Rethinking of the Image Salient Object Detection: Object-level Semantic Saliency Re-ranking First, Pixel-wise Saliency Refinement Latter

Guangxiao Ma[1] Shuai Li[1] Chenglizhao Chen[1,2*] Aimin Hao[1] Hong Qin[3]
[1]Qingdao Research Institute & State Key Laboratory of VRTS, Beihang University
[2]Qingdao University [3]Stony Brook University
Code & Data, https://github.com/gxma/TIP_RISOD

***Abstract*—The real human attention is an interactive activity between our visual system and our brain, using both low-level visual stimulus and high-level semantic information. Previous image salient object detection (SOD) work conducts its saliency predictions in a multi-task way, which performs pixel-wise saliency regression and segmentation-like saliency refinement at the same time. However, this multi-task mythology has one critical limitation, where the semantical information embedded in feature backbones might be degenerated during the training process. However, our visual attention is mainly determined by the semantical information, which can be evidenced by the fact that, in general, we tend to pay more attention to semantically salient regions even these regions are not the most perceptually salient ones at the first glance. This fact clearly contradicts with the widely-used multi-task methodology mentioned above. To improve, this paper divides the SOD problem into two sequential steps. We firstly devise a lightweight, weakly supervised deep network to coarsely locate those semantically salient regions. Next, as a post-processing refinement, we selectively fuse multiple off-the-shelf deep models on those semantically salient regions determined by the previous step to formulate the pixel-wise saliency map. Compared with the state-of-the-art (SOTA) models focusing on learning pixel-wise saliency in single image using perceptual clues solely, our method aims to investigate the object-level semantic ranks between multiple images, of which the methodology is more consistent with the real human attention mechanism. Our method is simple yet effective, which is the first attempt to consider the salient object detection mainly as an object-level semantic re-ranking problem.**

***Index Terms*—Image Salient Object Detection, Weakly Supervised Learning, Object-level Ranking.**

Fig. 1: The conventional methods consider the image saliency mainly from the perceptual perspective, while the real human vision system may pay more attention to those semantically salient objects.

I. Introduction

SALIENT object detection aims to fast locate the most eye-catching objects in a given scene. As an early perceptual tool to support the high-level recognition activities of human brain, a desired **s**alient **o**bject **d**etection (SOD) method should be lightweight designed [1], [2], [3], [4], [5], [6]. After entering the deep learning era, the **s**tate-**o**f-**t**he-**a**rt (SOTA) SOD models have achieved high-quality detection in an extremely fast end-to-end manner.

Generally speaking, a SOD deep model [7], [8], [9], [10] usually consists of an encoder, which can be any off-the-shelf feature backbone (e.g., the vanilla VGG16 or ResNet50), followed by a decoder with elaborate network design. The feature backbones are usually pre-trained via large-scale training sets with abundant semantic information (e.g., object categories). Thus, these backbones are semantical-aware in essence, which are able to span high discriminative feature spaces for the subsequent decoder to performance saliency estimation. However, the current SOTA models have followed the multi-task methodology, which performs the pixel-wise saliency regression and the segmentation-like saliency refinement at the same time, where the training process easily degenerates their backbones in providing semantical information.

In fact, the real human attention mechanism [11] is not solely dependent on the low-level visual stimulus provided by our visual system. On the contrary, the high-level semantic information rooted in our brain plays an extremely important role to determine where we really look at. For example, as shown in Fig. 1-A, the salient *road sign* in the image-1 will become non-salient when a person with a dog appears in the image-2. In such case, though the perceptual aspects of

Corresponding author: Chenglizhao Chen, cclz123@163.com.

