Neurocomputing 358 (2019) 166–176

Contents lists available at ScienceDirect

Neurocomputing

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

An easy-to-hard learning strategy for within-image co-saliency detection

Shaoyue Song[a,*], Hongkai Yu[b,*], Zhenjiang Miao[a,*], Dazhou Guo[c], Wei Ke[d], Cong Ma[a], Song Wang[c,e]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[b] Department of Computer Science, University of Texas-Rio Grande Valley, Edinburg, TX 78539, USA
[c] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
[d] Human Sensing Laboratory, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[e] School of Computer Science and Technology, Tianjin University, Tianjin 300350, China

ARTICLE INFO

Article history:
Received 20 November 2018
Revised 2 April 2019
Accepted 5 May 2019
Available online 10 May 2019

Communicated by Prof. Junwei Han

Keywords:
Within-image co-saliency
Easy-to-hard learning
Multiple instance learning

ABSTRACT

Within-image co-saliency detection is to detect/highlight the common saliency (similar-appearance salient objects) in a single image. Ideally, it can be solved by detecting each individual salient object first and then comparing them, which is possible for some images with simple representations. However, in practice, this way is not accurate and robust for some images with complex representations. In this paper, we propose an easy-to-hard learning strategy to solve this problem. By directly localizing and comparing salient objects in simple images, superpixel confidences as co-salient objects are inferred by an easy learning method, which provide promising but also noisy supervisions for complex images. Therefore, within-image co-saliency detection in complex images can be modeled as a hard learning problem with noisy labels. A multi-scale Multiple Instance Learning (MIL) model together with a new sampling method is proposed to solve this hard learning problem with noisy labels. Experimental results show that the proposed method achieves the best performance on a public benchmark dataset and two synthetic datasets.

1. Introduction

In recent years, saliency detections have demonstrated wide applications in the computer vision community, such as object segmentation [1,2], video analysis [3], person re-identification [4], object tracking [5], image classification [6], etc. As introduced in [7], saliency detection can be divided into three types: detections for within-image saliency, cross-image co-saliency, and within-image co-saliency. Most existing methods are developed for detecting all the salient objects in a single image [1,5,8–10] or the common salient objects across multiple images [11–16].

In this paper, we focus on the problem of within-image co-saliency detection, which highlights the similar-appearance salient objects in a single image. Within-image co-saliency detection has many potential applications in computer vision, for example, it could help to detect repeated instances of a salient object class to synthesize the realistic animation from a still picture [17], estimate the number of instances of the same object class [18,19], reduce the redundancy of the images [7], etc.

Within-image co-saliency detection is not uncomplicated because (1) the classes of co-salient objects are unknown, (2) the number of co-salient objects is unconstrained. Therefore, object detectors or instance-level object segmentation by supervised learning like Faster RCNN [20] and Mask RCNN [21] cannot accurately localize the salient objects since the salient objects might not match to any labeled objects in the training set. Recently, some methods [22,23] could be used to localize instance-level salient object for unconstrained salient object detection without the requirement of knowing object classes, however its localization result might be not accurate and robust enough sometimes, i.e., with missing detections, no detections, inaccurate detections in some images. Fig. 1 displays the success and failure cases of instance-level salient object detection by Zhang et al. [22] to detect each salient object in single images. We can see that it is difficult to accurately localize each salient object and compare one with each other to solve the problem of within-image co-saliency detection. However, without the need of knowing object classes, the instance-level salient object detection described in [22] provides meaningful information in some images with simple representations. By a

* Co-corresponding author.
E-mail addresses: 14112060@bjtu.edu.cn (S. Song), hongkai.yu@utrgv.edu (H. Yu), zjmiao@bjtu.edu.cn (Z. Miao).

https://doi.org/10.1016/j.neucom.2019.05.009
0925-2312/

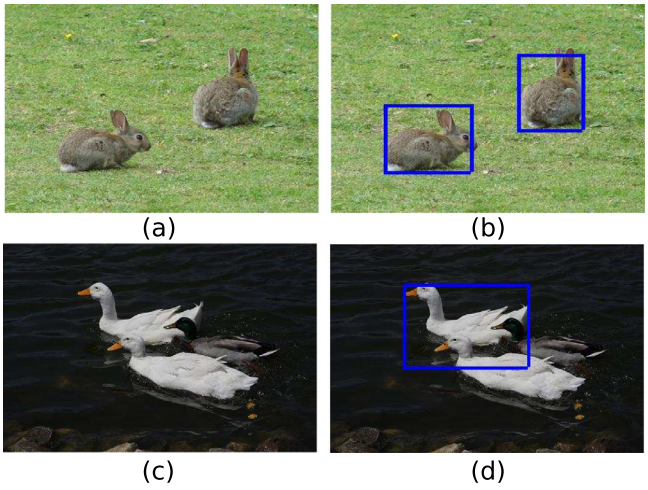

Fig. 1. Sample results of the instance-level salient object detection [22]. (a) An image with simple representation. (b) Success detection. (c) An image with complex representation. (d) Missing or inaccurate detections. The outputs of [22] are shown in blue bounding boxes. It demonstrates that directly detecting and comparing salient objects cannot solve the problem of within-image co-saliency detection.

proposed easy learning method in simple images, we could infer each pixel's confidence of its within-image co-saliency. Because superpixel provides superior local feature consistency and acceleration for computing, like the methods in [14,24], we use superpixel to replace the original pixel and treat each superpixel as an instance for computing in this paper.

The superpixel confidences inferred from the results of instance-level salient object detection on simple images are reasonable but still have some noises due to the inaccurate instance-level salient object localization. Following an easy-to-hard learning procedure like humans, we model detection on complex images as a hard learning problem with noisy labels. A multi-scale Multiple Instance Learning (MIL) model for data fusion of existing saliency detection methods is proposed to solve this hard learning problem with noisy labels. In our model, two bags are selected from each image (one positive and one negative bag). The positive bag denotes the superpixels/instances as the co-salient regions in that image. The negative bag indicates the superpixels/instances as regions without co-saliency in that image. As a weakly supervised learning model, MIL only requires the bag labels and does not need the instance labels for training. MIL requires that positive bag has at least one positive instance, which can be easily satisfied in our problem. MIL assumes that a negative bag only has negative instances, so we use a new sampling method to construct the negative bags by selecting the superpixels with maximum feature distance in potential regions without co-saliency to the potential co-salient regions. Superpixels in different scales provide rich information for the different descriptions of the image structure, so we train multiple MIL models in different scales independently and fuse the results together to achieve the final co-saliency detection in a single image.

The proposed method only applies the pretrained Convolutional Neural Networks (CNN) model to initialize the instance-level salient object detection, and then does not need any human annotations, so it is a weakly supervised method. In summary, this paper has three major contributions: (1) imperfect instance-level salient object detection is integrated into an easy-to-hard learning strategy for within-image co-saliency detection. (2) MIL together with a new sampling method is proposed to solve the learning problem with noisy labels. (3) A multi-scale MIL based data fusion method is proposed for hard learning in complex images.

Experiments on the benchmark dataset [7] for simple images, complex images and all images show that the proposed method achieves the state-of-the-art performance in terms of most evaluation metrics. Furthermore, the performance of high Precision is more impressive, which demonstrates that the proposed method is capable of detecting more accurate co-salient objects in a single image. We also synthesize two datasets to compare the proposed method with other comparison methods.

2. Related work

Saliency detection can be divided into three types: within-image saliency detection [1,10,25,26], cross-image co-saliency detection [11,15,16,27,28] and within-image co-saliency detection [7]. Most of the existing methods are designed for solving the first two problems. For example, Zhang et al. [26] proposed a symmetrical deep neural network architecture to detect the salient object in an end-to-end way, Han et al. [10] leveraged both of the depth view and RGB view of the image by a designed CNN model to help the salient object detection for the RGB image, Han et al. [15] introduced a unified metric learning-based framework which combines discriminative feature learning and co-salient object detector training together to solve the cross-image co-salient object detection, Hsu et al. [16] proposed an unsupervised CNN-based method which decomposed the cross-image co-saliency into single-image saliency detection and cross-image co-occurrence region discovery parts.

