Group-wise Deep Object Co-Segmentation with Co-Attention Recurrent Neural Network

Bo Li
Nanjing University
njumagiclibo@gmail.com

Zhengxing Sun
Nanjing University
szx@nju.edu.cn

Qian Li
Nanjing University
public_liqian@163.com

Yunjie Wu
Nanjing University
JiejiangWu@outlook.com

Anqi Hu
Nanjing University
dz1533009@smail.nju.edu.cn

Abstract

Effective feature representations which should not only express the images individual properties, but also reflect the interaction among group images are essentially crucial for real-world co-segmentation. This paper proposes a novel end-to-end deep learning approach for group-wise object co-segmentation with a recurrent network architecture. Specifically, the semantic features extracted from a pre-trained CNN of each image are first processed by single image representation branch to learn the unique properties. Meanwhile, a specially designed Co-Attention Recurrent Unit (CARU) recurrently explores all images to generate the final group representation by using the co-attention between images, and simultaneously suppresses noisy information. The group feature which contains synergetic information is broadcasted to each individual image and fused with multi-scale fine-resolution features to facilitate the inferring of co-segmentation. Moreover, we propose a group-wise training objective to utilize the co-object similarity and figure-ground distinctness as the additional supervision. The whole modules are collaboratively optimized in an end-to-end manner, further improving the robustness of the approach. Comprehensive experiments on three benchmarks can demonstrate the superiority of our approach in comparison with the state-of-the-art methods.

1. Introduction

Object co-segmentation aims at discovering and segmenting the co-occurring objects from the given set of images containing the same or similar objects. Unlike the single image object segmentation methods which individually segment image only using the information within one single image, object co-segmentation methods can further exploit the synergetic relationship among the multiple relevant images for higher accuracy. With such property, object co-segmentation can benefit various computer vision applications and beyond, such as image matching [4], weakly supervised learning [32], video object segmentation [18], and 3D reconstruction [36].

In order to segment the co-occurring objects accurately, two issues should be concerned: 1) extract and learn effective feature representations of images in the group; 2) model the synergetic relationship among the common objects to generate the final co-segmentation results. For 1), feature representation in the co-segmentation task should not only express the images individual properties, but also reflect the relevance and interaction between group images. For 2), the synergetic relationship such as common objects, similar categories, and related scenes should be fully exploited at group level. Therefore, the co-segmentation job can apply the single and the group-wise information to model the interaction between the images so that they mutually facilitate each other for the final co-segmentation results.

Conventional co-segmentation approaches [6, 9, 20, 21, 22, 24, 28] utilize handcrafted features to represent images, such as color, texture and SIFT descriptors etc., and these methods rely on researchers prior knowledge like objectness and saliency to model the interaction between the group images. However, low-level features and predefined prior knowledge are too subjective to face the multiple challenges including background clutter, appearance variance of co-object across images, and similarity between co-object and non-common object, etc. Encouraged by the success of deep learning in many computer vision tasks, recent researches [12, 19, 33] improve object co-segmentation by using deep neural network (DNN) to learn visual representation or segmentation inference module in a data driven manner. However, as feature extraction and object segmentation are treated as separate steps in these approaches, the learned features are not tailored for segmenting the co-

occurring objects, resulting in suboptimal performance. As mentioned by Han *et al*. [8], it is hard to design the network architecture for co-segmentation task, since the dimension of DNNs input data should usually be constant, whereas the number of the contained images is not constant both in training and testing. The very recent works [3, 15] integrated the process of feature learning as well as co-segmentation inferring as an organic whole and proposed end-to-end deep learning methods for co-segmentation. For designing a feasible network, they simply fixated the number of input images to two and then segment the co-object in a pairwise manner. However, seeking co-segment objects in two images at a time can only utilize the limited relationship between the image pair, which damages the robustness of co-segmentation when extending beyond pairwise relations and limits its practical application value. So, how to design an end-to-end network architecture for group-level object co-segmentation is still an unsolved challenge.

In this paper, we propose a novel end-to-end deep learning approach for group-wise object co-segmentation with a recurrent network architecture. Unlike the previous pairwise methods, our aim is to create a robust and effective co-segmentation network by making use of all available information including individual image properties and the group-level synergetic relationships to meet the need for real-world applications. Specifically, our network first extracts the semantic features of all images, then the features are processed by two branches. The *single image representation* branch processes each image individually to learn the unique properties. Meanwhile, the *group-wise representation* branch can gradually explore all images in the group to learn a robust group-wise representation by introducing the recurrent architecture. The group feature is then broadcasted to each individual image and fused with single image feature, which allows the network to sufficiently exploit the complementarity and interaction of group and single representation to facilitate the final co-segmentation reasoning. Particularly, instead of using conventional recurrent neural network like LSTM [11] and GRU [5], we specially design a novel Co-Attention Recurrent Unit (CARU) to handle the variation of co-object in appearance and the location across images by using the spatial and channel co-attention between images. Moreover, to make full use of the interactive relationships of whole images in the training group, we further propose a group-wise training objective as the additional supervision in our end-to-end training process.

Our main contributions can be summarized as follows. 1) We make one of the earliest efforts to formulate the object co-segmentation in an end-to-end manner. As the first attempt to introduce the recurrent architecture into deep co-segmentation, our network can simultaneously segment the co-occurring objects at group-level. 2)We design a novel Co-Attention Recurrent Unit to recurrently learn a robust group feature, which can handle the variation of co-object in appearance and the location across images and meanwhile suppresses noisy information like irrelevant background and non-common object. 3) We propose a group-wise training objective to utilize the co-object similarity and figure-ground distinctness as the additional supervision. 4) Comprehensive experiments on three widely used datasets for object co-segmentation have demonstrated the superiority of the proposed approach as compared to the state-of-the-art methods.