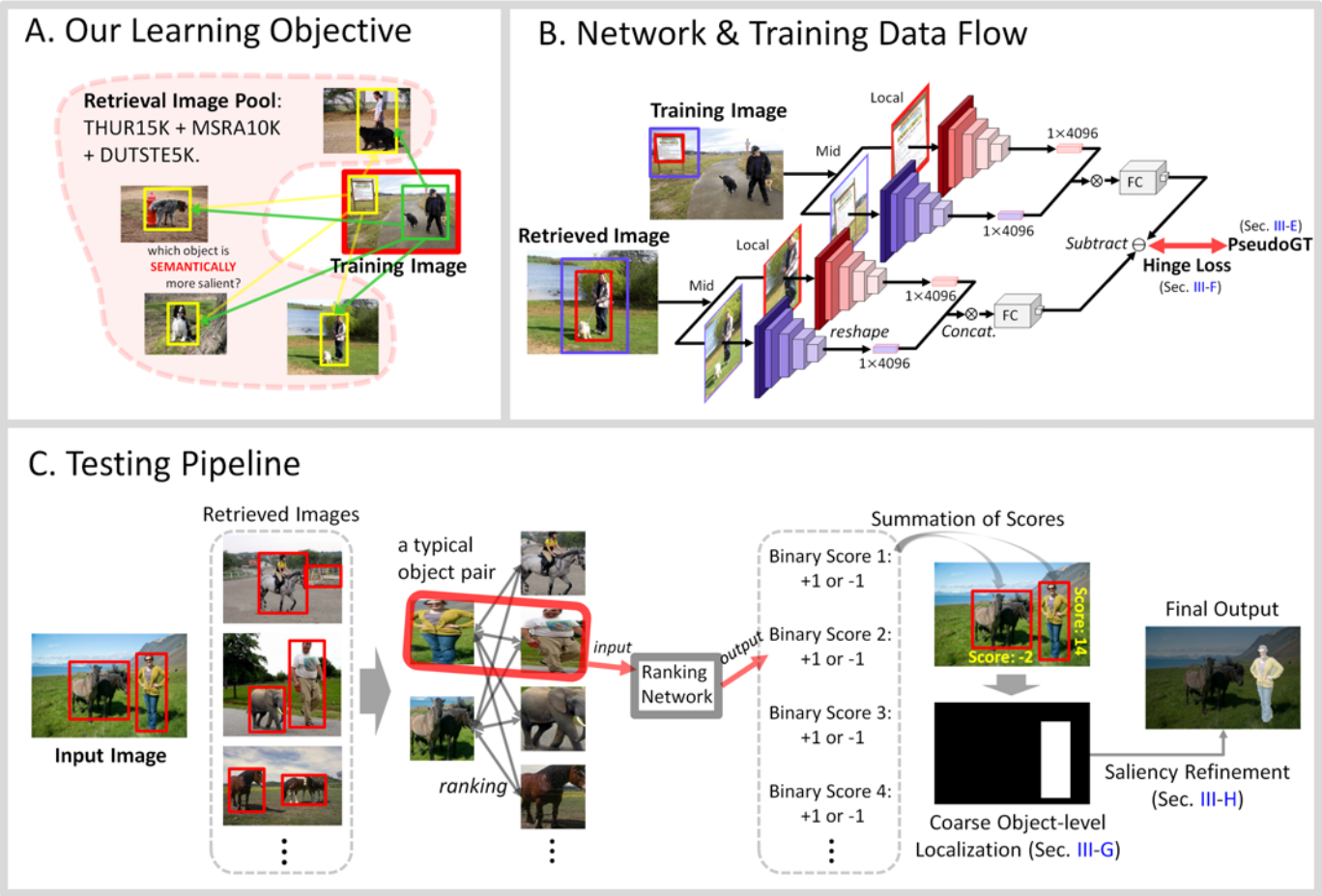

Fig. 2: Method Overview. The sub-figure A demonstrates our learning objectives; because the high-level semantical information rooted in our brain plays an extremely important role to determine where we look at, we aim to learn the semantical ranks for each object pair; the sub-figure B shows the network details, where our network is lightweight designed and weakly supervised with fixed backbones; the sub-figure C shows the testing process.

the *road sign*, e.g., appearance, location and contrast, are not even changed, we may still tend to pay more attention to those regions which are semantically more salient, i.e., the person and the dog in image-2; however, the conventional methods [12], [13] may still assign large saliency value to the *road sign* incorrectly (see left column in Fig. 1-B). To solve this problem, this paper proposes a novel SOD learning framework, of which the major highlight is its capability of emphasizing on the use of "semantic saliency", which is not the saliency as firstly defined by Itti et al. in [14]. For example, as can be seen in Fig. 1-C and D, the red phone booth might be the most salient object if we follow the conventional definition of saliency (i.e., the perceptual saliency). However, from the perspective of the SOD related applications such as image compression, treating the phone booth as the salient object might be questionable, and it may be more useful to detect the BigBen as the salient object because the BigBen is more meaningful than the phone booth in this scene. Thus, we aim to emphasize the importance of the semantic saliency, which may be more helpful than the perceptual saliency in practice.

Moreover, compared with the conventional methods [15] focusing on learning perceptual saliency via fully supervised manner, our learning scheme learns to rank different objects according to their semantical saliency in a weakly supervised fashion (see Fig. 2-A). We show the object-level semantical saliency ranking network overview in Fig. 2-B, where the proposed network consists of two sequential parts, i.e., a feature backbone followed by a lightweight binary classifier. This network simultaneously takes two resized rectangular object proposals as input, where the feature backbones will compute the semantic deep features for each of these objects. The learning objective is also simple and intuitive—to make binary decision on which object in the given object pair is semantically more salient. Notice that if two objects in the given object pair share similar semantic information, we will resort the conventional object-level perceptual contrast (i.e., {Local + Mid}, Fig. 2-B) to avoid the learning ambiguity.

Once the above mentioned classifier has been trained, we directly use it to conduct inter-object semantic saliency ranking, aiming to coarsely locate salient objects. The detailed testing dataflow can be found in Fig. 2-C, where, as a post-processing procedure, we selectively fuse multiple off-the-shelf deep saliency models to achieve the pixel-wise saliency refinement for the final saliency.