The problem of within-image co-saliency detection is first proposed in [7], and a dataset with 364 images with the assumption that each image has and only has one class of co-salient objects for within-image co-saliency detection problem is publicized. Yu et al. [7] proposed an unsupervised bottom-up method to address the within-image co-saliency detection problem by fusing and optimizing a set of proposal groups showing good common saliency in the original image, and their method obtained the state-of-the-art performance in detecting within-image co-saliency.

Another work mostly related to our problem is the instance-level salient object detection or segmentation [22,23]. Different from the object detection/segmentation problems like [20,29,30] which can only detect/segment objects of known classes, instance-level salient object detection methods can detect or segment salient objects of unknown classes from unconstrained images, in which the number of salient objects is not given. Intuitively, the instance-level salient object detection can provide meaningful prior knowledge as the start point for us, and we may directly extend these methods to within-image co-saliency detection by detecting and comparing the salient objects. However, their results are imperfect with certain-level detection errors. Therefore, an easy-to-hard strategy is developed in this paper to solve the learning problem with noisy labels.

Learning or improving saliency detection from current noisy saliency maps by existing unsupervised or supervised saliency methods are recently introduced in [31–33]. Zhang et al. [31] designed "intra-image" and "inter-image" based reliability measures to iteratively improve the current noisy saliency maps by several unsupervised saliency detection methods. [32] updates a Fully Convolutional Networks (FCN) with an estimated noise model to improve the latent saliency prediction from FCN in each iteration. Li et al. [33] uses image class labels as a weak supervision to help pixel-level saliency prediction. These methods all use estimated or improved pseudo ground truth to iteratively train the deep CNN based models, however our method just uses MIL based learning method which is much more simple and efficient in training. In addition, our method divides within-image co-saliency detection into two steps: easy learning in simple images and hard learning in complex images. The proposed easy-to-hard learning strategy is

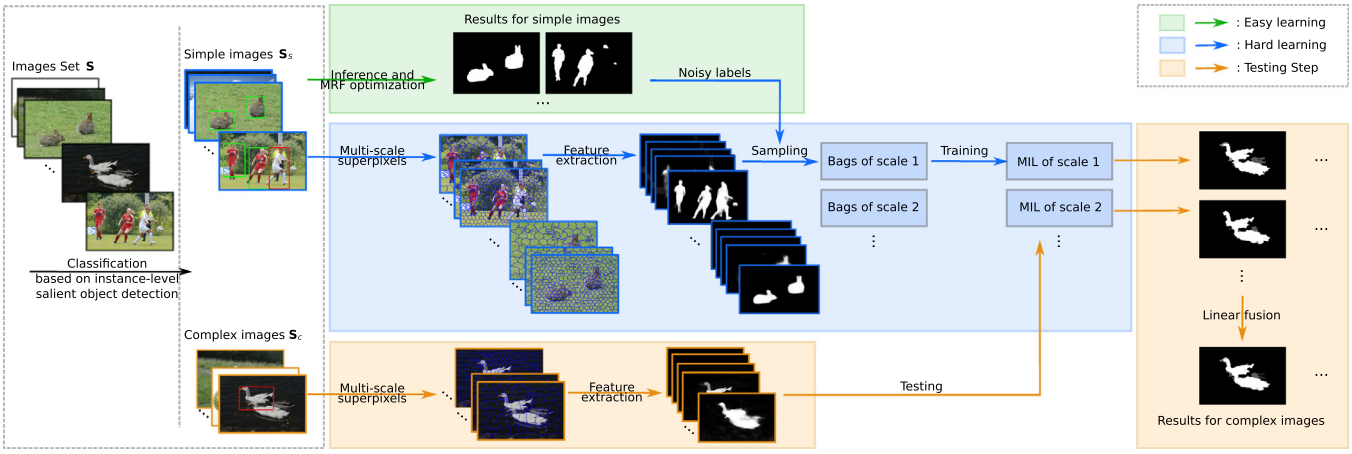

Fig. 2. An overview of the proposed easy-to-hard learning strategy. Imperfect instance-level salient object detection is integrated to provide weak supervision, and a multi-scale MIL model is used for learning with noisy labels. Superpixel is taken as the instance in the multi-scale MIL model.

more similar to the human learning, therefore our motivation and framework are totally different with them.

Self-paced learning (SPL) based frameworks such as [34–36] share the similar spirit with our easy-to-hard learning framework. However, these two solutions are quite different. SPL framework usually begins with simple examples and gradually takes complex examples into consideration in a self-paced fashion iteratively to learn the model. While in our easy-to-hard learning framework, we only obtain the pseudo labels for one time and a MIL algorithm based learning method is followed to alleviate the bad influence of the noisy labels when predicting the saliency maps of hard images.

Weakly supervised learning methods are widely used to solve the learning problem with incomplete labels. Among these methods, MIL models [24,37,38] are frequently used for within-image saliency detection and cross-image co-saliency detection. MIL model requires that negative bags only contain negative instances, but it is challenging to be satisfied in our problem. Therefore, we propose a MIL model together with a new sampling method to relieve this difficulty.

3. Methodology

In this section, we will first give an overall statement to the whole framework, and then introduce how to classify images into simple and complex ones by the imperfect instance-level salient object detection, followed by detailed elaborations for the easy learning in simple images and hard learning in complex images.

3.1. Problem statement

Fig. 2 shows an overview of the proposed easy-to-hard learning strategy. Given a set of images $\mathbf{S}$, our task is to detect the common saliency within each image of $\mathbf{S}$. Some simple images of $\mathbf{S}$, denoted as a subset $\mathbf{S}_s$, can be solved by directly applying the pretrained CNN model for instance-level salient object detection followed by a simple instance-level object comparison, while other complex images of $\mathbf{S}$, denoted as a subset $\mathbf{S}_c$, cannot be solved in such a naive way. In this paper, on $\mathbf{S}_s$, the naive way by comparing instance-level salient objects is called easy learning. Based on the results of easy learning on $\mathbf{S}_s$, we propose a multi-scale Multiple Instance Learning (MIL) based method to detect the co-salient objects on $\mathbf{S}_c$, which is named as hard learning.

3.2. Classification for simple and complex images

As shown in Fig. 1, directly applying the pretrained CNN model for instance-level salient object detection [22] can localize each salient object in a single image, but it might be imperfect, e.g., no detections, with missing or inaccurate detections in some images. As the assumption in the benchmark dataset [7] for within-image co-saliency detection, it assumes that each image has and only contains one class of co-salient objects, so we design the following method for image classification.

Let n denotes the number of detected salient objects (bounding boxes) by directly applying the pretrained CNN model for instance-level salient object detection [22] in the image I. Suppose the detected salient object (bounding box) is O_i, where $i = 1, \ldots, n$. If $n \leq 1$, the image I is classified as a complex image since current information is not enough to discover the co-salient objects in I. If $n \geq 2$, we calculate the L_2 distance (denoted as d_{ij}) of the 128×3 bin normalized RGB color histograms between each pair of detected salient objects (O_i and O_j), and the pair is matched if $d_{ij} < \lambda$, where λ is a predefined threshold. When $n \geq 2$, if no pairs of detected salient objects are matched, I is classified as a complex image, otherwise I is classified as a simple image. In a simple image, if one detected salient object O_k is not matched with any other detected salient objects, O_k is considered as a noisy salient object (without showing co-saliency) in a single image.

Fig. 3 shows some simple and complex examples determined by our feature comparing based classification algorithm. From Fig. 3, we can see that the classification result on the image (simple or complex) may different from human's feeling, fully depending on the feature comparing results.