2. Related Work

The concept of co-segmentation was first introduced by Rother *et al*. [21], who used histogram matching to simultaneously segment out the common object from a pair of images. Following this work, a number of researchers have made further efforts to develop more effective object co-segmentation models by comparing foreground color histograms [28] or adopting more diverse features like Gabor filters [10] and SIFT [22]. For example, Rubinstein *et al*. [22] combined a visual saliency and dense SIFT matching to capture the sparsity and visual variability of the common object in a group of images. In order to better explore the correspondence relationship among common objects, some existing methods additionally introduced prior constraints to better distinguish them from the undesired image backgrounds. References [6, 24] first extracted prior information of the common objects from the object class by using shape templates [6] and part detectors [24], and then applied the extracted prior information to segment the common object of each image. References [13, 29] exploited the objectness [29] prior and saliency prior [13] to constrain the obtained foreground segments in each image. However, these methods cannot obtain robust performance in real-world scenarios, where the handcrafted low-level features are too subjective to face the multiple challenges including intra-class variations and background clutters and the predefined prior knowledge cannot always provide adequate and precise constraint on the common objects.

Deep learning has recently emerged and demonstrated success in many computer vision applications. Recent researches [12, 19, 33] use deep visual features to improve object co-segmentation and they also try to learn more robust synergetic properties among images in a data driven manner. Yuan *et al*. [33] introduced a DNN-based dense conditional random field framework for object co-segmentation by cooperating co-occurrence maps which are generated using selective search [27]. Hsu *et al*. [12] proposed a DNN-based method which uses the similarity between images in deep features and an additional object proposals algorithm [14] to segment the common objects. These methods achieved state-of-the-art results by substituting the features learned by DNN for engineered features.

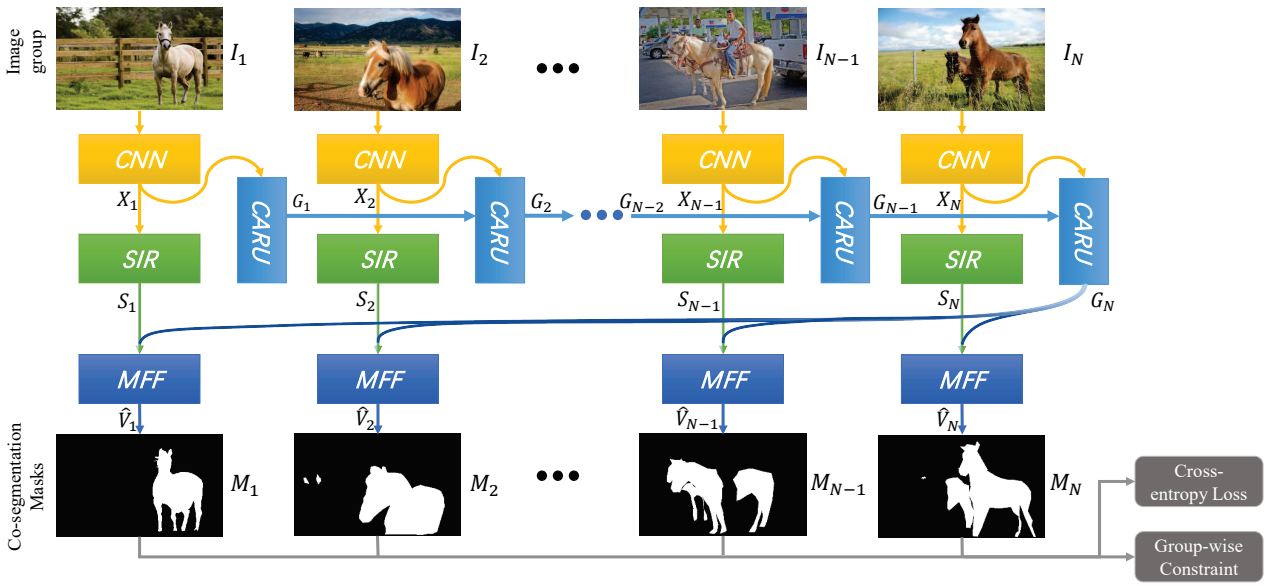

Figure 1. Illustration of the proposed recurrent network architecture for object co-segmentation.

However, as feature learning and object segmentation are somehow separated in these approaches, the learned features are not tailored for segmenting the co-occurring objects, resulting in suboptimal performance. The very recent works [3, 15] proposed end-to-end deep learning methods for co-segmentation by integrating the process of feature learning and co-segmentation inferring as an organic whole. By introducing the correlation layer [15] or a semantic attention learner [3], they can utilize the relationship between the image pair and then segment the co-object in a pairwise manner. However, their siamese network structures limit their use of group-wise information which contains more sufficient synergetic relationships than image pairs. Consequently, co-segmenting common objects from image pairs has very limited robustness and practical application value when extending beyond pairwise relations. Unlike the previous methods, by introducing the recurrent architecture, our co-segmentation network is able to make use of all available information including individual image properties and the group-level synergetic relationships to meet the need for real-world applications.

One closely related topic to object co-segmentation is co-saliency detection [34], which aims at generating co-saliency maps for each of the images from the given image collection to highlight the common and salient objects. Compared with co-saliency detection, object co-segmentation only aims at segmenting the co-occurring objects without constraining those objects to be (co-)salient. Even though, the co-saliency maps generated by co-saliency detection can still be used as prior knowledge for object co-segmentation [2, 16, 26]. As an end-to-end work, our method needs no such predefined prior knowledge as additional information in object co-segmenting.

