object and part segmentation using deep learned potentials," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1573–1581.

[8] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[9] 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.
[10] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[11] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[13] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. CVPR*, Jun. 2010, pp. 73–80.
[14] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 297–312.
[15] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2417–2424.
[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. (2013). "Rich feature hierarchies for accurate object detection and semantic segmentation." [Online]. Available: https://arxiv.org/abs/1311.2524
[17] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. Int. Conf. Comput. Vis.*, Jul. 2001, pp. 105–112.
[18] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
[19] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 845–852.
[20] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, "Weakly-supervised dual clustering for image semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2075–2082.
[21] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1908–1915.
[22] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, Feb. 2013.
[23] L. Zhang, Y. Yang, Y. Gao, Y. Yu, C. Wang, and X. Li, "A probabilistic associative model for segmenting weakly supervised images," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4150–4159, Sep. 2014.
[24] T. Ma and L. J. Latecki, "Graph transduction learning with connectivity constraints with application to multiple foreground cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1955–1962.
[25] F. Meng, H. Li, K. N. Ngan, L. Zeng, and Q. Wu, "Feature adaptive co-segmentation by complexity awareness," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4809–4824, Dec. 2013.
[26] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 542–549.
[27] G. Kim and E. P. Xing, "On multiple foreground cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 837–844.
[28] F. Meng, H. Li, and J. Cai, "On multiple image group cosegmentation," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 258–272.
[29] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based RGBD image co-segmentation with mutex constraint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4428–4436.
[30] 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.
[31] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2011, pp. 2217–2224.
[32] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. ECCV*, 2014, pp. 391–405.
[33] X. Dong, J. Shen, L. Shao, and M.-H. Yang, "Interactive cosegmentation using global and local energy optimization," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3966–3977, Nov. 2015.
[34] L. Gorelick, O. Veksler, Y. Boykov, and C. Nieuwenhuis, "Convexity shape prior for binary segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 258–271, Feb. 2017.
[35] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 2028–2035.

Fanman Meng (S'12–M'14) received the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China, Chengdu, China, in 2014. From 2013 to 2014, he joined the Division of Visual and Interactive Computing, Nanyang Technological University, Singapore, as a Research Assistant. He is currently an Associate Professor with the School of Electronic Engineering, University of Electronic Science and Technology of China . He has authored or co-authored numerous technical articles in well-known international journals and conferences. His research interests include image segmentation and object detection. He received the Best Student Paper Honourable Mention award for the 12th Asian Conference on Computer Vision, Singapore, in 2014, and the Top 10% Paper Award in the IEEE International Conference on Image Processing, Paris, France, in 2014. He is a member IEEE CAS society.

Hongliang Li (SM'12) received the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, China, in 2005. From 2005 to 2006, he joined the Visual Signal Processing and Communication Laboratory, The Chinese University of Hong Kong, (CUHK) as a Research Associate a Post-Doctoral Fellow the Visual Signal Processing and Communication Laboratory, CHUK, from 2006 to 2008. He is currently a Professor with the School of Electronic Engineering, University of Electronic Science and Technology of China. He has authored or co-authored numerous technical articles in well-known international journals and conferences. He is a co-editor of a Springer book titled *Video segmentation and its applications*. His research interests include image segmentation, object detection, image and video coding, visual attention, and multimedia communication system. He was involved in many professional activities. He is a member of the Editorial Board of the Journal on Visual Communications and Image Representation, and the Area Editor of the *Signal Processing: Image Communication* and the *Elsevier Science*. He served as a Technical Program Co-Chair of VCIP2016 and ISPACS 2009, a General Co-Chair of the ISPACS 2010, a Publicity Co-Chair of IEEE VCIP 2013, the Local Chair of the IEEE ICME 2014, and the TPC Member in a number of international conferences, including, ICME 2013, ICME 2012, ISCAS 2013, PCM 2007, PCM 2009, and VCIP 2010.

Qingbo Wu (S'12–M'13) received the B.E. degree in education of applied electronic technology from Hebei Normal University in 2009, and the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China in 2015. From 2014 to 2014, he was a Research Assistant with the Image and Video Processing Laboratory, The Chinese University of Hong Kong. From 2014 to 2015, he served as a Visiting Scholar with the Image and Vision Computing Laboratory, University of Waterloo. He is currently a Lecturer with the School of Electronic Engineering, University of Electronic Science and Technology of China. His research interests include image/video coding, quality evaluation, and perceptual modeling and processing.

Bing Luo received the B.Sc. degree in communication engineering from The Second Artillery Command College, Wuhan, China, in 2009, and the M.Sc. degree in computer application technology from Xihua University, Chengdu, China, in 2012. He is currently pursuing the Ph.D. degree with the University of Electronic Science and Technology of China, Chengdu, under the supervision of Prof. H. Li. His research interests include image and video segmentation and machine learning.

King Ngi Ngan (F'00) received the Ph.D. degree in electrical engineering from the Loughborough University, U.K. He is currently a Chair Professor with the Department of Electronic Engineering, The Chinese University of Hong Kong. He was a Full Professor with the Nanyang Technological University, Singapore, and The University of Western Australia, Australia. He has been appointed a Chair Professor with the University of Electronic Science and Technology, Chengdu, China, under the National Thousand Talents Program since 2012. He has authored extensively, including three authored books, seven edited volumes, over 380 refereed technical papers, and edited nine special issues in journals. In addition, he holds 15 patents in the areas of image/video coding and communications. He is a member of IET, U.K., and IEAust, Australia, and the IEEE Distinguished Lecturer from 2006 to 2007. He holds honorary and Visiting Professorships of numerous universities in China, Australia and South East Asia. He served as an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Journal on Visual Communications and Image Representation*, *EURASIP Journal of Signal Processing: Image Communication*, and *Journal of Applied Signal Processing*. He chaired and co-chaired a number of prestigious international conferences on image and video processing including the 2010 IEEE International Conference on Image Processing, and served on the advisory and technical committees of numerous professional organizations.