3.3. Easy learning for simple images

Fig. 4 shows the main steps for the easy learning in simple images. With the instance-level salient object detection, we first search for the matched and unmatched salient objects. Meanwhile, we apply the method DCL [39] to detect the within-image saliency maps as M_1. If one detected salient object/bounding box O_k is a noisy salient object (unmatched), we set the saliency values in the area of O_k to zeros in M_1 as an initial within-image co-saliency map M_i; otherwise, we directly use M_1 as the initial map M_i. The reason to choose DCL for this step is that DCL using CNN to detect all the salient objects in the image has a high recall. Because the detected bounding box for the salient object is not accurate enough, removing the saliency values in O_k might have some

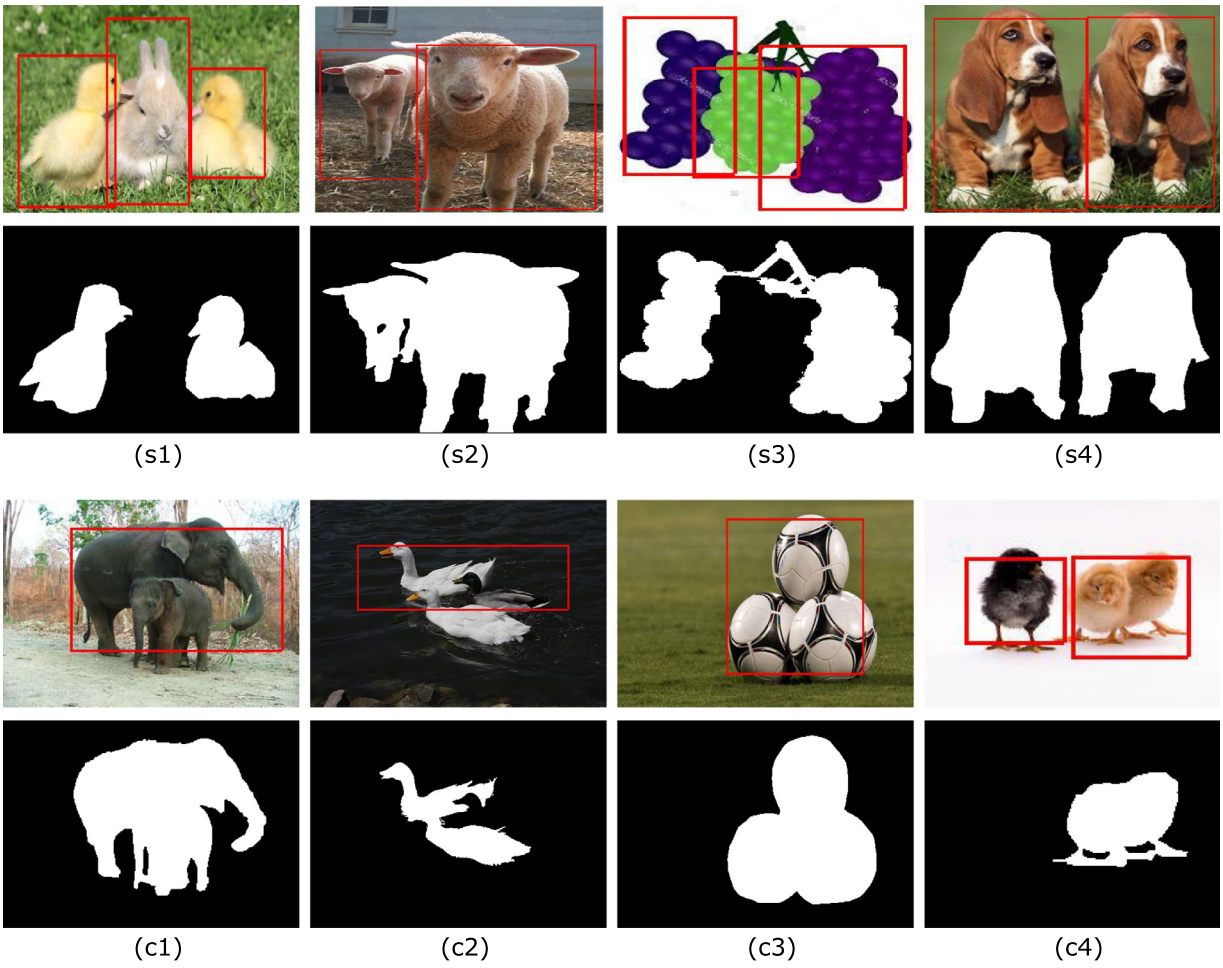

Fig. 3. The top two rows are examples of simple images and the bottom two rows are examples of complex images from the "WICOS" dataset [7]. The bounding boxes are the results of the instance-level salient object detection [22]. Ground truth is also displayed for reference. Simple images (s1–s4): two of the detected bounding boxes are matched for each image; complex images (c1–c3): only one detected bounding box is generated for each image; complex image (c4): two detected bounding boxes are generated but not matched. Complex images are the images that the common salient objects cannot be found by simply comparing the features of the detected bounding boxes.

errors. To reduce the removing errors, we refine the saliency map as an energy minimization problem in Markov Random Fields (MRF).

Let $X = \{x_i\}_{i=1}^{n}$ be the binary labels for pixel i, where $x_i = 1$ for co-salient regions and $x_i = 0$ for regions without co-saliency. Let θ denotes the appearance models including Gaussian Mixture Model (GMM) for co-salient regions and regions without co-saliency, respectively. The refinement is accomplished by minimizing the following energy function:

$$E_1 = \sum_i D(x_i, \theta) + \sum_{i,j \in \mathcal{L}} S(x_i, x_j) \tag{1}$$

where $\mathcal{L}$ defines a local neighborhood for each pixel (4 or 8 neighbor connectivity). $D(x_i, \theta)$ is the data term and $S(x_i, x_j)$ is the smooth term. These two terms are determined by

$$D(x_i, \theta) = -\mathbf{log}[P(z_i \mid \theta_f)x_i + P(z_i \mid \theta_b)(1 - x_i)] \tag{2}$$

$$S(x_i, x_j) = [x_i \neq x_j]\mathbf{exp}(-\beta||z_i - z_j||^2) \tag{3}$$

where $P(\cdot)$ is the Gaussian probability for pixel i's feature z_i given foreground and background GMM models θ, $[\phi]$ denotes the indicator function taking a value of 1 or 0 for a true or false predicate ϕ, and β is a scale parameter that is set to 1.0 in our experiments. The data term $D(x_i, \theta)$ measures the cost of labeling pixel i as co-salient regions or regions without co-saliency according to the appearance GMM models θ. The smooth term $S(x_i, x_j)$ enables the smoothness of the labels by penalizing the discontinuity among the local neighboring pixels with different labels. The optimization is initialized by thresholding the initial map M_i (0.2 as threshold in all experiments). Max-flow algorithm [40] is then iteratively applied to estimate X and θ simultaneously so as to minimize the energy function E_1 until convergence as that in [41]. Superpixel confidences (binary labels) as co-salient objects are then extracted.

3.4. Hard learning for complex images

The refined within-image co-saliency maps on simple images are promising but still contain some false positive and false negative errors. We treat the simple images as the training set, and view the complex images as the testing set. Therefore, this is a learning problem with incomplete labels. MIL as a widely used model for weakly supervised learning does not require the labels for each instance during training, so it is selected to finish our task. Considering the superior local feature consistency and computation efficiency of superpixels, each superpixel is thought as an instance in our model. Two bags are selected from each image: co-salient regions (positive bag) and regions without co-saliency (negative bag). Our goal is to score each superpixel for complex images by learning a reliable MIL model from simple images.