3. Proposed Approach

3.1. Problem Formulation

Object co-segmentation aims at discovering and segmenting the co-occurring objects from a group of N relevant images $\mathcal{I} = \{I_n\}_{n=1}^N$. The co-occurring object masks $\mathcal{M} = \{M_n\}_{n=1}^N$ are produced by a co-segmentation model:

$$\mathcal{M} = F(\mathcal{I}; \Theta), \tag{1}$$

where $F()$ is the model function that takes an image group as input and outputs a group of co-segmentation results simultaneously. Θ represents model parameters which are optimized by an end-to-end learning scheme in this work. The core idea of this work is trying to make full use of all available information to learn the effective feature representations which can not only express the images individual properties, but also reflect the interaction among group images for robust co-segmentation referring. The overall architecture of the proposed approach is illustrated in Figure 1. For an input image group with an arbitrary size, our network first extracts the semantic features of all images. Then the *single image representation* (SIR) branch processes each image individually to learn the unique properties. Meanwhile the Co-Attention Recurrent Unit (CARU) in the *group-wise representation* branch recurrently explores all images in the group to learn the robust group representation. The two branches are later merged in the Multi-scale Features Fusion (MFF) module for the final object co-segmentation referring.

3.2. Single Image Representation

As a basic rule in object segmentation, it is important to learn the unique properties of each image to capture poten-

tial co-occurring objects in the individual image. In the proposed approach, for each image I_n in the input group $\mathcal{I}$ we first use a pre-trained convolutional neural network (CNN) VGG19 [23] to extract the pooled features ($Pool5$) as its semantic features $X_n \in \mathcal{R}^{H\times W\times C}$. Then we construct an SIR block with 3 convolutional layers to encode the individual properties for each image $S = \{S_n\}_{n=1}^{N}$, which is defined as follows:

$$S_n = f_S(X_n; \Theta_S), \tag{2}$$

where Θ_S are the parameters learned from the convolutional process f_S.

3.3. Group Representation with CARU

As images within a co-segmentation group are contextually associated with each other in different ways such as common objects, similar categories, and related scenes, learning a robust group representation which contains the relevance and interaction between group images is extremely important for co-segmentation referring. In this work, we proposed to use a recurrent architecture to learn the group representation G_N for an arbitrary size group $\mathcal{I}$. It is defined as follows:

$$G_N = f_G(\{X_n\}_{n=1}^{N}; \Theta_G), \tag{3}$$

where Θ_G are the parameters learned from the recurrent convolutional process f_G. Since there is much noise information of irrelevant background and non-common object in the group and the appearance as well as the location of co-occurring object varies across images, we specifically design a novel recurrent unit CARU to gradually explore all the synergetic relationships between images for the group representation. As illustrated in Figure 2(a), for step n, CARU takes two inputs: the current image feature map X_n and the group representation G_{n-1} of all the images been explored. In the first step, G_0 is initialized with X_1. We construct two gates in our CARU, the reset gate g_d is used for denoising current image and the update gate g_z is used to decide how to update current group representation G_n.

Since all images in the group share the common objects, we want to use the synergetic relationships between the explored images and the current image to suppress the noise data in current image, like the irrelevant background and non-common object. The reset gate D is defined as:

$$D = g_d([G_{n-1}, X_n]). \tag{4}$$

Then the denoised feature map $\tilde{X}_n$ is computed as:

$$\tilde{X}_n = X_n \odot D = X_n \odot sigmoid(\mathbf{W}_d \times [G_{n-1}, X_n]), \tag{5}$$

where $[G_{n-1}, X_n] \in \mathcal{R}^{H\times W\times 2C}$ is the concatenated feature map of G_{n-1} and X_n, $\times$ is matrix multiplication and $\odot$ denotes element-wise multiplication. We use a FC layer $\mathbf{W}_d$ to reduce the feature dimension.

As the appearance and the location of co-occurring object varies across images, we want to fully explore the spatial-channel-wise variation of the co-occurring object object with co-attention mechanism to determine what group information should be retained in G_{n-1} and what new information should be updated from $\tilde{X}_n$. So $Z_n = (G_{n-1} - X_n)$ is used as an input of g_z to model the cross-images variation in each step. The update gate Z is defined as follows:

$$Z = g_z(Z_n; \Theta_z) = g_z(G_{n-1} - X_n). \tag{6}$$

Figure 2(b) illustrates the structure of update gate model g_z.

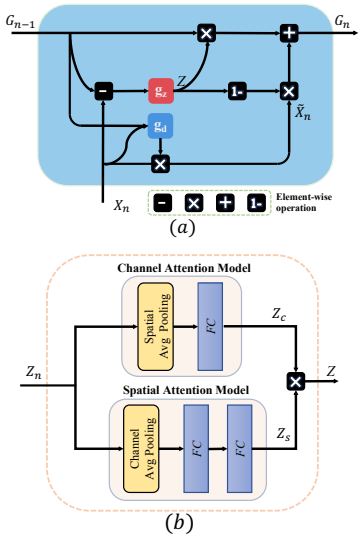

Figure 2. (a) The architecture of co-attention recurrent unit (CARU). (b) The architecture of proposed update gate model g_z.

For spatial attention model, we first use a global cross-channel average pooling layer to get the overall response in each spatial position.Then two FC layers (the first layer is followed by ReLU) are applied to generate the spatial attention maps $Z_s \in \mathcal{R}^{H\times W\times 1}$. It is formulated as:

$$Z_s = \mathbf{W}_s^2 \times ReLU(\mathbf{W}_s^1 \times Z_n^{H,W}), \tag{7}$$

where $Z_n^{H,W} \in \mathcal{R}^{H,W}$ is the result of Z_n after cross-channel average pooling.

