Quality-Guided Fusion-Based Co-Saliency Estimation for Image Co-Segmentation and Colocalization

Koteswar Rao Jerripothula, *Member, IEEE*, Jianfei Cai, *Senior Member, IEEE*, and Junsong Yuan, *Senior Member, IEEE*

***Abstract*—Despite the advantage of exploiting interimage information by performing joint processing of images for co-saliency, co-segmentation, or co-localization, it introduces a few drawbacks: 1) its necessity in scenarios where the joint processing might not perform better than individual image processing; 2) increased complexity over individual image processing; and 3) complex parameter tuning. In this paper, we propose a simple cosaliency estimation method where we fuse saliency maps of different images using the dense correspondence technique. More important, the co-saliency estimation is guided by our proposed quality measurement that helps decide whether the saliency fusion really improves the quality of the saliency map or not. Our basic idea for developing the quality metric is that a high-quality saliency map should have well-separated foreground and background, as well as a concentrated foreground like ground-truths. Extensive experiments on several benchmark datasets including the large-scale dataset, ImageNet, for the applications of foreground co-segmentation and co-localization show that our proposed framework is able to achieve very competitive results.**

***Index Terms*—Co-saliency, co-segmentation, co-localization, fusion, foreground, quality.**

I. Introduction

FOREGROUND segmentation or localization is a very useful and important step for many vision and multimedia applications such as recognition and streaming, since it separates the object of interest from the background and thus facilitates more efficient subsequent processing or understanding. When dealing with only a single image, visual saliency has been a common cue used for highlighting the foreground. However, single-image saliency has obtained limited success when dealing with images that have cluttered backgrounds, or images where the foreground has similar attributes as the background, which cause the object of interest to be less salient. Recognizing the limitations of individual processing, in recent years various joint processing works such as co-saliency [1]–[5], co-segmentation [6]–[12], co-localization [13], [14], co-skeletonization [15], knowledge transfer [16], [17] have been proposed, and have been demonstrated quite effective in extracting foregrounds in a batch mode. The basic idea of all these works is to exploit the commonness across a set of images that contain some common object, which gives inter-image prior information about the common object, a clear advantage that certainly lacks in the individual processing.

Despite such an advantage, the existing joint processing algorithms also bring in new challenges. 1) Due to the way of co-labeling pixels [18] or co-selection of bounding boxes [13] in a set of images, most of the existing high-performance joint processing algorithms are usually complicated with large numbers of variables, which unavoidably have the scalability issue. 2) As shown in [18], [19], joint processing of images might not perform better than individual processing. The recently proposed video co-localization work [20] also cannot perform better than the individual processing [21]. This certainly raises up the question: *Given a set of images for foreground segmentation or localization, should we process them jointly or individually?* 3) For effective co-segmentation or co-localization, the existing joint processing algorithms usually require tuning quite a few parameters, which further increases the complexity, especially when dealing with large and diverse datasets.

To address the above challenges, in this paper we propose a co-saliency framework, where we explore inter-image information via co-saliency and then perform co-saliency based segmentation or localization on individual images. Such an approach of performing the joint process on saliency maps, i.e., co-saliency, while performing individual processing on the targeted applications (segmentation and localization), helps us overcome the increased complexity of the conventional way of co-segmentation or co-labeling on multiple images simultaneously. At the heart of our co-saliency framework are two key components (saliency quality measurement and fusion based co-saliency) to handle the remaining two challenges, which are joint *vs* individual processing issue and the parameters dependence, respectively.

Manuscript received June 20, 2017; revised October 21, 2017; accepted December 9, 2017. Date of publication January 25, 2018; date of current version August 14, 2018. This work was supported in part by Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE2015-T2-2-114. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ivan V. Bajic. *(Corresponding author: Koteswar Rao Jerripothula.)*

K. R. Jerripothula is with the Graphic Era University, Dehradun 248002, India (e-mail: krjimp@geu.ac.in).

J. Cai and J. Yuan are with Nanyang Technological University, Singapore 639798 (e-mail: asjfcai@ntu.edu.sg; jsyuan@ntu.edu.sg).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes additional results. Contact krjimp@geu.ac.in for further questions about this work.

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

Digital Object Identifier 10.1109/TMM.2018.2798294

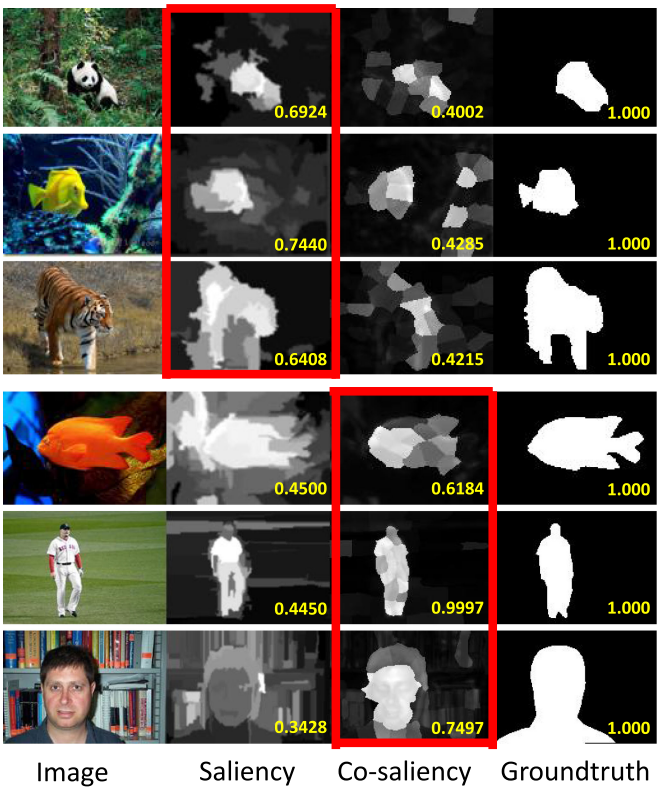

Fig. 1. The desired saliency map that highlights the object could either be the original saliency map itself (e.g., the first three rows) or the jointly processed co-saliency map (e.g., the last three rows). Our quality measurement gives appropriate scores (shown at the bottom-right of each saliency map) for selecting the right ones (inside red boxes). The selected ones are basically the ones closer to the ground-truths' quality without having the ground-truths.

Quality measurement: In the first component, we propose a metric to measure and compare the quality of each saliency map with that of its corresponding co-saliency map, so as to answer the second challenge, i.e., joint processing or not. Our quality metric is developed based on two empirical observations: 1) a better saliency map should have a better separation between the foreground and the background; 2) a better saliency map should have a better foreground concentration, i.e., preferring the foreground to be a concentrated saliency region. These observations can certainly be made for ground-truth masks, and the attempt here is to measure the closeness of saliency maps to the quality of such ground-truths without having them. Fig. 1 gives several examples, wherein the first three examples, the object regions are better highlighted in the individually processed saliency maps [22] compared to the jointly processed co-saliency maps [23], while in the last three examples, the co-saliency maps look better than the individually processed saliency maps. For all these examples in Fig. 1, our proposed metric generates appropriate quality scores as shown at the bottom-right of each map. These scores help us select between the jointly processed and the individually processed saliency maps, which tackles the joint *vs* individual processing issue. It can also be seen that our method gives score 1 for all the ground-truths as per the expectation. Note that [24] previously compared different saliency maps of an image obtained by different saliency detection methods in a supervised manner, whereas we make comparison between a saliency map and a co-saliency map, and we do it in a completely unsupervised way.

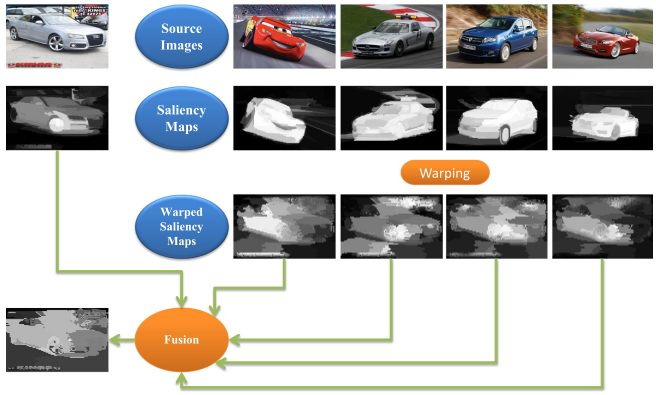

Fig. 2. An example to show that through the joint processing *via* warping and fusion, the last four images having salient common objects (car) could render help to the first image where the common object (car) has only weak saliency.