However, one difficulty in this problem is the commonly-used assumption by MIL for negative bags during training. MIL assumes

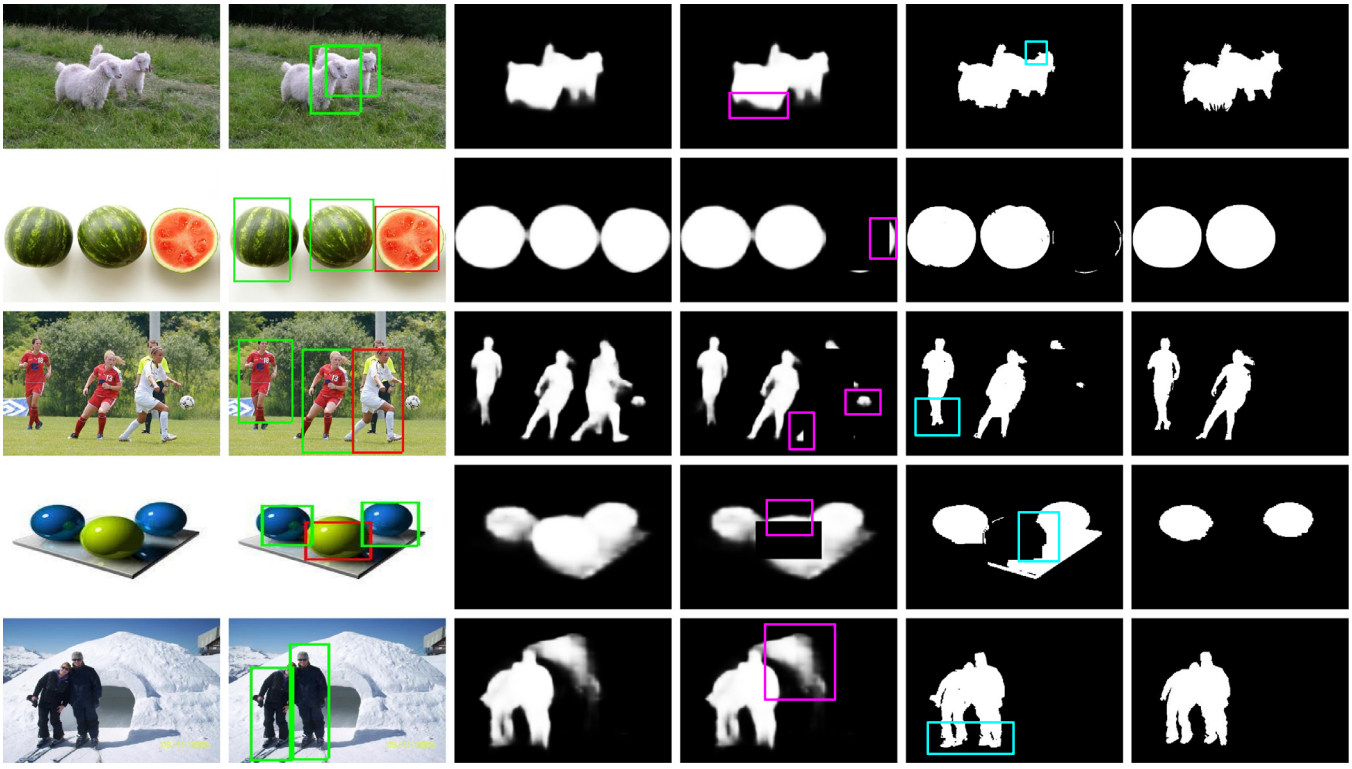

Fig. 4. Easy learning for simple images (four examples) on the "WICOS" dataset [7]. (a) Color images. (b) Instance-level salient object detection by Zhang et al. [22]. (c) Within-image saliency map M_1 by DCL [39]. (d) Initial within-image co-saliency map M_i. (e) Refined within-image co-saliency map M_r. (f) Ground truth. The matched salient objects/bounding boxes are shown in green and the unmatched ones are shown in red. The removing of MRF refinement step is highlighted in pink box and the compensating is highlighted in cyan box.

that the positive bag contains at least one positive instance and the negative bag only contains negative instances. MIL's assumption for positive bags is excellent for our problem since it solves our problem of incomplete positive bags. However, we cannot guarantee that the negative bags only contain negative instances in our problem. In order to relieve this challenge, we use *a sampling method considering feature distance*. Specifically, in each simple image, the image can be divided into potential positive regions (pixels as 1) and potential negative regions (pixels as 0) based on its M_r. In each simple image, we sample m superpixels in potential positive regions considering the sorted saliency values in M_1 into the positive bag, and we also sample m superpixels in potential negative regions having top m largest L_2 feature distance to the potential positive regions into the negative bag. The normalized RGB color histogram is used to calculate the above feature distance.

For each sampled superpixel/instance, designing its handcrafted feature is challenging, especially to express the within-image co-saliency. Some existing saliency detection methods CWS [11], RC [1], DCL [39], RFCN [25] and CDS [7] could provide reasonable prior knowledge for superpixels. The last method is for within-image co-saliency detection. The first four methods are for within-image saliency detection. Different methods have various advantages and disadvantages, so we expect that they could make up for each other to obtain a better detection. Therefore, we model our MIL learning problem as a data fusion problem. For each superpixel, we extract a 5D vector as its instance feature by the corresponding saliency values from five saliency methods (CWS, RC, DCL, RFCN, CDS). The training set is from simple images and has a set of bags $B_1, B_2, \ldots, B_N$, and the corresponding bag labels are $y_1, y_2, \ldots, y_N$, where y_i is 0 or 1 as a binary classification problem. Each bag B_i contains m instances $\mathbf{b_{ij}}$, where $j = 1, \ldots, m$. The instance $\mathbf{b_{ij}}$ is represented as a 5D feature vector. In order to assign the designed bag labels to the corresponding bags, the classical MIL training process is to minimize the following log-likelihood energy function [42,43]:

$$E_2 = -\sum_{i}^{N} y_i \mathbf{log}(P_i) + (1 - y_i)\mathbf{log}(1 - P_i) \tag{4}$$

where P_i is the bag probability assigned to its bag label y_i. The bag probability P_i is computed from the instance probability:

$$p_{ij} = (1 + e^{-(\mathbf{w^T b_{ij}})})^{-1}, \tag{5}$$

and a noisy-or combination rule [42,43]: $P_i = 1 - \prod_j (1 - p_{ij})$, where $\mathbf{w}$ is the weight vector during training the MIL model. Using the learned weight vector $\mathbf{w}$ on simple images, then we test the trained MIL model on complex images. During testing, we also extract the 5D feature vector for each superpixel as the instance feature, and then the probability (score) for each superpixel/instance could be computed by Eq. (5).

Superpixels in different scales provide rich information for different descriptions of the image structure, so we train multiple MIL models in different scales independently and then fuse their testing results linearly to achieve the final co-saliency detection in each complex image as shown in Fig. 2.

4. Experiments

In this section, we will introduce the benchmark dataset, experiment settings, evaluation metrics, comparison methods and results.

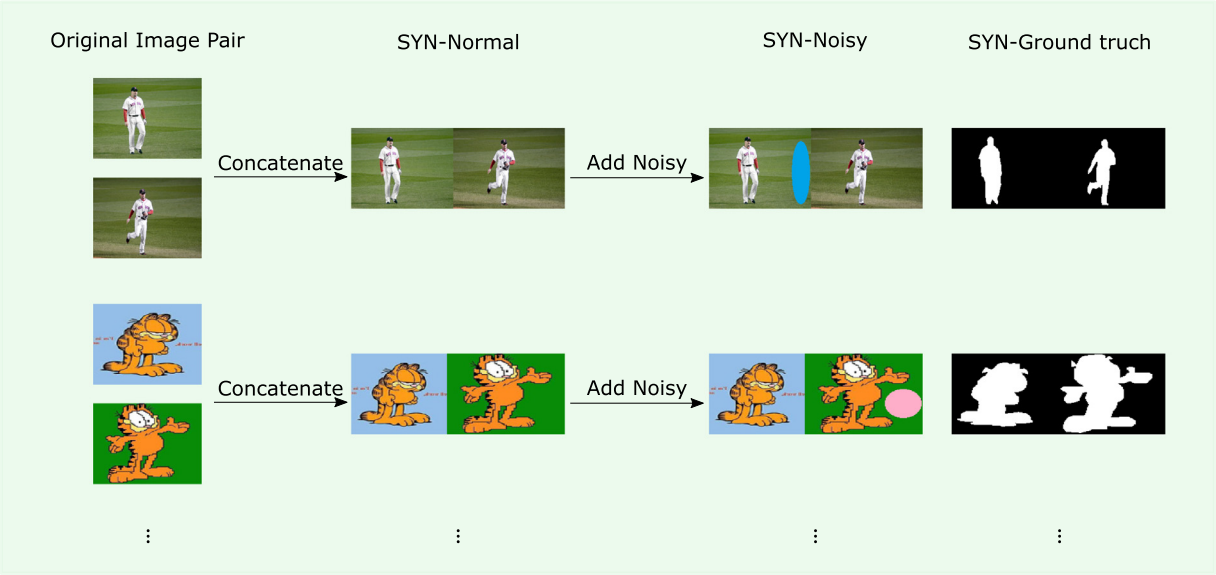

Fig. 5. Example images in the new synthetic datasets. Given two images containing similar object(s), we concatenate the two images into one single image, leading to the new synthetic dataset "SYN-Normal" (86 images). Adding ellipse-shaped noises in some random background locations generate another new single image, leading to the new synthetic dataset "SYN-Noisy" (86 images). The last column of images shows the corresponding ground truth annotation.