For channel attention model, we first introduce a global spatial space average pooling layer to get overall response of each channel. Then a FC layer is applied to get the channel attention maps $Z_c \in \mathcal{R}^{1\times 1\times C}$, which is formulated as:

$$Z_c = \mathbf{W}_c \times Z_n^C, \tag{8}$$

where $Z_n^C \in \mathcal{R}^C$ is the result of Z_n after global spatial space average pooling.

The overall attention maps of current input feature are the product of spatial attention maps and channel attention maps. After a $sigmoid$ operation, the overall attention maps are normalized into the range between 0 and 1,

$$Z = sigmoid(\, Z_s \odot Z_c \,), \tag{9}$$

where $Z \in \mathcal{R}^{H\times W\times C}$. Then the CARU updates G_n by

$$G_n = Z \odot G_{n-1} + (1 - Z) \odot \tilde{X}_n. \tag{10}$$

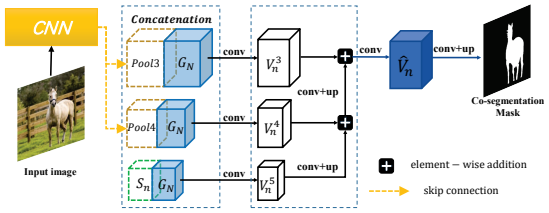

Figure 3. Illustration of the multi-scale features fusion.

The value of each position in Z denotes the probability for the activation value in corresponding position of G_{n-1} to be reserved and position of $\tilde{X}_n$ to be updated. Higher probability value indicates that the update gate model g_z considers that last group feature G_{n-1} in this location has high quality group information and should be reserved. While location with lower probability means there is new useful synergetic information in $\tilde{X}_n$ should be updated to G_n. After exploring all the images in group, we use the last output of CARU G_N as the group-wise representation. Unlike the conventional recurrent units (GRU or LSTM), our CARU can use co-attention to recurrently learn robust group representation for co-segmentation referring meanwhile reducing the noise data. We will provide justification on this issue in the later experiments section.

3.4. Co-Segmentation with Fused Representation

As described previously, the group feature is then broadcasted to each individual image, which allows the network to leverage the synergetic information and unique properties between the images. So the interaction of group and single representation are sufficiently exploited to facilitate the robust co-segmentation reasoning. Regarding visual features, coarse-resolution features from high layers of neural network emphasize the abstraction of visual content and contain the context with large receptive field to the summary of object, while fine-resolution features from low layers emphasize the appearance details and are more conducive to the location of objects. In our work, besides the single and group semantic features, we also utilize the low layer features ($Pool3, Pool4$) in the backbone network as the complementary and fuse the visual features from multiple layers to provide a comprehensive representation for object co-segmentation. Specifically, we concatenate the group-wise representation G_N with the single image representation S_n as well as the fine-resolution features of each image, and a 1×1 convolutional layer is used to reduce the dimensionality. The resultant feature , called $\{V_n^3, V_n^4, V_n^5\}_{n=1}^N$, are jointly used for co-segmentation as shown in Figure 3. Starting from V_n^5 ,the coarser-resolution feature map is upsampled by a factor of 2 using a deconvolutional layer. The upsampled map is then merged with the finer-resolution one by element-wise addition. This process iterates until the finest map $\hat{V}_n$ is obtained. To alleviate the aliasing effect of upsampling, we add a 3×3 convolutional layer after each merging operation. So the co-segmentation results M_n can be obtained with the fused features by applying 3×3 convolutional layers and deconvolutional layers followed with a sigmoid activation function. It is formulated as:

$$M_n = f_C(\hat{V}_n; \Theta_C), \tag{11}$$

where Θ_C are the parameters learned from the fusion and convolutional process f_C.

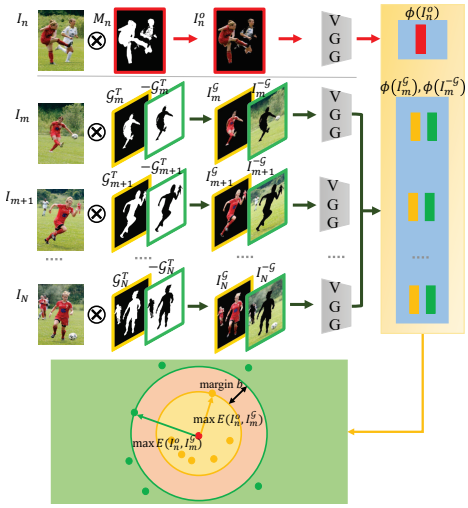

Figure 4. Illustration of the group-wise training objective.

4. Loss Function

Let $\mathcal{I} = \{I_n\}_{n=1}^N$ and their groundtruth $\{\mathcal{G}_n^T\}_{n=1}^N$ denote a collection of training samples where N is the number of images. After object co-segmenting, the co-segmentation results are $\{M_n\}_{n=1}^N$. We use the cross-entropy loss as the individual supervision for each image I_n:

$$L_s(I_n; \Theta) = -\big(\mathcal{G}_n^T log(M_n) + (1 - \mathcal{G}_n^T) log(1 - M_n) \big). \tag{12}$$