Fusion based co-saliency: The second component of our framework is to perform joint processing in such a manner that it doesn't introduce unnecessary parameters for tuning, i.e., the last challenge. Therefore, for each image, we fuse the saliency map with the saliency information from other images so as to boost up the common object saliency while suppressing the background saliency. Our basic idea is to make use of the existing techniques on dense correspondences [25] to align individual object pixels for saliency fusion. Fig. 2 illustrates the proposed joint process for generating co-saliency map. In particular, for one image, the individual saliency maps of its neighbors are warped to align with its own saliency map and then all the aligned saliency maps are fused together. In this way, although most parts of the car in the first image are weakly salient originally, they become salient by acquiring the general saliency across the images through fusion. The underlying assumption here is that the common object or its parts are salient in general, if not in every image. Such an assumption allows us to eradicate any parameter dependence as far the joint processing is concerned. Interestingly, our method with the default setting (without tuning any parameter) itself gives competitive results. In fact, we have applied our method on entire 1 million images of ImageNet for co-localization without tweaking any parameter, and we comfortably outperform earlier state-of-the-art [13], [26] for this task.

We would like to point out that our preliminary studies on the saliency quality measurement and the fusion-based co-saliency have been reported in [27] and [28], respectively. This journal extension combines the two early works together to have a unified framework for selecting and generating a better saliency map for each image for the applications of foreground segmentation or localization. In addition to the overall integration, more explanations, and more comprehensive experimental results, we also add the following extensions to the two individual components. In particular, unlike [27], which only considers the separation criterion in the quality metric, in this paper we further incorporate the foreground concentration criterion to improve

the saliency quality measurement. Also, in [28], we only considered geometric mean for saliency fusion, while in this paper we compare different ways for saliency fusion.

In summary, this paper makes the following major contributions: 1) designing a metric for saliency quality measurement to combine the strengths of both the individual processing and the joint processing; 2) developing a simple saliency fusion based co-saliency estimation method for overcoming the complexity and the parameter setting challenges; 3) achieving good results comparable to state-of-the-art methods in the applications of foreground segmentation and localization on several benchmark datasets including the large-scale dataset, ImageNet. Our entire framework can be easily modified to cope with other scenarios such as when there are groud-truth segmentation/localization maps available for some of the images or to reduce the complexity when dealing with large-scale datasets.

II. Related Work

Our method is closely related to co-saliency, co-segmentation and co-localization research.

A. Co-saliency

Co-saliency typically refers to the common saliency existing in a set of images containing similar objects. The term co-saliency was first coined in [29] in the sense of what is unique in a set of similar images, and the concept was later linked to extracting common saliency, which is very useful for many practical applications [23], [30]. For example, co-saliency object priors have been efficiently used for co-segmentation in [31]. A cluster based co-saliency method using various cues was proposed in [1], which learns the global correspondence and obtains cluster saliency quite well. However, application of co-saliency method [1] is limited to images of the same object captured at different viewpoints or instances. It cannot well handle image-sets with huge intra-class variation as it uses the color feature. We try to overcome this challenge by grouping (or clustering) of images followed by histogram analysis. In [32], it introduced deep intra-group semantic information and wide cross-group heterogeneousness information for co-saliency detection. In this way, they can capture the concept-level properties of the co-salient objects and suppress the common backgrounds in the image group. In contrast to this supervised learning based method, our method is unsupervised (weakly supervised to be particular). Also, none of the previous methods account for the issue of whether to process jointly or not.

B. Co-segmentation

The concept of co-segmentation was first introduced by Rother *et al.* [33], who used histogram matching to simultaneously segment out the common object from a pair of images. Since then, many co-segmentation methods have been proposed to either improve the segmentation in terms of accuracy and processing speed [34]–[38] or scale from image pair to multiple images [34], [39], [40]. [34] proposed a discriminative clustering framework and [39] used optimization for co-segmentation. [41] combined co-segmentation with co-sketch for effective co-segmentation. [18] adopted dense SIFT matching to discover common objects and co-segment them. These methods are quite complicated and require parameter tuning for high performance, whereas our method, although simple, achieves competitive performance even without tuning the parameters.

For large scale foreground extraction, [42] showed how they could segment half-million images using the transfer of human annotated segmentation masks. So far, this method is state-of-the-art in applying co-segmentation on Imagenet [43] dataset. Previously as well, there has been such attempts like in [39], but the evaluation was very limited as they use the bounding box for the evaluation whereas [42] uses proper segmentation masks for subsets of images. [44] is another co-segmentation method that is highly scalable to perform foreground extraction in large-scale datasets where they improve upon the results obtained by GrabCut [45] by finding the optimal hyperplane that can separate foreground and background in feature space. A simple modification (as suggested in our work [46]) to our original idea makes it more efficient (reducing complexity from quadratic to linear) for the large-scale application.

C. Co-localization

Co-localization is also similar to co-segmentation in terms of idea, i.e., using multiple images, but the output is a bounding box around the object instead of object segment. This has been introduced by [13] along with handling noisy datasets, where it is able to avoid assigning the bounding box if the image does not contain the common object. The performance of this method has been further improved in [20]. However, it's a joint framework optimizing over all the images, whereas we explore inter-image information via co-saliency and then perform co-saliency based localization on individual images. Another work [47] proposes a generic co-localization where objects across the images need not be common. Our method assumes that all the images contain the common object and aims for higher performance compared to the existing methods. Slightly different from the co-localization, there are some bounding-box propagation algorithms [17], [48] where some images already have bounding boxes and they are utilized to localize unannotated images. It is like a supervised scenario. Our method can effectively adapt to this by considering ground truth bounding box as the high-quality saliency map already.

III. Proposed Method

In this section, we have five points to discuss: 1) our objective and the proposed solution, 2) the quality measurement system, 3) how images interact, 4) more efficient way of interacting, and 5) applications.

A. Objective and Proposed Solution

Let $\mathbf{I} = \{I_1, I_2, \ldots, I_m\}$ be an image-set containing m similar images. Denote set of their corresponding saliency maps as $\mathbf{S} = \{S_1, S_2, \ldots, S_m\}$. Functions $\phi(\cdot)$ and $\psi(\cdot)$ denote quality functions for the separation measure (between foreground and

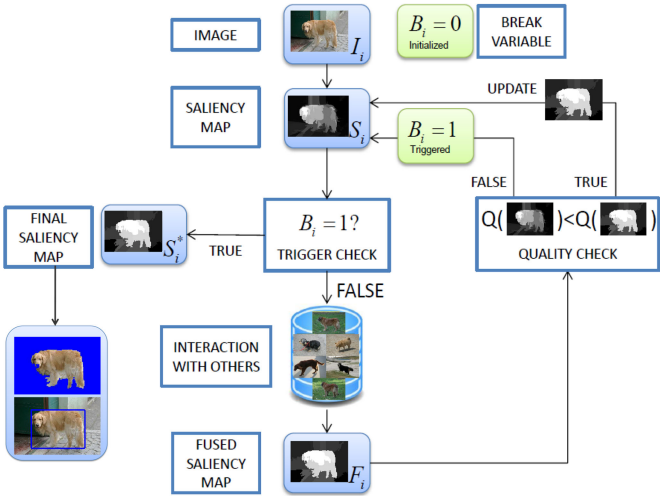

Fig. 3. Flowchart of the proposed method (for an image I_i): Saliency map S_i is iteratively updated by the fused saliency map F_i as long as the fused saliency map F_i is of higher quality. Drop in the quality (Q) triggers the break variable B_i to stop any further updates for the image.

background) and the concentration measure (of foreground), respectively. We define the total quality of any saliency map S_i as the product of these two measures, i.e., $\phi(S_i)\psi(S_i)$. In the pursuit of common object discovery through interaction, we assume that higher quality saliency maps are better. Therefore, we define our objective as

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} \sum_{i=1}^{m} \phi(S_i)\psi(S_i)$$
$$\text{s.t. } S_i \in \{S_i^k | k = 1, \ldots, K\}, \quad (1)$$

where we want to achieve saliency map set $\mathbf{S}^*$ such that the total quality of comprising saliency maps is maximum, and S_i can be any saliency map of an image ranging from the original saliency map to the saliency map obtained after K interactions, where K is set as 5 by default. During the interaction process, saliency maps from other similar images fuse together with the saliency map of each image to develop its fused saliency map. Denote $\mathbf{F} = \{F_1, F_2, \ldots, F_m\}$ as the set of such fused saliency maps resulted by such interaction of similar images.