4.1. Dataset, setting and evaluation

Most of the existing widely-used image datasets related to saliency detection like MSRA [44], iCoseg [45], HKU-IS [46] are mainly designed for evaluating within-image saliency detection or cross-image co-saliency detection, which do not pay attention to within-image co-saliency in a single image, which are not suitable to evaluate our problem. The recent research [7] is the first work to discuss the problem of within-image co-saliency, and a new benchmark dataset is publicized to evaluate the within-image co-saliency detection together with their work. Therefore, we perform our experiments on this new benchmark dataset [7], and we define this benchmark dataset as "WICOS" in this paper. "WICOS" dataset contains 364 color images with certain level of within-image co-saliency in each image. In the benchmark dataset, it assumes that each image has and only contains one class of co-salient objects.

By detecting and comparing the instance-level salient objects in "WICOS", 254 images are classified as simple images and the remaining 110 images are thought as complex images. Specifically, the pretrained GoogLeNet is used as the baseline network to detect the instance-level salient object as suggested by the publicized code of [22], and all the parameters follow their default setting. We set the predefined threshold $\lambda = 20$ to classify simple and complex images. λ is set based on human supervisions inferred from a small number of sample images in iCoseg dataset [45]. In the easy learning step, to smooth the saliency map, we linearly fuse the refined binary map M_r after minimizing the Eq. (1), the DCL [39] map and RFCN [25] map with the fusion weights {0.8, 0.1, 0.1} as the final within-image co-saliency map for each simple image.

Following the experiment setting in [7], seven within-image saliency detection methods and one within-image co-saliency detection method are chosen as comparison methods: LRK [47], SR [48], FT [49], CWS [11], RC [1], DCL [39], RFCN [25] and CDS [7]. The first five are traditional feature-based methods, DCL and RFCN are deep CNN based methods, the last one (CDS) is a bottom-up method to detect within-image co-saliency.

We evaluate our performance on the commonly used precision-recall (PR) curve, maximum *F*-measure (maxF) with changing thresholds. MAE error defined as the average absolute difference between the result and ground truth [7,39] is also evaluated. Furthermore, the average precision, recall and *F*-measure with adaptive thresholds (twice the mean saliency value) [39,49] are also reported in our experiment. The *F*-measure is computed as $F_\gamma = \frac{(1+\gamma^2)\times Precision\times Recall}{\gamma^2\times Pecision+Recall}$, where γ^2 is set to 0.3 as defined in [39,49].

There is only one publicized dataset ("WICOS") for the problem of within-image co-salient object detection [7], so we synthesize new datasets to further evaluate our method. There are some existing publicized dataset for cross-image co-saliency detection like [45,50], we select dataset in [50] to synthesize the new datasets. Li and Ngan [50] collected a dataset includes 105 image pairs where each image of a pair contains one or more similar co-salient objects. Each image's scale is less than 200×200 pixels. We simply concatenate two images of a pair with similar object(s) into a new image. The new synthetic images are aligned horizontally. We remove the ones that contain multi-class co-salient objects (e.g., two black cows and two yellow cows), leading to 86 images finally. This dataset is named "SYN-Normal" (86 images).

In order to better evaluate the methods, we randomly add ellipse-shaped noises to background locations of each synthetic image, leading to another new dataset named "SYN-Noisy" (86 images). Fig. 5 shows how we generate the new synthetic dataset. Then, we report the evaluation results of our method and other comparison methods on "SYN-Normal" and "SYN-Noisy" respectively.

4.2. Results on "WICOS" dataset

In this section, we report the experimental results on the "WICOS" dataset. We separately evaluate our method on simple images, complex images and all the images. Fig. 6 shows the PR curves of the proposed method and other eight comparison methods on simple, complex and all images, respectively. From the figure, we see that the proposed method achieves the best PR curve in detecting the within-image co-saliency. Table 1 shows the maximum *F*-measure and MAE error results. From this table, we find that the proposed method achieves the best performance among the three image sets. Specifically, the proposed method get higher

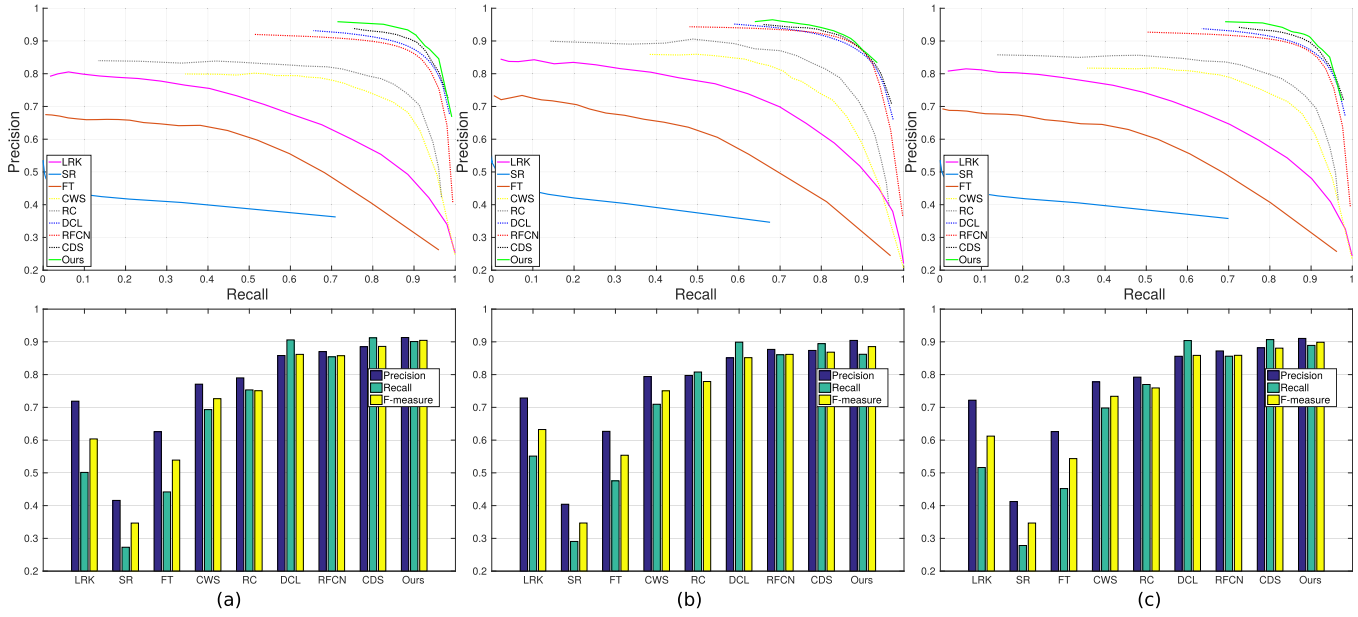

Fig. 6. PR curves with changing thresholds (first row) and average Precision, Recall and *F*-measure with adaptive threshold (second row) by the proposed method ('Ours') and other comparison methods on the "WICOS" dataset. We show results for: (a) easy learning on simple images (254 images). (b) Hard learning on complex images (110 images). (c) Easy-to-hard learning on all images (364 images). The baseline algorithms for superpixels' feature extraction are shown in dash lines.

Table 1
The maximum *F*-measure (maxF) and MAE error of the proposed method ('Ours') and the comparison methods on simple images (254 images), complex images (110 images) and all images (364 images) of the "WICOS" dataset [7]. Larger maxF and smaller MAE error indicate better performance.