In addition to the cross-entropy losses, we propose a group-wise training objective to further explore the interactive relationships of whole images in the training group. Two criteria are jointly considered in the design of group-wise training objective, including 1) high cross-image similarity between the co-occurring objects and 2) high distinctness between the detected co-occurring objects and the rest of the images like background and non-common objects. We apply triplet loss as the group-wise constraint. Specifically, for a image I_n, we can generate three masked images with its co-segmentation mask M_n and groundtruth $\mathcal{G}_n^T$

$$I_n^o = M_n \otimes I_n,\ I_n^{\mathcal{G}} = \mathcal{G}_n^T \otimes I_n \ and \ I_n^{-\mathcal{G}} = (\mathcal{G}_n^{-T}) \otimes I_n, \tag{13}$$

where $\otimes$ denotes element-wise multiplication and $\mathcal{G}_n^{-T} = 1 - \mathcal{G}_n^T$. The masked image I_n^o means our current detected co-segmentation objects of I_n while image $I_n^{\mathcal{G}}$ and $I_n^{-\mathcal{G}}$ mean the real co-segmentation objects and non-common segmentation regions of I_n. Then we apply the extractor ϕ to all masked images $\{I_n^o, I_n^{\mathcal{G}}, I_n^{-\mathcal{G}}\}_{n=1}^N$ and obtain their

corresponding features $\{\phi(I_n^o), \phi(I_n^{\mathcal{G}}), \phi(I_n^{-\mathcal{G}})\}_{n=1}^N$. We apply triplet loss L_c on each I_n^o as the group-wise training objective as shown in Figure 4, formulated as:

$$L_c(I_n, I_{m\neq n}; \Theta) = \frac{1}{N-1} \sum_{m\neq n} \big[b + max \;\; E(I_n^o, I_m^{\mathcal{G}}) - min \;\; E(I_n^o, I_m^{-\mathcal{G}})\big]_+, \tag{14}$$

where b is the margin and $E(\cdot,\cdot)$ denotes the Euclidean distance between two feature vectors. The group-wise training objective uses the hinge function $[b + \bullet]_+$ to forces co-segmentation result M_n to be more similar to real co-segmentation objects than non-common regions. In co-segmentation task, it can be beneficial to pull together co-occurring objects as much as possible. For this purpose, it is possible to replace the hinge function by a smooth approximation using the softplus function: $\ln\big(1 + exp(\bullet)\big)$. The softplus function has similar behavior to the hinge, but it decays exponentially instead of having a hard cut-off, we hence refer to it as the soft-margin formulation. Note that all parts of our network are trained jointly, and the overall loss function is given as:

$$L = \frac{1}{N} \sum_{n=1}^{N} \big(\, L_s(I_n; \Theta) + \lambda \cdot L_c(I_n, I_{m\neq n}; \Theta) \, \big), \tag{15}$$

where λ is the tradeoff parameter and $\Theta = \{\Theta_S, \Theta_G, \Theta_C\}$ is the all learnable parameters set of our network.

5. Experiments

5.1. Experimental Setup

Datasets. We evaluate the proposed method and compare it with existing methods on three benchmarks for object co-segmentation, including the Internet dataset [22], the iCoseg dataset [1], and the PASCAL-VOC dataset [7]. These datasets are composed of real-world images with large intra-class variations, occlusions and background clutters. The Internet dataset contains images of three object categories including airplane, car and horse. Thousands of images in this dataset were collected from the Internet. Following the same setting of the previous work [12, 22, 25] , we use the same subset of the Internet dataset where 100 images per class are available. iCoseg consists of 38 groups of total 643 images which are challenging for object co-segmentation task because of the large variations of viewpoints and multiple co-occurring object instances. The PASCAL-VOC dataset contains total 1,037 images of 20 object classes from PASCAL-VOC 2010 dataset. The PASCAL-VOC dataset is more challenging and difficult than the Internet dataset due to extremely large intra-class variations and subtle figure-ground discrimination.

Implementation Details. We select the widely used pre-trained VGG19 net [23] (over the MS COCO dataset [17]) as the backbone network to extract the semantic features X_n and the fine-resolution features for each image. The deconvolutional layers are initialized with simple bilinear interpolation parameters. All images and groundtruth maps are resized to 224×224. The proposed models are optimized by standard SGD in which the momentum parameter is chosen as 0.99, the learning rate is set to 1e-5, and the weight decay is 0.0005. We need about 5000 training iterations for convergence. And for group-wise constraint, we use the activated layers $Relu3\_1$, $Relu4\_1$, $Relu5\_1$ of VGG as the feature vectors to calculate the Euclidean distance. Since the direct use of group-wise constraints may lead the model into the local optimal, we activate the triplet loss after 100 training iterations with cross-entropy loss only. And the loss tradeoff parameter λ is set to be 0.1 in our work. Training a deep neural network requires a lot of data. However, existing co-segmentation datasets are either are too small or have a limited number of object classes. Inspired by [7, 15, 30] we generated our training data from existing image dataset (COCO dataset [17]). Our training dataset contains 9k images belonging to 118 groups. The final binary co-segmentation masks are obtained by the self-adaptive threshold T [35] .

Evaluation metrics. Two widely used measures, *precision* ($\mathcal{P}$) and *Jaccard index* ($\mathcal{J}$), are adapted to evaluate the performance of object co-segmentation. Precision measures the percentage of correctly segmented pixels including both object and background pixels. Jaccard index is the ratio of the intersection area of the detected objects and the ground truth to their union area. The background pixels are taken into account in precision, so the images with larger background areas tend to have a higher performance in precision. Therefore, precision may not very faithfully reflect the quality of object co-segmentation results. Compared with precision, Jaccard index is considered more reliable to measure the quality of results. It provides a more appropriate evaluation as it only focuses on objects.