We propose the following approach to achieve our objective. After the interaction, if the quality of saliency map improves by the fusion process, then only corresponding fused saliency maps can update the current saliency map. Otherwise, current saliency map is considered as the final one. In this manner, total saliency quality of set increases progressively. Different images may obtain their final saliency maps at different iterations. To track them and avoid further fusion for them, we define break variable B_i (set as 0 initially), which gets triggered at such an occurrence for image I_i. Fig. 3 depicts the flowchart for this. However, in the supervised scenario, saliency maps of images having annotations are replaced with the annotations, and their B_i is triggered right in the beginning.

Therefore, saliency map S_i^k, fused saliency map F_i^k at kth iteration, and B_i help in determining S_i^{k+1} (saliency map at next iteration) in the following manner:

$$S_i^{k+1} = \begin{cases} S_i^k, & \text{if } B_i = 1; \\ F_i^k, & \text{if } B_i = 0 \text{ and } \phi(F_i^k)\psi(F_i^k) > \phi(S_i^k)\psi(S_i^k); \\ S_i^k, & \text{if } B_i = 0 \text{ and } \phi(F_i^k)\psi(F_i^k) < \phi(S_i^k)\psi(S_i^k), \end{cases} \quad (2)$$

where the first case denotes that image has already achieved its final saliency map and there is no need for an update. The second case denotes that image has not yet achieved its final saliency map, and since the quality has improved by fusion, fused saliency map updates the current saliency map. The third case denotes that although the image has not yet achieved its high quality saliency map, but since quality has decreased by fusion, there is no need for an update and current saliency map is taken as final one. And it's the third case that triggers B_i.

Since we ensure that no way a lower quality fused saliency map can update the current saliency map, total saliency quality of the set $\mathbf{S}$ therefore can only get higher after any given iteration, and algorithm eventually stops when either $\forall B_i = 1$ or $k = K$. At this point, we have our $\mathbf{S}^*$.

B. Quality Measurement System

In this section, we propose two measures for determining the quality of any given saliency map S: (i) separation measure, which measures the separation between foreground and background; and (ii) concentration measure, which measures how concentrated the foreground pixels are. In order to assign likelihoods (foreground or background), we apply Otsu's threshold on S.

1) Separation Measure (ϕ): A high-quality saliency map should have well-separated foreground and background likelihoods like a ground-truth binary mask. Assuming distributions of these likelihoods to be of Gaussian in nature, we attempt to measure the separation between the two. Let $\mu_f(S)$, $\mu_b(S)$, $\sigma_f(S)$, and $\sigma_b(S)$ denote foreground mean, background mean, foreground standard deviation, and background standard deviation, respectively, computed based on the two likelihood distributions (obtained by Otsu thresholding). Let us denote $D_f(z; S)$ and $D_b(z; S)$ as foreground and background Gaussian distributions, respectively, where z takes saliency value ranging between 0 and 1. Specifically,

$$D_f(z; S) = \frac{e^{-\left(\frac{z-\mu_f(S)}{\sigma_f(S)}\right)^2}}{\sigma_f(S)\sqrt{2\pi}} \text{ and } D_b(z; S) = \frac{e^{-\left(\frac{z-\mu_b(S)}{\sigma_b(S)}\right)^2}}{\sigma_b(S)\sqrt{2\pi}}, \quad (3)$$

plotted in Fig. 4 as an example. It is clear that the less the two distributions overlap with each other, the better the saliency map is, i.e., the foreground and background are more likely to be separable. In order to calculate such overlap, it is needed to figure out the intersecting point z^* (see Fig. 4). It can be obtained by equating the two functions, i.e., $D_f(z; S) = D_b(z; S)$, which

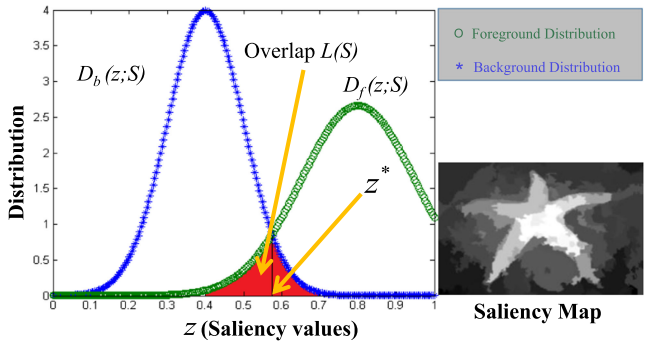

Fig. 4. Separation measure (ϕ) of quality of saliency map is measured using overlap of estimated likelihood distributions of the two classes: Foreground and Background.

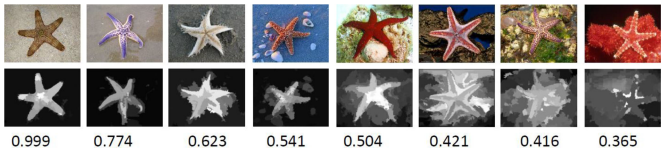

Fig. 5. Sample Images with their saliency maps and separation measures (ϕ) of quality. Saliency maps with low scores fail to highlight the starfish effectively.

finally leads to

$$z^2\left(\frac{1}{\sigma_b^2}-\frac{1}{\sigma_f^2}\right)-2z\left(\frac{\mu_b}{\sigma_b^2}-\frac{\mu_f}{\sigma_f^2}\right)+\frac{\mu_b^2}{\sigma_b^2}-\frac{\mu_f^2}{\sigma_f^2}+2\log\left(\frac{\sigma_b}{\sigma_f}\right)=0. \tag{4}$$

Note that we have omitted expressing "(S)" along with the means and variances for clarity. When we solve the above quadratic equation, we get

$$z^*=\frac{\mu_b\sigma_f^2-\mu_f\sigma_b^2}{\sigma_f^2-\sigma_b^2}\pm\frac{\sigma_f\sigma_b}{\sigma_f^2-\sigma_b^2}\times\Big((\mu_f-\mu_b)^2-2\left(\sigma_f^2-\sigma_b^2\right)\left(\log\left(\sigma_b\right)-\log\left(\sigma_f\right)\right)\Big)^{\frac{1}{2}}. \tag{5}$$

Having obtained z^*, overlap $L(S)$ can now be computed as

$$L(S)=\int_{z=0}^{z=z^*}D_f(z;S)\,dz+\int_{z=z^*}^{z=1}D_b(z;S)\,dz\,. \tag{6}$$

And finally, separation measure ϕ for saliency map S is calculated as

$$\phi(S)=\frac{1}{1+\log_{10}\left(1+\gamma\,L(S)\right)}. \tag{7}$$

where γ is set as number of bins used for representing the two distributions. In Fig. 5, we show a set of images with their saliency maps and separation measures. It can be seen that saliency maps become unfit to highlight starfish as separation measure decreases from top-left to the bottom-right.

2) Concentration Measure (ψ): A high-quality saliency map should also have concentrated foreground pixels like in a ground-truth, especially if there is a single foreground object.

TABLE I
ILLUSTRATION OF OUR CONCENTRATION MEASURE ψ BY VARYING THE $\mathbf{O}(S)$

$\mathbf{O}(S)$	$\psi(S)$
$\{100\}$	1
$\{90,10\}$	0.95
$\{90,4,3,2,1\}$	0.92
$\{80,10,10\}$	0.867
$\{50,50\}$	0.75
$\{50,30,20\}$	0.667
$\{25,25,25,25\}$	0.438

Note: Numeric values here mean areas of the comprising object components and total foreground area is 100. It can be seen that it decreases as largest component's contribution decreases and number of object components increases.