Images	Metrics	LRK	SR	FT	CWS	RC	DCL	RFCN	CDS	Ours
Simple	maxF (%)	65.8	40.9	57.7	76.1	79.3	88.8	87.8	90.4	**92.2**
	MAE error	0.247	0.255	0.252	0.171	0.147	0.060	0.086	0.051	**0.047**
Complex	maxF (%)	70.8	39.0	59.1	78.4	82.6	88.8	89.4	89.9	**90.6**
	MAE error	0.229	0.225	0.225	0.152	0.131	0.056	0.077	0.050	**0.048**
All	maxF (%)	67.4	40.3	58.2	76.7	80.2	88.8	88.3	90.3	**91.3**
	MAE error	0.241	0.246	0.244	0.165	0.142	0.059	0.083	0.050	**0.047**

maxF and lower MAE error no matter on simple, complex or all images.

We also calculate the average Precision, Recall and *F*-measure by adaptive thresholds [39,49] and show them in Fig. 6. The proposed method achieves the best average Precision and *F*-measure among all the methods. The high Precision indicates the effectiveness and accuracy of the proposed method for within-image co-saliency detection. For the Recall performance, the proposed method achieves comparable results with the best one. That is because our easy-to-hard learning strategy tends to focus more on the accurate within-image co-salient regions and slightly ignore some regions with low within-image co-saliency scores. In summary, the state-of-the-art performance is accomplished by the proposed method in terms of best PR curve, highest maxF, lowest MAE error, highest Precision, highest *F*-measure, and comparable Recall.

Fig. 7 shows some sample results of within-image co-saliency detection from the proposed method and other eight comparison methods. The top three rows are examples of the complex image and the bottom three rows show the results of the simple image. Because the first seven methods mainly focus on solving the within-image saliency detection problem, they tend to highlight all the salient regions in an image but not de-emphasize the salient regions without co-saliency.

Compared to the CDS [7] method specially designed for within-image co-saliency detection, the proposed method is better at highlighting the co-salient regions and de-emphasizing the salient regions without co-saliency. The reason of our method achieves a better result than the CDS method is that our method can accurately remove most of the noisy objects in the simple images selected by the feature comparing based classification method in the easy learning step, and the MRF refinement step combine with the MIL algorithm can somehow relieve the influence of the inaccurate supervision information from the easy learning results.

In the hard learning step, we select scale number as 5 and five MIL models are trained on five different superpixel scales independently. We use the SLIC over-segmentation method [51] to generate the superpixels in five scales by changing the desired superpixel number as {100, 200, 400, 800, 1600}. The 254 simple images are used as the training set and the remaining 110 images are utilized as the testing set. During training the MIL model, one positive and one negative bags with $m = 30$ instances are selected by the sampling method from each simple image for training. During testing, we define each complex image as a testing bag and treat all the inside superpixels as testing instances. As a result, trained MIL models generate five saliency maps in different scales. Finally, we linearly fuse the five saliency maps to get our final detection result for each complex image.

4.3. Ablation study on "WICOS" dataset

To show the effectiveness of some key steps in our easy to hard learning strategy, we compare the results on three different kinds of experiment setups: (1) remove the hard learning step, and get final saliency maps of complex images with the same method as simple images; (2) remove the MRF refinement step in easy learning step and directly use the results without

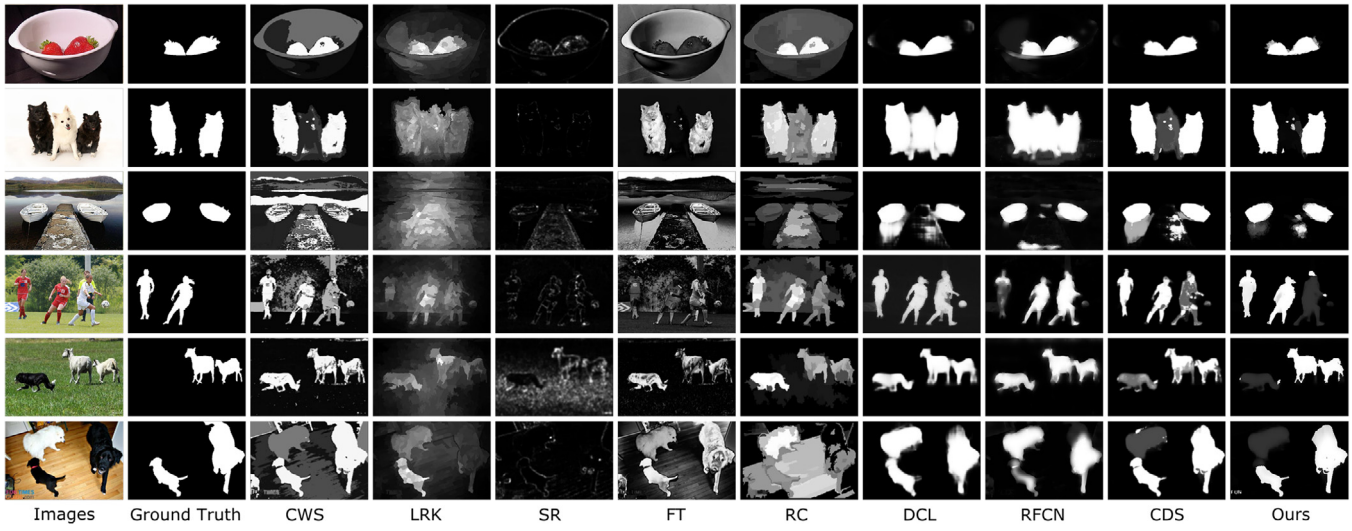

Fig. 7. Sample results by the proposed method ('Ours') and other comparison methods on the "WICOS" dataset. The top three rows are examples of the complex image and the bottom three rows show the results of the simple image.

Table 2
Results of ablation study on "WICOS" dataset. HL denotes the hard learning step, MRF indicates the MRF refinement step in easy learning, and MIL refers to the MIL step in hard learning.

Dataset	Precision (%)	Recall (%)	*F*-measure (%)	MAE error
w/o HL	89.4	90.0	88.9	0.050
w/o MRF	88.8	89.1	88.1	0.052
w/o MIL	89.8	**90.2**	89.2	0.061
Ours	**91.1**	88.9	**89.9**	**0.047**

refinement as the label for hard learning step; (3) replace the MIL step with directly linearly fusing the saliency maps of the five selected saliency methods. Table 2 shows the results on all the images on the "WICOS" dataset using adaptive thresholds.

In experiment (1) w/o HL, the results of complex images concern more about all the salient objects in the image but do not consider enough about the co-saliency information, so it results of high recall but low precision. For experiment (2) w/o MRF, the results without the MRF refinement step contain much more noisy information, so low precision and low recall results are obtained. For experiment (3) w/o MIL, we directly linearly fusing the saliency maps of the five selected saliency methods. The methods CWS, RC, DCL and RFCN are for within-image saliency detection and the method CDS is for within-image co-saliency detection. This fusion method tends to highlight all the salient objects in the image, leading to a higher recall. However, the proposed method tends to highlight the co-salient objects and de-emphasize other salient objects without co-saliency in the image, so the proposed method might remove some pixels, leading to a slightly lower recall but a higher precision. Because precision is more important in saliency detection, the proposed method is more advanced. The results show that the steps of hard learning, MRF, and MIL in the proposed method are necessary to help to improve the performance.

4.4. Results on "SYN-Normal" and "SYN-Noisy" datasets

We detect the within-image co-salient objects by the proposed easy-to-hard leaning strategy on the synthetic datasets. Some sampled results on the synthetic datasets are displayed in Fig. 8. The evaluation results are shown in Table 3. From the experimental results, we can see that our method achieved the best performance in terms of *F*-measure, MAE error and maxF on

Table 3
The average Precision, Recall, *F*-measure of a adaptive threshold, maximum *F*-measure (maxF) and MAE error of the proposed method ('Ours') and the comparison methods on the synthetic "SYN-Normal" dataset (86 images) and "SYN-Noisy" dataset (86 images).