Method	Airplane		Car		Horse		Avg.	
	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$
Rubinstein13 [22]	88.0	0.56	85.4	0.64	82.8	0.52	82.73	0.427
Hati16 [9]	77.7	0.33	62.1	0.43	73.8	0.20	71.20	0.320
Jerripothula16 [13]	90.5	0.61	88.0	0.71	88.3	0.61	88.93	0.643
Quan16 [19]	91.0	0.56	88.5	0.67	89.3	0.58	89.60	0.603
Tao17 [25]	79.8	0.43	84.8	0.66	85.7	0.55	83.43	0.547
Yuan17 [33]	92.6	0.66	90.4	0.72	90.2	0.65	91.07	0.677
Han18 [8]	92.3	0.60	88.7	0.68	89.3	0.58	90.1	0.620
Ren18 [20]	88.3	0.48	83.5	0.62	83.2	0.49	85.0	0.530
Hsu18 [12]	94.2	0.67	93.0	0.82	89.7	0.61	92.29	0.698
Chen18 [3]	-	0.71	-	0.80	-	0.71	-	0.740
Li18 [15]	94.6	0.64	94.0	0.83	91.4	0.65	93.3	0.707
Ours	97.5	0.83	97.8	0.93	96.1	0.76	97.1	0.840

Table 1. The performance of object co-segmentation on the Internet dataset. The numbers in red and green respectively indicate the best and the second best results.

5.2. Comparison to the State-of-the-Arts

We compare the proposed method with the state-of-the-art methods on the Internet, iCoseg, and PASCAL-VOC

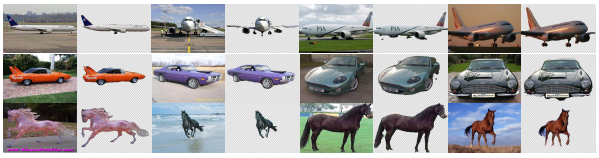

Figure 5. The co-segmentation results generated by our approach on the Internet dataset. In the three examples (rows), the common object categories are airplane, car, and horse, respectively.

datasets, and report their performances in Table 1, Table 2, and Table 3, respectively. The compared methods include the conventional methods and the most recent deep learning based methods. For fair comparison, we use either the implementations with recommended parameter settings or the co-segmentation results provided by the authors. Our method takes whole group with random images order as the input and achieves the state-of-the-art performance on the three datasets under two metrics. Specifically, on the Internet dataset, our method outperforms the second best results [3, 15] a large margin 13.5% in $\mathcal{J}$ and a margin 4% in $\mathcal{P}$. On the iCoseg dataset, our method improves upon the second best results [3, 12] by margins around 2% and 1.5% in terms of $\mathcal{J}$ and $\mathcal{P}$ respectively in Table 2. Note that for our network over 70% categories in iCoseg are unseen, the results indicate that our method can adapt itself well to unseen images with large variations. The same conclusion can be obtained from the co-segmentation examples on iCoseg in Figure 6. In Table 3, although the PASCAL-VOC dataset has higher variations than the Internet and iCoseg datasets, our proposed method also outperforms the best competing results [12, 31] by 5% and 3.4% in terms of $\mathcal{J}$ and $\mathcal{P}$ respectively. Some visual co-segmentation examples of our methods on three datasets are shown in Figure 5, Figure 6 and Figure 7. As can be seen, our method can generate promising object segments under different types of intra-class variations, such as colors, sharps, views, scales and background clutters. Even on the PASCAL-VOC dataset which contains images with higher intra-class variations and subtle figure-ground differences than the other two datasets, our method can infer the common object segments of high quality. For example, the sofa in the third row are of dissimilar colors and have clutter backgrounds. The effectiveness of our method mainly results from two properties: 1) The robust feature representations which not only express the images complementary individual properties, but also reflect the interaction among group images. 2) The effective training supervision which makes full use of the group-wise interactive relationships in the training group.

5.3. Ablation Studies

In this section, we conduct evaluation on PASCAL-VOC dataset to investigate the effectiveness of various components of the proposed model. This dataset has relative larger

Method	Avg. $\mathcal{P}$	Avg. $\mathcal{J}$	Method	Avg. $\mathcal{P}$	Avg. $\mathcal{J}$
Jerripothula16 [13]	91.8	0.72	Ren18 [20]	-	0.73
Quan16 [19]	93.3	0.76	Han18 [8]	94.4	0.78
Tao17 [25]	90.8	0.74	Chen18 [3]	-	0.87
Wang17 [31]	93.8	0.77	Hsu18 [12]	96.5	0.77
Yuan17 [33]	94.4	0.82	Tsai19 [26]	90.8	0.72
Li18 [15]	-	0.84	Ours	97.9	0.89

Table 2. The performance of object co-segmentation on the iCoseg dataset. The numbers in red and green respectively indicate the best and the second best results.

Figure 6. The co-segment results generated by our approach on the iCoseg dataset. From the first row to the last row, the classes are Taj Mahal, Pyramids, Cheetah, Woman Soccer, and kendo respectively.

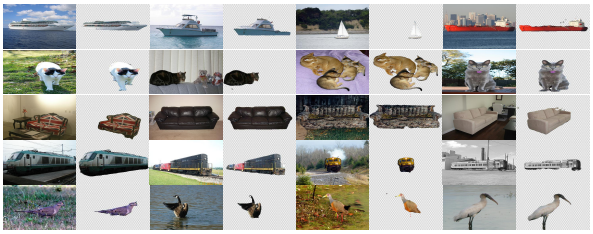

Figure 7. The co-segment results generated by our approach on the PASCAL-VOC dataset. From the first row to the last row, the classes are boat, cat, sofa, train, and bird respectively.

operable group size and is more challenging with various common objects poses and sizes, greater appearance variations and complex backgrounds. The results are shown in Table 4. We set the baseline approach by only using the single feature learning branch and training it with the cross-entropy loss L_s alone.