Often they get distributed into multiple object components spatially. We use *bwconncomp* function of MATLAB to generate these object components from the Otsu thresholded saliency maps. Ideally, there should be one largest object component and other components (if any) will disperse from that component. Bigger the contribution of this largest component to the foreground, higher will be the concentration of foreground. At the same time, lesser the dispersion of foreground into several object components, higher will be the foreground concentration again. Let $\mathbf{O}(S)=\{O_1(S),O_2(S),\ldots,O_{|\mathbf{O}(S)|}(S)\}$ denote set of these object components. Contribution $C_u(S)$ of $O_u(S)$ towards the total foreground is measured as

$$C_u(S)=\frac{[O_u(S)]}{\sum_{u=1}^{|\mathbf{O}(S)|}[O_u(S)]}, \tag{8}$$

where $[\cdot]$ denotes area of $O_u(S)$ and $|\cdot|$ denotes cardinality. Essentially, it is the fraction of the total foreground area covered by the object component. Now, if $u^*=\arg\max_u C_u(S)$, then concentration measure ψ for S is calculated by

$$\psi(S)=C_{u^*}(S)+(1-C_{u^*}(S))\,\frac{1}{|\mathbf{O}(S)|}, \tag{9}$$

where the first term measures contribution made by the largest component and second term measures lowness of dispersion in the foreground by calculating the reciprocal of the total number of components. These two terms are adaptively balanced by the sum of remaining contributions, i.e., $1-C_{u^*}(S)$. This ensures that concentration measure always lies between $C_{u^*}(S)$ and 1. In Table I, we illustrate how a saliency map S having 100 pixels with foreground likelihood measures while varying its object components set $\mathbf{O}(S)$. It can be observed how concentration measure ψ decreases from top to bottom as the largest component's contribution decreases. And it also decreases with the increasing dispersion of the foreground. For instance, the third set shows lower concentration value than the second one because of higher dispersion, although both have same largest object component's contribution. Similar observations can be made from the visual examples in Fig. 6.

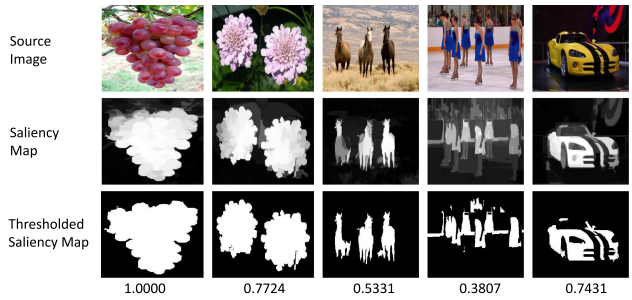

Fig. 6. Sample Images with their thresholded saliency maps and concentration measures (ψ) of quality. Saliency maps with dissipated foregrounds and concentrated foregrounds get lower scores and higher scores, respectively.

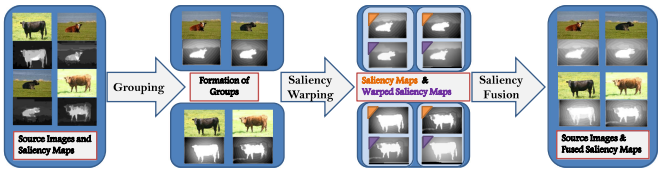

Fig. 7. Interaction process includes three steps: grouping, saliency warping, and saliency fusion.

C. Interaction

Images interact with each other hoping for saliency quality improvement. Our interaction process consists of 3 steps: grouping, saliency warping, and saliency fusion as shown in Fig. 7. Note that we employ the pre-processing step as described in our preliminary work [28], where we brighten up the initial saliency map and ensure that the saliency map highlights sufficient regions of the salient object before interaction is carried out. This is mainly due to the functions like geometric mean we use for saliency fusion, which over-penalize in the presence of low saliency values. Thus, we add this pre-processing step to avoid the bias.

1) Grouping: Considering intra-class variation that can exist in terms of the viewpoint, size, color, location etc, of the common objects, we divide images into a number of image groups so that images within the same group have somewhat similar appearances. Specifically, weighted GIST descriptor [49] (weighted by saliency map following [18]) is used to represent each image. We use k-means clustering for this grouping. Let there be N clusters. Denote Z_n as a set of images in the nth cluster, where $n \in \{1, \ldots, N\}$. In general, 10 images per group are good enough for our approach, and we set N accordingly. This grouping can also assist in feature selection for the matching purpose. Shape features such as SIFT may be reliable for feature matching in general, but it is not the same with color feature due to the possible intra-class variation in terms of color. But when it is the same object instance across images, color plays a very vital role, such as in the iCoseg dataset. Therefore, in order to adaptively detect such a case, for any group, we calculate a metric δ that measures the color histogram variance (across images in group) averaged over histogram bins using

$$\delta(Z_n) = \frac{1}{N_b} \sum_{j=1}^{N_b} \sqrt{\frac{1}{|Z_n|-1} \sum_{I_i \in Z_n} \left(H_{I_i}^{S_i}(j) - \hat{H}_n(j) \right)^2}, \quad (10)$$

where $H_{I_i}^{S_i}$ denotes normalized color histogram (with $N_b = 512$ bins indexed by j) of only the salient pixels in I_i. $\hat{H}_n$ denotes average of such histograms in Z_n. Higher the δ, more the color feature becomes unreliable for interaction. We consider only salient pixels because we assume that common objects' pixels are generally salient, and this gives some information about the common object. So, we concatenate color feature to the dense SIFT feature of images in the group Z_n, only if $\delta(Z_n) < \epsilon$ (See Section V for ϵ-setting).

2) Saliency Warping: Warping [25] basically is a process of aligning one image w.r.t. another by establishing dense correspondence. The idea behind saliency warping is that by alignment of corresponding pixels in other images to the pixels of an image, saliency information across corresponding pixels can be shared to estimate a suitable saliency value for the pixel. Following [18], masked dense correspondence [25] (masked by Otsu thresholded saliency map) is used to find the corresponding pixels. The difference is that feature used in our approach may also include the color feature in addition to the SIFT feature, depending upon the $\delta(Z_n)$ value.

Particularly, if w_{ij} denotes flow field, warped saliency map W_{ij} of I_j for I_i is formed by $W_{ij}(p) = S_j(p + w_{ij}(p))$. In this manner, warped saliency maps of other images in the group are formed for every image in the group. These warped saliency maps are considered as candidate saliency maps comprising of candidate saliency values for each pixel in an image. Let $\mathbf{W}_i^k$ be set of all the candidate saliency maps for image $I_i \in Z_n$ at kth iteration including its own saliency map, and thus it is defined as

$$\mathbf{W}_i^k = \begin{cases} \{S_i^k, W_{ij}^k | I_j \in Z_n \backslash I_i\}, & \text{if } B_i = 0; \\ \{S_i^k\}, & \text{else,} \end{cases} \quad (11)$$

where the set consists of warped saliency maps in addition to the saliency map if break variable is not yet triggered. Hence, break variables become crucial in avoiding the costly warping processes when they are not required.

3) Saliency Fusion: Now that we have collected candidate saliency maps for I_i in the set $\mathbf{W}_i^k$, we can fuse them in any number of ways, such as average, geometric mean or median, etc. Also, we can make use of the quality scores as weights to improve the chances of fused saliency map to have better quality. Let $\mathbf{Q}_n^k = \{\phi(S_j^k)\psi(S_j^k) | I_j \in Z_n\}$ be set of quality scores of saliency maps of group Z_n at kth iteration. Let fusion function be denoted as $\mathcal{F}$ and we define fused saliency map of $I_i \in Z_n$ at kth iteration as

$$F_i^k = \mathcal{F}(\mathbf{W}_i^k, \mathbf{Q}_n^k) \quad (12)$$

where fusion function takes set of candidate saliency maps of the image and set of quality scores of its group as inputs.