Images	Methods	Precision (%)	Recall (%)	*F*-measure (%)	MAE error	maxF(%)
SYN-Normal	LRK	28.9	16.7	23.1	0.375	34.9
	SR	47.1	32.7	40.9	0.255	44.0
	FT	42.0	29.8	36.7	0.298	43.9
	CWS	73.2	50.7	63.9	0.302	68.9
	RC	71.9	67.7	69.0	0.188	72.8
	DCL	**89.0**	55.1	74.4	0.191	80.3
	RFCN	85.1	76.8	82.2	0.118	84.5
	CDS	85.2	**81.5**	83.7	0.089	85.6
	Ours	87.4	80.1	**84.5**	**0.086**	**85.7**
SYN-Noisy	LRK	18.5	10.5	15.1	0.348	33.5
	SR	44.1	31.0	38.3	0.255	41.0
	FT	34.4	25.8	30.4	0.316	36.3
	CWS	35.5	29.6	32.6	0.332	50.5
	RC	54.9	50.1	52.1	0.235	53.3
	DCL	59.4	46.3	53.9	0.228	60.6
	RFCN	70.3	65.2	67.8	0.162	68.6
	CDS	76.4	**72.4**	74.4	0.117	78.7
	Ours	**80.3**	67.6	**75.7**	**0.114**	**78.9**

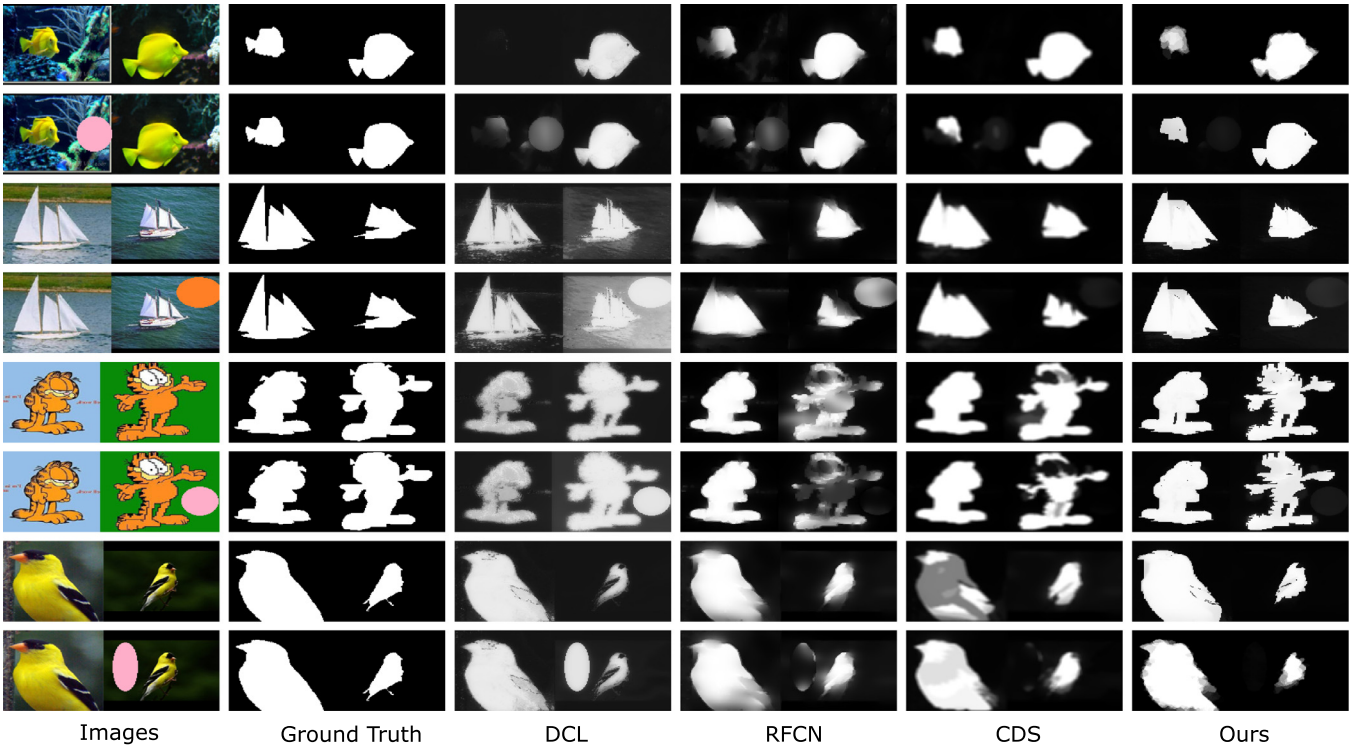

Fig. 8. Some example results on the synthetic "SYN-Normal" and "SYN-Noisy" datasets. Due to the page limit, we only show the results of the best four methods (Ours, CDS, DCL, RFCN) here.

"SYN-Normal" dataset and our method obtained the best performance in terms of Precision, *F*-measure, MAE error and maxF on "SYN-Noisy" dataset. The performance on "SYN-Noisy" dataset is worse than that on "SYN-Normal" dataset. We also see that the improvement by the proposed method on "SYN-Noisy" is more than that on "SYN-Normal". Taking *F*-measure as an example, our method gets 75.7% compared to the second best 74.4% on "SYN-Noisy", while our method gets 84.5% compared to the second best 83.7% on "SYN-Normal".

4.5. Running time

We calculate the average running time of each step per image on "WICOS" dataset. Our experiment were run on a workstation with 2.2 GHz CPU and our code was implemented in Matlab. The instance-level salient object detection [22] was performed on a NVIDIA Tesla P40 GPU card. With a parallel computing for each scale independently on the "WICOS" dataset, testing on a single image needs 25.7 s per image, and the easy learning step takes 5.7 s per image, and the hard learning step takes 26.2 s per image. During training and testing, most of the running time is spent on the SLIC superpixel-level feature extraction.

5. Conclusion

In this paper, we propose a new easy-to-hard learning method for within-image co-saliency detection. By incorporating pretrained model for instance-level salient object detection, the images are classified into simple and complex images. In simple images, we use an easy learning method to detect and refine the co-salient objects. Based on the imperfect labels obtained in simple images, we incrementally detect the within-image co-saliency in complex images. We model it as a hard learning problem with noisy labels for data fusion. A new multi-scale MIL model together with a sampling method is proposed to solve it. Experimental results show that the proposed easy-to-hard learning method achieves the state-of-the-art performance in terms of most evaluation metrics on the benchmark datasets.

Conflict of interest

We declare that we have no conflict of interest to this work.

Acknowledgments

This work is supported in part by the NSFC 61672089, 61273274, 61572064, 61672376, NSFC-U 1803264, National Key Technology R&D Program of China 2012BAH01F03, and NSF 1658987. Shaoyue Song and Cong Ma are supported by China Scholarship Council.

References

[1] M.-M. Cheng, N. Mitra, X. Huang, P. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[2] H. Yu, M. Xian, X. Qi, Unsupervised co-segmentation based on a new global GMM constraint in MRF, in: Proceedings of the IEEE International Conference on Image Processing, 2014, pp. 4412–4416.
[3] Y. Lin, Y. Tong, Y. Cao, Y. Zhou, S. Wang, Visual-attention-based background modeling for detecting infrequently moving objects, IEEE Trans. Circuits Syst. Video Technol. 27 (6) (2017) 1208–1221.
[4] R. Zhao, W. Oyang, X. Wang, Person re-identification by saliency learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 356–370.
[5] V. Mahadevan, N. Vasconcelos, Saliency-based discriminant tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1007–1013.
[6] H. Guo, K. Zheng, X. Fan, H. Yu, S. Wang, Visual attention consistency under image transforms for multi-label image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[7] H. Yu, K. Zheng, J. Fang, H. Guo, W. Feng, S. Wang, Co-saliency detection within a single image, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, 2018, pp. 7509–7516.
[8] R. Huang, W. Feng, J. Sun, Saliency and co-saliency detection by low-rank multiscale fusion, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2015, pp. 1–6.