As can be seen, the baseline approach can't well handle the co-segmentation task. After applying the CARU, with the group representation, the performance of the baseline approach is improved by 5.8% and 3.8% in terms of $\mathcal{J}$ and $\mathcal{P}$. This shows that the group representation is essentially crucial for object co-segmentation. When we replace our CARU with the conventional recurrent units GRU, the performance drops a lot. That means our CARU is more effective to capture the synergetic information in a group compared with conventional recurrent units. However, the GRU still improves the performance of baseline, which means the recurrent architecture is naturally suitable for co-saliency detection task. We then add the group-wise training objective L_c to form the completed version of our approach. The

Method	Avg.$\mathcal{P}$	Avg.$\mathcal{J}$	A.P	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	D.T	Dog	Horse	M.B	P.S	P.P	Sheep	Sofa	Train	TV
Faktor13 [7]	84.0	0.46	0.65	0.14	0.49	0.47	0.44	0.61	0.55	0.49	0.20	0.59	0.22	0.39	0.52	0.51	0.31	0.27	0.51	0.32	0.55	0.35
Hati16 [9]	72.5	0.25	0.44	0.13	0.26	0.31	0.28	0.33	0.26	0.29	0.14	0.24	0.11	0.27	0.23	0.22	0.18	0.17	0.33	0.27	0.41	0.18
Quan16 [19]	89.0	0.52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Jerripothula16 [13]	85.2	0.45	0.64	0.20	0.54	0.48	0.42	0.64	0.55	0.57	0.21	0.61	0.19	0.49	0.57	0.50	0.34	0.28	0.53	0.39	0.56	0.38
Wang17 [31]	84.3	0.52	0.75	0.26	0.53	0.59	0.51	0.70	0.59	0.70	0.35	0.63	0.26	0.56	0.63	0.59	0.35	0.28	0.67	0.52	0.52	0.48
Han18 [8]	90.1	0.53	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Hsu18 [12]	91.0	0.60	0.77	0.27	0.70	0.61	0.58	0.79	0.76	0.79	0.29	0.75	0.25	0.63	0.66	0.65	0.37	0.42	0.75	0.67	0.68	0.51
Ours	94.1	0.63	0.78	0.29	0.71	0.66	0.58	0.82	0.79	0.81	0.35	0.78	0.26	0.65	0.78	0.69	0.39	0.45	0.77	0.70	0.73	0.55

Table 3. The performance of object co-segmentation on the PASCAL-VOC dataset under Jaccard index and Precision. The class-wise results are measured in Jaccard index. The numbers in red and green respectively indicate the best and the second best results.

triplet loss helps to further improve the performance comparing with CARU version by 3.8% and 2.2% in terms of $\mathcal{J}$ and $\mathcal{P}$. This indicates that the group-wise training supervision can better model the synergetic relationships between the common objects and help the network to learn more effective group representation, which in turn boost the co-segmentation task. We then replace the fused features $\hat{V}_n$ with the single-scale feature V_n^5 to investigate the effectiveness of multi-scale visual features. Consequently, the performance drops by 3% and 1.4% in terms of $\mathcal{J}$ and $\mathcal{P}$ when only using the single-scale feature. This result demonstrates that fusing visual features at multi-scales produces a comprehensive representation characterizing both visual abstraction and details of foregrounds and is useful for co-segmentation.

We evaluated the effects of CARU with different component settings. As shown, for update gate g_z, the combination of Z_s and Z_c achieves better performances than using them alone. When we remove the reset gate g_d from CARU, the performance declines on two metrics. This indicates the reset gate g_d is able to suppress the noise information in the group. In order to verify the adaptability of our recurrent architecture to different group sizes, we construct new testing groups by randomly selecting a sub-group with different size 5, 10 and 15 from the original groups. As reported in Table 4, our approach achieves good performance on different size groups and still consistently outperforms all the state-of-the-art methods. And the performance raises along with the group size, which emphasizes the importance of the group information completeness to robust object co-segmentation. To further justify the denoising ability of our CARU, we add a image which is randomly selected from COCO dataset to the testing groups in size 5, 10 and 15 for simulating the noise data in real-world applications. As shown in results, although the noise data damages our performance a little, we still outperform all the state-of-the-art methods, which demonstrate the robustness of the proposed method. As for the pair-wise methods, the noise data can badly damage their performance. To estimate the influence of the input order of images within a group on our network, we randomly generate three different input orders of the dataset for testing. There are minor variations between three results, which shows our method is not sensitive to the input orders.

Method	Avg.$\mathcal{P}$	Avg.$\mathcal{J}$	Method	Avg.$\mathcal{P}$	Avg.$\mathcal{J}$
Baseline	88.7	0.570	Ours(5)	93.5	0.617
Baseline+CARU	92.1	0.603	Ours(10)	93.9	0.621
Baseline+GRU	89.9	0.589	Ours(15)	94.0	0.623
Baseline+CARU+L_c	94.1	0.626	Ours(5+noise)	93.1	0.612
Baseline+CARU+L_c−MFF	92.8	0.607	Ours(10+noise)	93.7	0.620
Baseline+L_c+CARU(s)	93.8	0.621	Ours(15+noise)	93.9	0.623
Baseline+L_c+CARU(c)	93.6	0.620	Ours(order1)	94.1	0.625
Baseline+L_c+CARU($-gd$)	93.1	0.615	Ours(order2)	94.0	0.623
			Ours(order3)	94.1	0.622

Table 4. Ablation study of our method on PASCAL-VOC.

6. Conclusion

In this paper, we propose a novel end-to-end deep learning approach for group-wise object co-segmentation with a recurrent network architecture. The proposed approach explores single image properties and robust group representation simultaneously, which are essentially crucial for real-world co-segmentation. Specifically, the semantic features extracted from a pre-trained CNN of each image are first processed by single image representation branch to learn the unique properties. Then, by using the spatial and channel co-attention between images, the special designed CARU recurrently explores all images in the group to learn the robust group representation and meanwhile suppresses noisy information. The group feature is broadcasted to each individual image and fused with multi-scale fine-resolution features to facilitate the inferring of co-segmentation. Moreover, we propose a group-wise training objective to utilize the co-object similarity and figure-ground distinctness as the additional supervision. The whole modules are collaboratively optimized in an end-to-end manner, further improving the robustness of the approach. Extensive experimental results demonstrate the superiority of our approach. To the best of our knowledge, this is the first attempt to address group-wise object co-segmentation task with the recurrent architecture.