In order to show the importance of fusing warped saliency maps of other similar images, and to choose appropriate fusion function, we compare Otsu thresholded saliency maps with ground truth masks on various co-segmentation datasets. We report overall precision (Pre.), recall (Rec.) and f-measure (FM) results in the Table II. It can be seen that fusing saliency maps greatly improves the performance over using original saliency

TABLE II
DIFFERENT SCORES, *viz* PRECISION (PRE.), RECALL(REC) AND F-MEASURE (FM), OBTAINED USING INITIAL SALIENCY MAPS AND USING OUR FUSED SALIENCY MAPS THROUGH VARIOUS FUSION FUNCTIONS ON DIFFERENT DATASETS

	MSRC			iCoseg			CosegRep			Weizmann Horses			Internet Images			Total		
	Pre.	Rec.	FM	Pre.	Rec.	FM	Pre.	Rec.	FM	Pre.	Rec.	FM	Pre.	Rec.	FM	Pre.	Rec.	FM
Initial Saliency	77.9	32.4	58.8	68.8	42.5	60.5	55.3	30.2	46.6	62.1	39.1	56.8	55.4	29.0	48.0	63.8	34.8	54.1
Average	80.9	**51.2**	75.3	67.5	51.4	63.4	71.3	**51.4**	67.2	75.0	57.7	73.2	73.0	**53.8**	**69.2**	73.1	**53.4**	69.1
Geometric Mean	**82.1**	50.3	**75.9**	**69.2**	50.6	**65.0**	**72.7**	49.4	**68.6**	**76.1**	56.4	**73.4**	**73.5**	52.4	69.0	**74.5**	51.6	**70.0**
Median	81.5	50.9	**75.9**	68.1	**51.8**	64.6	71.8	51.2	67.9	75.7	**58.6**	**73.4**	72.1	**53.8**	68.1	73.6	**53.4**	**70.0**
RMS	80.9	47.9	72.7	64.6	49.4	59.3	69.5	49.4	65.1	75.0	56.4	71.8	71.8	51.9	67.6	72.2	50.8	66.9
Harmonic Mean	82.4	45.0	75.0	68.4	45.6	63.4	72.9	43.9	67.4	76.1	51.8	72.5	73.2	48.2	68.1	74.5	46.8	69.1

It can be seen that fusion improves the performance.

maps on all the datasets. Also, average, geometric mean and median functions perform better than others. Eventually, we choose weighted median function in our experiments due to its robustness to the outliers. Moreover, the way median filtering is used for removing the salt and pepper noise in images inspires us to adopt median filter for application on the corresponding pixels across images (quite different from neighboring pixels in an image).

We use regularization to make saliency scores consistent within a superpixel region. Specifically, [50] is adopted to generate superpixels, and each pixel's saliency score is replaced with average saliency score of its superpixel.

D. Improving Efficiency

We have seen that above approach of interaction involves aligning each and every image w.r.t. each and every image in a group, but so many of computationally expensive alignments while considering large datasets is certainly undesirable. In order to overcome this, we modify our approach slightly.

Let's assume that dense correspondence is precise, which means that same sets of corresponding pixels will get together every time we try to collect them for different images in the group. In that case, collecting candidate saliency values at each pixel for each image becomes repetitive. Instead, an efficient way would be to collect candidate saliency values for one time only and propagate the fused result. Therefore, for every cluster, we choose the nearest image to the cluster center as the key image, say $\mathcal{I}_n$ in Z_n, for which alone we obtain corresponding pixels and calculate the fused saliency map first. Then this fused saliency map is aligned back to different group members to form candidate saliency maps for the group members (as shown in Fig. 8). Let $\mathcal{W}_{in}^k$ denote warped fused saliency map of $\mathcal{I}_n$ to image I_i at kth iteration. $\mathbf{W}_i^k$ can now be redefined in the following way:

$$\mathbf{W}_i^k = \begin{cases} \{S_i^k, W_{ij}^k | I_j \in Z_n \backslash I_i\}, \text{ if } I_i = \mathcal{I}_n \text{ and } B_i = 0; \\ \{S_i^k, \mathcal{W}_{in}^k\}, \text{ if } I_i \neq \mathcal{I}_n \text{ and } B_i = 0; \\ \{S_i^k\}, \text{ if } B_i = 1, \end{cases} \tag{13}$$

where the first case is for the key image, the second case is for member images, and the third case is when break variable is already triggered. And as far as fusion is concerned, it is performed in the following way:

$$F_i^k = \begin{cases} \mathcal{F}\left(\mathbf{W}_i^k, \mathbf{Q}_n^k\right), \text{ if } I_i = \mathcal{I}_n; \\ \mathcal{F}\left(\mathbf{W}_i^k, \left\{\phi(S_i^k)\psi(S_i^k), \sum_{I_j \in Z_n} \phi(S_j^k)\psi(S_j^k)\right\}\right), \text{ else,} \end{cases} \tag{14}$$

where fusion for the key image remains the same as earlier. But for member images, since it's fusion between saliency map and warped fused saliency map of the key image, we give high weight as much as the sum of all the quality scores in the group to the warped fused saliency map, because it's highly reliable for having been developed using multiple images.

Let's analyze how this modification affects the time-complexity. In the original method, in a group consisting of x images, every time we calculate fused saliency map for an image, saliency maps of all other $(x-1)$ images need to warp to this image, which requires computing the costly dense correspondences for $(x-1)$ times. Thus, the interaction of all x images will need computing dense correspondences for $x(x-1)$ times. This suggests $O(x^2)$ complexity, which is time-consuming and it is undesirable while dealing with large-scale datasets. But after the modification, we need to compute dense correspondence for $(x-1)$ times for the key image alone to generate its fused saliency map first, and then warp it back to member images, requiring computing dense correspondences for another $(x-1)$ times. Thus, in total it turns out to be only $2(x-1)$, suggesting $O(x)$ complexity. Thus, this modification resulted in reducing the time-complexity of proposed method from quadratic to linear.

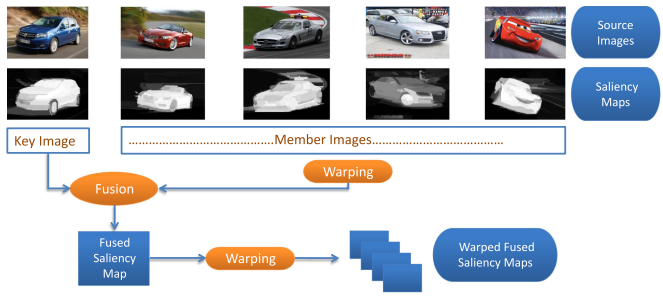

Fig. 8. We improve efficiency of our interaction process by collecting candidate saliency maps only for the key image and then aligning back the fused saliency map to other member images.

E. Applications

The obtained final high-quality saliency maps are utilized for object-level segmentation and localization in the following manner:

Segmentation: Based on the final saliency map S_i^*, we obtain the final object mask using GrabCut algorithm [45], in which foreground ($\mathcal{FG}_i$) and background ($\mathcal{BG}_i$) seed locations are determined by

$$p \in \begin{cases} \mathcal{FG}_i, & \text{if } S_i^*(p) > \tau; \\ \mathcal{BG}_i, & \text{if } S_i^*(p) < \lambda_i, \end{cases} \tag{15}$$

where pixel p will be considered as background seed location if its final saliency value is less than λ_i (Otsu's threshold [51] value of S_i^*). Similarly, pixel p will be considered as foreground seed location if its saliency value is greater than τ, which we call foreground threshold parameter. By default, we set τ as 0.75.

Localization: For localization, we first threshold final saliency map S_i^* with some threshold, say τ (same as in segmentation application above), and identify sparsely located spatial group (same as the object components such as $O_u(S_i^*)$) of white pixels (having saliency values greater than τ) as the candidate objects. Out of them, we only choose dominant objects that make at least half the contribution made by the largest object, i.e., $C_u(S_i^*) \geqslant 0.5 \times C_{u^*}(S_i^*)$. Such a criterion allows localization of multiple dominant objects if they are of somewhat similar size. By this, we also ensure that insignificant objects in the image that are present (may be due to complex backgrounds) are not considered for localization. Also, since these dominant objects may not be having similar edges as ones in the image, we identify nearest edge locations to the pixels in the concerned dominant object and adjust the bounding box to extreme edge locations in the four directions.