[9] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, A. Borji, Detect globally, refine locally: a novel approach to saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[10] J. Han, H. Chen, N. Liu, C. Yan, X. Li, CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion, IEEE Trans. Cybern. 48 (11) (2018) 3171–3183.
[11] H. Fu, X. Cao, Z. Tu, Cluster-based co-saliency detection, IEEE Trans. Image Process. 22 (10) (2013) 3766–3778.
[12] D. Zhang, J. Han, C. Li, J. Wang, Co-saliency detection via looking deep and wide, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2994–3002.
[13] R. Huang, W. Feng, J. Sun, Color feature reinforcement for cosaliency detection without single saliency residuals, IEEE Signal Process. Lett. 24 (5) (2017) 569–573.
[14] D. Zhang, H. Fu, J. Han, A. Borji, X. Li, A review of co-saliency detection algorithms: fundamentals, applications, and challenges, ACM Trans. Intell. Syst. Technol. 9 (4) (2018) 1–31.
[15] J. Han, G. Cheng, Z. Li, D. Zhang, A unified metric learning-based framework for co-saliency detection, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2018) 2473–2483.
[16] K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, X. Qian, Y.-Y. Chuang, Unsupervised CNN-based co-saliency detection with graphical optimization, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 485–501.
[17] X. Xu, L. Wan, X. Liu, T.-T. Wong, L. Wang, C.-S. Leung, Animating animal motion from still, in: Proceedings of the ACM Transactions on Graphics, 27, 2008, p. 117.
[18] X. He, S. Gould, An exemplar-based CRF for multi-instance object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 296–303.
[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 740–755.
[20] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[21] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2017, pp. 2980–2988.
[22] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, R. Mech, Unconstrained salient object detection via proposal subset optimization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5733–5742.
[23] G. Li, Y. Xie, L. Lin, Y. Yu, Instance-level salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 247–256.
[24] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, J. Han, A self-paced multiple-instance learning framework for co-saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 594–602.
[25] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Saliency detection with recurrent fully convolutional networks, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 825–841.
[26] P. Zhang, W. Liu, H. Lu, C. Shen, Salient object detection with lossless feature reflection and weighted structural loss, IEEE Trans. Image Process. PP (99) (2019) 1.
[27] C. Ge, K. Fu, F. Liu, L. Bai, J. Yang, Co-saliency detection via inter and intra saliency propagation, Signal Process. Image Commun. 44 (2016) 69–83.
[28] H. Li, F. Meng, B. Luo, S. Zhu, Repairing bad co-segmentation using its quality evaluation and segment propagation, IEEE Trans. Image Process. 23 (8) (2014) 3545–3559.
[29] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012, (http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html).
[30] D. Guo, L. Zhu, Y. Lu, H. Yu, S. Wang, Small object sensitive segmentation of urban street scene with spatial adjacency between object classes, IEEE Trans. Image Process. 28 (6) (2019) 2643–2653.
[31] D. Zhang, J. Han, Y. Zhang, Supervision by fusion: towards unsupervised learning of deep salient object detector, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4048–4056.
[32] J. Zhang, T. Zhang, Y. Dai, M. Harandi, R. Hartley, Deep unsupervised saliency detection: a multiple noisy labeling perspective, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9029–9038.
[33] G. Li, Y. Xie, L. Lin, Weakly supervised salient object detection using image labels, Thirty-Second AAAI Conference on Artificial Intelligence, 2018, arXiv:1803.06503.
[34] D. Zhang, J. Han, L. Zhao, D. Meng, Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework, Int. J. Comput. Vis. 127 (4) (2019) 363–380.
[35] L. Han, D. Zhang, H. Dong, X. Chang, J. Ren, S. Luo, J. Han, Self-paced mixture of regressions, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.
[36] D. Zhang, J. Han, L. Yang, D. Xu, SPFTN: a joint learning framework for localizing and segmenting objects in weakly labeled videos, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2018) 1.
[37] Q. Wang, Y. Yuan, P. Yan, X. Li, Saliency detection by multiple-instance learning, IEEE Trans. Cybern. 43 (2) (2013) 660–672.
[38] B. Lai, X. Gong, Saliency guided end-to-end learning for weakly supervised object detection, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 2053–2059.
[39] G. Li, Y. Yu, Deep contrast learning for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[40] Y. Boykov, M.-P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in ND images, in: Proceedings of the IEEE International Conference on Computer Vision, 2001, pp. 105–112.
[41] C. Rother, V. Kolmogorov, A. Blake, GrabCut: interactive foreground extraction using iterated graph cuts, ACM Trans. Graph. 23 (3) (2004) 309–314.
[42] B. Babenko, P. Dollár, Z. Tu, S. Belongie, Simultaneous learning and alignment: multi-instance and multi-pose learning, in: Proceedings of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, 2008.
[43] S. Tsogkas, I. Kokkinos, Learning-based symmetry detection in natural images, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 41–54.
[44] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[45] D. Batra, A. Kowdle, D. Parikh, J. Luo, T. Chen, iCoseg: interactive co-segmentation with intelligent scribble guidance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3169–3176.
[46] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5455–5463.
[47] X. Shen, Y. Wu, A unified approach to salient object detection via low rank matrix recovery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 853–860.
[48] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[49] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.
[50] H. Li, K.N. Ngan, A co-saliency model of image pairs, IEEE Trans. Image Process. 20 (12) (2011) 3365–3375.
[51] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.

Shaoyue Song received the B.E. degree from Beijing Jiaotong University, Beijing, China, in 2014, and is currently working towards the Ph.D. degree at Beijing Jiaotong University. In 2018, she was a visiting student at the University of South Carolina, Columbia, supported by the China Scholarship Council (CSC). Her research interests include salient object detection and image classification.

Hongkai Yu received the Ph.D. degree in computer science and engineering from University of South Carolina, Columbia, SC, USA in 2018. He then joins the Department of Computer Science at University of Texas-Rio Grande Valley, Edinburg, TX, USA as an assistant professor. His research interests include computer vision, machine learning, deep learning and intelligent transportation system. He is a member of the IEEE.

Zhenjiang Miao (M'11) received the B.E. degree from Tsinghua University, Beijing, China, in 1987, and the M.E. and Ph.D. degrees from Northern Jiaotong University, Beijing, in 1990 and 1994, respectively. From 1995 to 1998, he was a Post-Doctoral Fellow with the école Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications, Institut National Polytechnique de Toulouse, Toulouse, France, and was a Researcher with the Institute National de la Recherche Agronomique, Sophia Antipolis, France. From 1998 to 2004, he was with the Institute of Information Technology, National Research Council Canada, Nortel Networks, Ottawa, Canada. He joined Beijing Jiaotong University, Beijing, in 2004. He is currently a Professor, Director of the Media Computing Center, Beijing Jiaotong University, and Director of the Institute for Digital Culture Research, Center for Ethnic & Folk Literature & Art Development, Ministry Of Culture, P.R. China. His current research interests include image and video processing, multimedia processing, and intelligent human-machine interaction.

Dazhou Guo is currently a Ph.D. candidate in Computer Science at University of South Carolina, USA. He received the B.S. degree in Electronic Engineering from Dalian University of Technology, Dalian, China, in 2008, the M.S. degree in Information and Informatics Engineering from Tianjin University, Tianjin China, 2010. His research interests include computer vision, medical image processing, and machine learning.

Wei Ke received the Ph.D. degree in electrical engineering from University of Chinese Academy of Sciences, Beijing, China, in 2018. He is currently a PostDoc Researcher with the School of Computer Sciences, Carnegie Mellon University. In 2016, he visited the Center for Machine Vision and Signal Analysis at University of Oulu as a joint Ph.D. student, supported by China Scholarship Council (CSC). His research interests include computer vision and deep learning. He has published 10 papers in refereed conferences and journals including IEEE CVPR and ECCV. He is the winner of the President Award of Chinese Academy of Sciences in 2017.

Cong Ma received the B.E. degree from Beijing Jiaotong University, Beijing, China, in 2013, and is currently working towards the Ph.D. degree at Beijing Jiaotong University. In 2018, he was a visiting student at the University of California, Merced, supported by the China Scholarship Council (CSC). His current research interests include visual tracking and video analysis.

Song Wang received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2002. He was a Research Assistant with the Image Formation and Processing Group, Beckman Institute, UIUC, from 1998 to 2002. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA, where he is currently a Professor. His current research interests include computer vision, image processing, and machine learning. He is currently serving as the Publicity/Web Portal Chair of the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society, an Associate Editor of IEEE Transaction on Pattern Analysis and Machine Intelligence, Pattern Recognition Letters, and Electronics Letters. He is a Senior Member of the IEEE and a member of the IEEE Computer Society.