Acknowledgments

This work was supported by National High Technology Re-search and Development Program of China (No. 2007AA01Z334), National Natural Science Foundation of China (Nos. 61321491 and 61272219), Program for New Century Excellent Talents in University of China (NCET-04-04605), The China Postdoctoral Science Foundation (Grant No. 2017M621700)and Innovation Fund of State Key Laboratory for Novel Software Technology (Nos. ZZKT2013A12, ZZKT2016A11 and ZZKT2018A09).

References

[1] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, pages 3169–3176, 2010. 6

[2] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 2129–2136, 2011. 3

[3] Hong Chen, Yifei Huang, and Hideki Nakayama. Semantic aware attention based deep object co-segmentation. In *Asian Conference on Computer Vision (ACCV)*, 2018. 2, 3, 6, 7

[4] Yun-Chun Chen, Po-Hsiang Huang, Li-Yu Yu, Jia-Bin Huang, Ming-Hsuan Yang, and Yen-Yu Lin. Deep semantic matching with foreground detection and cycle-consistency. In *Asian Conference on Computer Vision (ACCV)*, 2018. 1

[5] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111, 2014. 2

[6] Jifeng Dai, Ying Nian Wu, Jie Zhou, and Song-Chun Zhu. Cosegmentation and cosketch by unsupervised learning. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1305–1312, 2013. 1, 2

[7] Alon Faktor and Michal Irani. Co-segmentation by composition. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1297–1304, 2013. 6, 8

[8] Junwei Han, Rong Quan, Dingwen Zhang, and Feiping Nie. Robust object co-segmentation using background prior. *IEEE Trans. Image Processing*, 27(4):1639–1651, 2018. 2, 6, 7, 8

[9] Avik Hati, Subhasis Chaudhuri, and Rajbabu Velmurugan. Image co-segmentation using maximum common subgraph matching and region co-growing. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 736–752, 2016. 1, 6, 8

[10] Dorit S. Hochbaum and Vikas Singh. An efficient algorithm for co-segmentation. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 269–276, 2009. 2

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[12] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 748–756, 2018. 1, 2, 6, 7, 8

[13] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Image co-segmentation via saliency co-fusion. *IEEE Trans. Multimedia*, 18(9):1896–1909, 2016. 2, 6, 7, 8

[14] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 725–739, 2014. 2

[15] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. In *Asian Conference on Computer Vision (ACCV)*, 2018. 2, 3, 6, 7

[16] Yong Li, Jing Liu, Zechao Li, Hanqing Lu, and Songde Ma. Object co-segmentation via salient and common regions discovery. *Neurocomputing*, 172:225–234, 2016. 3

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755, 2014. 6

[18] Armin Mustafa and Adrian Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5583–5592, 2017. 1

[19] Rong Quan, Junwei Han, Dingwen Zhang, and Feiping Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 687–695, 2016. 1, 2, 6, 7, 8

[20] Yan Ren, Licheng Jiao, Shuyuan Yang, and Shuang Wang. Mutual learning between saliency and similarity: Image cosegmentation via tree structured sparsity and tree graph matching. *IEEE Trans. Image Processing*, 27(9):4690–4704, 2018. 1, 6, 7

[21] Carsten Rother, Thomas P. Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 993–1000, 2006. 1, 2

[22] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 1939–1946, 2013. 1, 2, 6

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4, 6

[24] Jian Sun and Jean Ponce. Learning dictionary of discriminative part detectors for image categorization and cosegmentation. *International Journal of Computer Vision*, 120(2):111–133, 2016. 1, 2

[25] Zhiqiang Tao, Hongfu Liu, Huazhu Fu, and Yun Fu. Image cosegmentation via saliency-guided constrained clustering with cosine similarity. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4285–4291, 2017. 6, 7

[26] Chung-Chi Tsai, Weizhi Li, Kuang-Jui Hsu, Xiaoning Qian, and Yen-Yu Lin. Image co-saliency detection and co-segmentation via progressive joint optimization. *IEEE Trans. Image Processing*, 28(1):56–71, 2019. 3, 7

[27] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 2

[28] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother. Cosegmentation revisited: Models and optimization. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II*, pages 465–479, 2010. 1, 2

[29] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 2217–2224, 2011. 2

[30] Chong Wang, Zheng-Jun Zha, Dong Liu, and Hongtao Xie. Robust deep co-saliency detection with group semantic. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*, 2019. 6

[31] Chuan Wang, Hua Zhang, Liang Yang, Xiaochun Cao, and Hongkai Xiong. Multiple semantic matching on augmented n-partite graph for object co-segmentation. *IEEE Trans. Image Processing*, 26(12):5825–5839, 2017. 7, 8

[32] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2314–2320, 2017. 1

[33] Ze-Huan Yuan, Tong Lu, and Yirui Wu. Deep-dense conditional random fields for object co-segmentation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3371–3377, 2017. 1, 2, 6, 7

[34] Dingwen Zhang, Huazhu Fu, Junwei Han, Ali Borji, and Xuelong Li. A review of co-saliency detection algorithms: Fundamentals, applications, and challenges. *ACM TIST*, 9(4):38:1–38:31, 2018. 3

[35] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *International Journal of Computer Vision*, 120(2):215–232, 2016. 6

[36] Dingwen Zhang, Junwei Han, Yang Yang, and Dong Huang. Learning category-specific 3d shape models from weakly labeled 2d images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3587–3595, 2017. 1