IV. Experiments

We conduct extensive experiments to evaluate our method in terms of the applications discussed in the previous section using [52] as initial saliency map. We use it because its code is publicly available and is considered as one of the state-of-the-art methods for salient object detection. Note that our framework is not limited to this particular saliency map; any other saliency map can also be used. Some sampled final saliency maps along with the initial saliency maps are presented in Fig. 9. Note that our model is able to retain the initial one as the final one when the initial one is already very good. In this section, we first provide details of different datasets and evaluation metrics used. Then we proceed with the evaluation.

A. Datasets

Several public co-segmentation and co-localization datasets are already available on which we can evaluate our final saliency maps.

In literature, most popularly used datasets for the co-segmentation evaluation are MSRC [53] and iCoseg [54] datasets. We evaluate on Coseg-Rep [41] and Internet images [18] datasets as well. MSRC dataset contains only

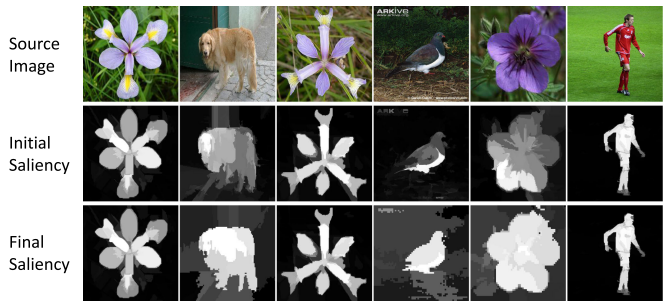

Fig. 9. Examples of some final saliency maps along with their initial saliency maps. Note that our model is able to retain the initial one as the final one when the initial one is already good.

14 categories with 419 images in total. iCoseg dataset contains 38 categories with 643 images in total. For fair comparison with the existing methods [18], [55], like them, we also use subset of the iCoseg dataset which includes 30 categories and a total of 530 images. Coseg-Rep dataset contains 23 categories and 572 images in total. Also, Internet images dataset released by [18] contains 3 categories: Airplane, Car, and Horse, with 4347, 6381 and 4542 images, respectively. All these datasets are small, and hence, suitable for our original interaction approach. For evaluating our efficient interaction approach for eventual segmentation in the large-scale scenario, ImageNet [56] setup of 0.5 million images in [16] is used.

Recently, for co-localization evaluation, [13] used tight bounding boxes across the ground truth segmentation masks of Internet images dataset as the ground truth bounding boxes. Following them, we also evaluate our method on the same setting. For the large-scale localization evaluation, we use ImageNet again in both unsupervised and supervised setups as suggested by previous methods, [13] and [17], respectively. Since our method can work in both supervised and unsupervised scenarios, for distinguishing between the unsupervised results and supervised results, suffixes (U) and (S) are used, respectively. As per the setup in [13], there are 1 million images for which bounding boxes are available in ImageNet, and they are spread over 3627 classes. In the supervised setup, [17] divides images with available ground truth bounding boxes into source sets (or training set) and target sets (or test set). For images that belong to source set, we replace saliency maps with ground-truth bounding boxes, and the task now is to obtain bounding boxes for remaining images in the group.

B. Evaluation

Following the literature [18], [55], we use Jaccard Similarity (Jacc.) and Accuracy (Acc.) for segmentation evaluation. Jaccard Similarity is defined as the intersection divided by the union of ground-truth and the segmentation result. Accuracy is defined as the percentage of correctly labeled pixels. Similarly, CorLoc score has been used for evaluation of localization which is defined as the percentage of images that satisfy the condition: $\frac{area(B_{gt} \cap B_{co})}{area(B_{gt} \cup B_{co})} > 0.5$, where B_{gt} and B_{co} are ground-truth and computed bounding boxes, respectively. Although there is no new parameter that gets introduced in our joint processing

TABLE III
COMPARISON ON COSEG-REP DATASET USING OVERALL VALUES OF JACCARD SIMILARITY (JACC.) AND ACCURACY (ACC.)

	Jacc.	Acc.
Co-segmentation&Co-sketch [41]	0.67	90.2
Ours (original)	0.73	91.9
Ours (efficient)	0.72	91.3
Ours (tuned τ/group)	**0.76**	**92.8**

TABLE IV
COMPARISON ON INTERNET IMAGE DATASET USING OVERALL VALUES OF JACCARD SIMILARITY (JACC.) AND ACCURACY (ACC.)

	Car		Horse		Airplane	
	Jacc.	Acc.	Jacc.	Acc.	Jacc.	Acc.
[34] (reported in [18])	0.37	58.7	0.30	63.8	0.15	49.2
[40] (reported in [18])	0.35	59.2	0.29	64.2	0.12	47.5
[18] (tuned threshold)	0.63	83.4	0.54	83.7	0.56	86.1
[57]	0.67	–	0.51	–	0.57	–
[26]	0.65	–	0.55	–	**0.62**	–
Ours (original)	0.71	87.0	0.57	84.7	0.55	85.7
Ours (efficient)	0.71	86.4	0.56	84.2	0.54	85.2
Ours (tuned τ/group)	**0.73**	**88.4**	**0.61**	**88.1**	0.59	**88.4**

TABLE V
COMPARISON ON MSRC AND ICOSEG DATASETS USING OVERALL VALUES OF JACCARD SIMILARITY (JACC.) AND ACCURACY (ACC.)

	MSRC		iCoseg	
	Jacc.	Acc.	Jacc.	Acc.
Initial Saliency	0.63	86.3	0.64	88.9
Discriminative [34] (reported in [18])	0.45	70.8	0.39	61.0
Multi-Class [40] (reported in [18])	0.51	73.6	0.43	70.2
Object Discovery [18] (tuned threshold)	0.68	87.7	0.69	89.8
Composition [55]	0.73	89.2	**0.73**	**92.8**
Clustering [1]	0.65	84.2	0.64	87.8
Ours (original)	0.72	88.9	0.67	89.3
Ours (efficient)	0.71	88.1	0.66	88.9
Ours (tuned τ/group)	**0.74**	**89.7**	0.72	91.8

approach, since other methods tune (or learn) their parameters for better performance, in addition to our original results, we also report results obtained by tuning the threshold parameter τ, which is basically the individual processing parameter.

C. Segmentation Evaluation

In Tables III–V, we compare our results of both original and efficient methods with state-of-the-art co-segmentation methods on different datasets. It can be seen that our methods obtain competitive performance compared to the existing methods. Although we base our model upon a simple co-saliency idea of fusing warped saliency maps, we still achieve better co-segmentation results compared to those obtained by the state-of-the-art co-saliency method [1]. Table V also gives comparison with the initial saliency map (seeds for GrabCut segmentation are provided by Otsu's thresholding) used in our method, which demonstrates the effectiveness of the joint processing

TABLE VI
COMPARISON ON IMAGENET DATASET USING OVERALL VALUES OF JACCARD SIMILARITY (JACC.) AND ACCURACY (ACC.)

Methods	Jacc.	Acc.
[42] (weights learnt)	–	77.3
[16] (weights learnt)	0.57	84.3
Ours (efficient)	0.56	84.1
Ours (tuned τ/group)	**0.59**	**86.4**

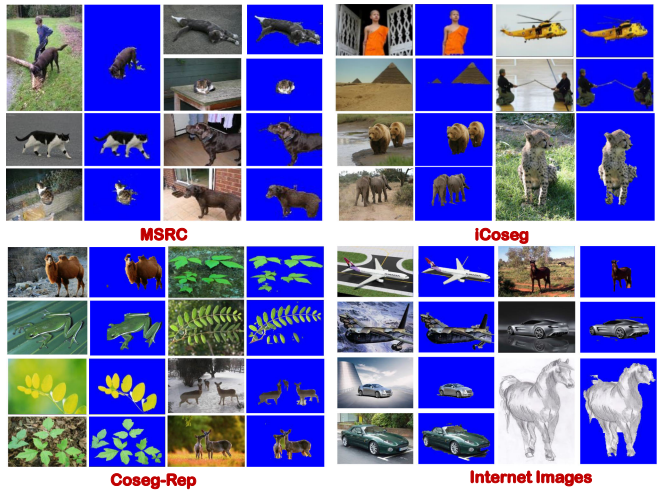

Fig. 10. Sample segmentation results from MSRC, iCoseg, Coseg-Rep, and Internet images datasets.

approach. Note that our method is much faster than state-of-the-art co-segmentation method [55]. Specifically, running on the same PC with Intel Core i5-3470@3.20 GHz CPU and 32 GB RAM, [55] (using their own source codes in Matlab) takes 29.2 hours to complete the entire segmentation process on MSRC dataset. However, our method (also in Matlab codes) takes only 4.9 hours. Also, we show some examples in supplementary material where our method performs better than [55]. Another thing to note here is that the performance difference is quite narrow between our original and efficient methods, which suggests that our assumption of warping process being precise is a viable assumption. Given this, we compare segmentation results of our efficient method with existing state-of-the-art results on ImageNet (large-scale dataset) in Table VI. We achieve comparable performance here as well. See Fig. 10 and supplementary material for sample segmentation results that we obtain on different datasets. It can be seen in these figures that proposed method is able to accurately segment both simple and complex images because we are able to guide the co-saliency estimation effectively using our saliency quality measurement.

D. Localization Evaluation

In this section, we discuss how we evaluate the proposed method for localization application on both unsupervised and supervised setups.

Unsupervised Setup: In the unsupervised setup, we compare our results with existing methods in Table VII on both

TABLE VII
CORLOC COMPARISON ON IMAGENET AND INTERNET IMAGES DATASETS IN UNSUPERVISED SETUP

	ImageNet	Internet
[18](U)	–	75.2 (tuned threshold)
[13](U)	53.2	76.6 (tweaked coefficients/group)
[47]	–	84.2
[26]	57.6	–
Ours (efficient)(U)	**64.6**	**84.5** (tuned τ/group)

TABLE VIII
CORLOC COMPARISON ON IMAGENET DATASET IN SUPERVISED SETUP

	CorLoc
[48](S)	58.5
[17](S)	66.5
[17]*(S)	68.3
Ours (efficient)(S)	**71.1**
Initial Saliency (U)	64.9
Ours (efficient)(U)	68.7

ImageNet and Internet images datasets. We achieve 21.8% and 12.1% improvements over [13] and [26], respectively, on ImageNet dataset. [47] also evaluates on Internet images dataset, and proposed method could marginally outperform [47] as well. However, the vital difference between [47] and proposed method is in terms of speed. Our method is much simpler and faster, and therefore it has large scale application as demonstrated on the ImageNet.

Supervised Setup: In the supervised setup, the problem that we try to address here is similar to the "Self" case in [17], where only images within the same class are used as source sets. In Table VIII, we compare our results on the target sets with two previous attempts in [48] and [17] to populate ImageNet with bounding boxes in a supervised manner. We achieve 21.4% and 6.9% improvement over [48] and [17], respectively. [17] also reports results using state-of-the-art features and object proposals, which we denote as [17]*. We achieve 4.1% improvement over state-of-the-art [17]* as well. Considering that the proposed method does not essentially need bounding boxes, unlike [17], we report unsupervised results obtained using the proposed method (Ours(efficient)(U)) and the initial saliency maps, where we do not use any ground-truth bounding boxes. Interestingly, we still obtain comparable results to [17]*(S). Again, the improvement over initial saliency demonstrates the effectiveness of the joint processing approach.

We show our localization results (red) along with the ground-truth (green) for visual comparison in Fig. 11. See supplementary material for more localization results. Thanks to our quality-guided approach to the joint processing, the proposed method is able to accurately provide bounding boxes as well for both simple and complex images.

V. DISCUSSION

ϵ-setting and δ-effectiveness: In a group Z_n, usage of the color feature depends upon $\delta(Z_n) < \epsilon$ criterion. In order to set

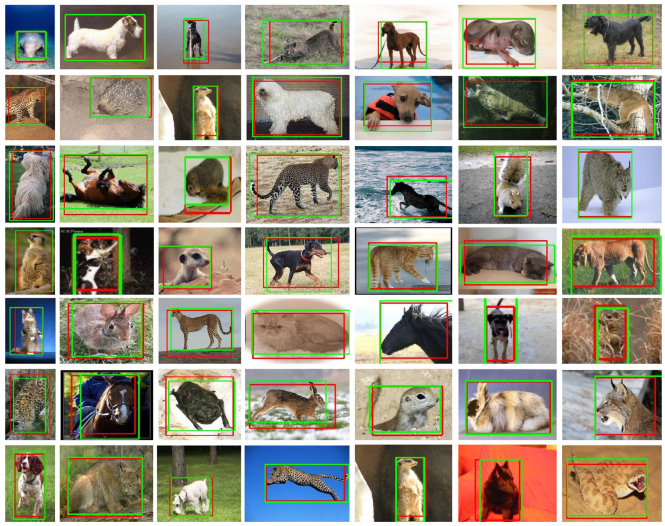

Fig. 11. Sample visual comparison between ground truth (green) and our results (red).

TABLE IX
AVERAGE δ AND TOTAL TIME TAKEN ON VARIOUS DATASETS FOR ORIGINAL AND EFFICIENT INTERACTION STRATEGY (1 ITERATION)

	δ	$\delta_{g.t.}$	Time (mins.) (Original)	Time (mins.) (Efficient)
MSRC	0.0026	0.0023	85	18
iCoseg	0.0017	0.0014	116	28
Coseg-Rep	0.0029	0.0029	186	39
Weizmann Horses	0.0039	0.0027	80	17
Internet Images	0.0028	0.0027	4306	901

the ϵ value properly, we show average δ values obtained for 5 datasets in Table IX, and only iCoseg dataset out of them requires color feature. As expected, a notable difference can be observed between iCoseg dataset and rest of the datasets in terms of their δ values. Noting this, we comfortably set $\epsilon = 0.0020$. We also show $\delta_{g.t}$ values where we make use of ground truth maps in the place of saliency maps. A high correlation of 0.796 between δ and $\delta_{g.t}$ suggests that our δ measurement is a good indicator.

Efficiency Comparison: It can also be observed in Table IX how the efficient strategy greatly reduces the time taken for interaction to $20\% - 25\%$ compared to the original strategy. Therefore, proposed modifications have certainly made large-scale application feasible while keeping the performance somewhat competitive (as we see in Tables III–V).

Limitations: The proposed method has some limitations. First, our separation measure only explores the saliency distribution, but not the spatial distribution. Fortunately, our concentration measure tries to cover this issue by exploring the foreground concentration, but partially for the lack of distance considerations. Second, our proposed method fails when our assumption (common object or its parts are salient in general, if not in every image) fails. Therefore, it heavily depends on the association of the image for our method to succeed. For example, only the beak portion of goose gets segmented or localized in Fig. 12(i), because other body parts are salient neither in the

Fig. 12. Our method has limitations in the following scenarios: (i) Wrong association, (ii) Difficult warping, and (iii) Multiple common objects.

considered image nor in the association. The third limitation is caused by poor warping process, i.e., when it struggles to align objects of very different sizes (a case that can arise due to the poor choice of clustering parameter, N, while grouping very few images). For example, high size variation in Fig. 12(ii) produces poor results. The fourth limitation is that our method may end up segmenting multiple object classes in some images, while ground truth masks may consist of only one object class. This can happen due to two reasons; one is that all images in a particular group (cluster) contain multiple common object classes, and another one is that saliency quality couldn't be improved by fusion against the (already) high-quality original saliency map. In such cases, our result may not match well with ground truth masks. Note how in Fig. 12(iii) our method captures multiple object classes. On the contrary, ground truth masks capture only one object class: plane, horse, windmill, or pyramid for those images.

VI. Conclusion

We propose a novel quality-guided fusion-based co-saliency estimation method, where saliency maps of different images are simply fused using dense correspondence technique. More importantly, this joint processing is guided by our proposed saliency quality measurement system, which helps us decide whether to choose the original or fused saliency map as the final one. The idea is to choose the saliency map with well-separated foreground and background, as well as a concentrated foreground. In this way, we attempt to address the individual versus joint processing issue. Our evaluation of final saliency maps w.r.t. segmentation and localization applications on several benchmark datasets including the large-scale dataset, Imagenet, show that proposed framework is able to achieve very competitive results.

Acknowledgment

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the Infocomm Media Development Authority, Singapore.

References

[1] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.
[2] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1896–1909, Dec. 2013.
[3] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.
[4] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: fundamentals, applications, and challenges," *ACM Trans. Intell. Syst. Technol.* vol. 9, no. 4, pp. 38:1–38:31, Jan. 2018.
[5] D. Zhang *et al.*, "A self-paced multiple-instance learning framework for co-saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 594–602.
[6] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1896–1909, Sep. 2016.
[7] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE Trans. Multimedia*, vol. 14, no. 5, pp. 1429–1441, Oct. 2012.
[8] Z. Tao, H. Liu, H. Fu, and Y. Fu, "Image cosegmentation via saliency-guided constrained clustering with cosine similarity," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4285–4291.
[9] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based RGBD image co-segmentation with mutex constraint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4428–4436.
[10] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, "Object-based multiple foreground video co-segmentation via multi-state selection graph," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3415–3424, Nov. 2015.
[11] F. Meng, J. Cai, and H. Li, "Cosegmentation of multiple image groups," *Comput. Vis. Image Understanding*, vol. 146, pp. 67–76, 2016.
[12] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *J. Visual Commun. Image Representation*, vol. 34, pp. 12–27, 2016.
[13] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2014, pp. 1464–1471.
[14] K. R. Jerripothula, J. Cai, and J. Yuan, "Cats: Co-saliency activated tracklet selection for video co-localization," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 187–202.
[15] K. R. Jerripothula, J. Cai, J. Lu, and J. Yuan, "Object co-skeletonization with co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3881–3889.
[16] M. Guillaumin, D. Kttel, and V. Ferrari, "Imagenet auto-annotation with segmentation propagation," *Int. J. Comput. Vis.*, vol. 110, no. 3, pp. 328–348, 2014.
[17] A. Vezhnevets and V. Ferrari, "Associative embeddings for large-scale knowledge transfer with self-assessment," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2014, pp. 1987–1994.
[18] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2013, pp. 1939–1946.
[19] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2011, pp. 2217–2224.
[20] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with Frank-Wolfe algorithm," in *Proc. Eur. Conf. Comput. Vis*. Springer, 2014, pp. 253–268.
[21] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1777–1784.
[22] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2011, pp. 409–416.
[23] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.
[24] L. Mai and F. Liu, "Comparing salient object detection results without ground truth," in *Proc. Eur. Conf. Comput. Vis*. Springer, 2014, pp. 76–91.
[25] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.
[26] S. Dutt Jain and K. Grauman, "Active image segmentation propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2864–2873.

[27] K. R. Jerripothula, J. Cai, and J. Yuan, "QCCE: Quality constrained co-saliency estimation for common object detection," in *Proc. IEEE Visual Commun. Image Process.*, 2015, pp. 1–4.
[28] K. R. Jerripothula, J. Cai, F. Meng, and J. Yuan, "Automatic image co-segmentation using geometric mean saliency," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 3282–3286.
[29] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *Proc. ACM Symp. User Interface Softw. Technol.*, 2010, pp. 219–228.
[30] H.-T. Chen, "Preattentive co-saliency detection," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 1117–1120.
[31] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2011, pp. 2129–2136.
[32] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. IEEE Comput. Vis. Pattern Recog.*, 2015, pp. 2994–3002.
[33] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2006, pp. 993–1000.
[34] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2010, pp. 1943–1950.
[35] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2009, pp. 2028–2035.
[36] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. Int. IEEE Conf. Comput. Vis.*, 2009, pp. 269–276.
[37] J. Yuan *et al.*, "Discovering thematic objects in image collections and videos," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2207–2219, Apr. 2012.
[38] G. Zhao and J. Yuan, "Mining and cropping common objects from images," in *Proc. ACM Multimedia*, 2010, pp. 975–978.
[39] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 169–176.
[40] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2012, pp. 542–549.
[41] J. Dai, Y. N. Wu, J. Zhou, and S.-C. Zhu, "Cosegmentation and cosketch by unsupervised learning," in *Proc. IEEE Int.Conf. Comput. Vis.*, 2013, pp. 1305–1312.
[42] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in imagenet," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 459–473.
[43] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[44] Y. Chai, V. Lempitsky, and A. Zisserman, "Bicos: A bi-level co-segmentation method for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2579–2586.
[45] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *Trans. Graph.*, vol. 23, no. 3., 2004, pp. 309–314.
[46] K. R. Jerripothula, J. Cai, and J. Yuan, "Group saliency propagation for large scale and quick image co-segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 4639–4643.
[47] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2015, pp. 1201–1210.
[48] M. Guillaumin and V. Ferrari, "Large-scale knowledge transfer for object localization in imagenet," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2012, pp. 3202–3209.
[49] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[50] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Proc. Eur. Conf. Comput. Vis*., Springer, 2008, pp. 705–718.
[51] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285–296, pp. 23–27, 1975.
[52] H. Jiang *et al.*, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2013, pp. 2083–2090.
[53] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2006, pp. 1–15.
[54] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "ICOSEG: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2010, pp. 3169–3176.
[55] A. Faktor and M. Irani, "Co-segmentation by composition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1297–1304.
[56] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[57] H. Cholakkal, J. Johnson, and D. Rajan, "Backtracking SCSPM image classifier for weakly supervised top-down saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5278–5287.

Koteswar Rao Jerripothula (S'15–M'17) received the B.Tech. degree from IIT Roorkee, Roorkee, India, in 2012, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2017. He interned with Visual Modeling & Analytics Team in ADSC, Singapore, during July and December 2016. He is a faculty with the Graphic Era University, Dehradun, India. His research interests include computer vision and multimedia. Dr. Jerripothula was the recipient of the "Top 10% paper" award at ICIP 2014 and published in top venues like CVPR, ECCV, TMM, etc.

Jianfei Cai (S'98–M'02–SM'07) received the Ph.D. degree from the University of Missouri, Columbia, MO, USA. He is a faculty with the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. He has served as the Head of Visual & Interactive Computing Division and the Head of Computer Communication Division with NTU. He has published more than 200 technical papers in international conferences and journals. His major research interests include computer vision, multimedia, and machine learning. Dr. Cai has served as the Leading Technical Program Chair for the IEEE International Conference on Multimedia & Expo 2012 and the Leading General Chair for Pacific-rim Conference on Multimedia 2012, as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING during 2013 and 2017, and for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY during 2007 and 2013. He is currently an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA. He also serves as the Chair for the IEEE CAS Visual Signal Processing and Communication Technical Committee. He is a Corecipient of the Paper Awards in ACCV, IEEE ICIP and MMSP.

Junsong Yuan (M'08–SM'14) received the Graduate degree from the Special Class for the Gifted Young of Huazhong University of Science and Technology (HUST), Wuhan, China, in 2002, the Ph.D. degree from Northwestern University, Evanston, IL, USA, and the M.Eng. degree from the National University of Singapore, Singapore. He is currently an Associate Professor with the School of Electrical and Electronics Engineering, Nanyang Technological University (NTU), Singapore. His research interests include computer vision, video analytics, gesture and action analysis, and large-scale visual search and mining. Prof. Yuan served as the Guest Editor for the *International Journal of Computer Vision*. He is currently a Senior Area Editor for the *Journal of Visual Communications and Image Representations*, Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is the Program Co-Chair of ICME18 and Area Chair of ACM MM, CVPR, ICIP, ICPR, ACCV, etc. He was the recipient of the Best Paper Award from the International Conference on Advanced Robotics, 2016 Best Paper Award from IEEE TRANSACTIONS ON MULTIMEDIA, Doctoral Spotlight Award from IEEE Conference on Computer Vision and Pattern Recognition, Nanyang Assistant Professorship from NTU, and Outstanding EECS Ph.D. Thesis award from Northwestern University.