It also should be noted that the conventional methods learning perceptual saliency and conducting pixel-wise refinement at the same time easily lead to an extremely large problem domain, thus their feature backbones need to be completely re-trained/finetunned on SOD training set to ensure the learning convergency, yet such re-training/finetunning procedure easily degenerates their feature backbones in providing semantic information. In sharp contrast, the problem domain of our semantic saliency learning is rather simple, which focuses on learning how to rank two objects according to their semantic saliency degrees; as a result, the feature backbone in our model can be fixed during model training, thus it can well retain strong semantic information, producing high-quality SOD eventually.

II. Related Work

A. Conventional SOD Methods

The conventional methods using handcrafted features can be categorized into two groups: bottom-up and top-down [16]. Due to the space limitation, we only list several most representative ones here.

Bottom-up models are mainly based on the center-surround prior using linear or non-linear combinations of low-level visual attributes (such as color, intensity, texture) to calculate various low-level saliency cues. According to the spatial scope of saliency calculation, these methods can be further divided into local methods and global methods. Local methods measure the saliency by considering the contrast between each pixel or image area. Itti et al. [14] calculated color and orientation contrasts at multiple scales to measure local saliency. Although it can identify prominent pixels, the results are usually blurry and contain a lot of false detections. Harel et al. [17] created feature maps using graph-based random walks for normalization. Since these methods only consider local contrast, they tend to detect the high-frequency features only (such as edges or noises) and suppress homogeneous areas inside salient objects simply. To further improve, Achanta et al. [18] proposed a method to estimate the saliency by considering the contrast of the entire image, which can directly estimate the pixel-wise saliency by calculating the color difference of the global color histogram. Similarly, Cheng et al. [19] took color histograms as the regional features and computed saliency on the basis of histogram dissimilarity. Yan et al. [20] proposed a hierarchical framework to address small-scale high-contrast patterns. Liu et al. [21] proposed an efficient regional histogram based computational model for saliency detection in natural images. First, the global histogram is constructed by performing an adaptive color quantization on the original image. Then, the pixel-level saliency map is generated by integrating the color spatial similarity measures with the distinctiveness and compactness measures. Olivier [22] proposed a computational model of visual attention using visual inferences. The dominant depth and the horizon line position are inferred from low-level visual features. This prior knowledge helps to find salient areas on still color images.

Top-down approaches usually require the integration of advanced knowledge in the calculation process of saliency cues, such as objectivity and object detectors. Borji [23] integrated both bottom-up and top-down features when learning their saliency models, which also took both person and car detectors as the high-level priors. Jia et al. [24] computed a high-level saliency prior using objectness, where a Gaussian MRF has been applied to enforce the consistency among salient regions. Léo [25] proposed a new unsupervised paradigm to compute motion saliency cues. The key ingredient is the flow in painting stage, where candidate regions are determined from the optical flow boundaries. The residual flow in these regions is given by the difference between the optical flow and the flow in painted from the surrounding areas. It provides the cue for motion saliency. The method is flexible and general by relying on motion information only. Olivier et al. [26] proposed a novel way to ensure high discrepancy between focal and ambient saliency maps, of which the key is to devise an automatic method for inferring the focalness degree. This method opens new avenues for the computational modelling of saliency models.

B. Deep Learning SOD Models

Recently, convolutional neural networks have achieved many successes in visual recognition tasks, including video salient object detection [27], [28], [29], [30], RGB-D salient object detection [31], [32], [33], [34], [35], image classification [36], object detection [37], and scene parsing [38]. Long et al. [39] proposed a fully convolutional network to predict pixel-wise saliency label. Hou et al. [13] proposed to use short connections to integrate the high-level semantic information, embedded in deep layers, into the shallower layers to take full use of the multi-scale deep features, improving the overall detection performance significantly. Then, to further make full use of multi-scale deep features, Zhang et al. [40] combined multi-level resolution feature maps by bidirectional aggregation, which integrated the multi-scale information in both the top-down and the bottom-up ways.

Many deep learning based models targeted the feature integration part. Wang et al. [41] proposed a pyramid pool module and a multi-stage refinement mechanism to collect context information and stage-wise results. Luo et al. [42] proposed a simplified convolutional neural network that combines global information and local information through a multi-resolution grid structure. Zhang et al. [43] utilized the deep uncertain convolutional features and proposed a reformulated dropout after specific convolutional layers to construct an uncertain ensemble of internal feature units. Li et al. [44] proposed a complementary perceptual network that combines cross-over models and cross-layer features to solve the saliency detection task of depth information. Albeit effective, these approaches do not typically consider the object-level semantic saliency information, and most of them focus on learning the perceptual saliency with a fully supervised method.

Our method differs greatly from models [45], [45], [46], [47] mentioned above, and our model is a weakly supervised object-level semantic saliency detection scheme. Moreover, our approach is also totally different to the co-saliency detection approaches [48], [49], [50], [51], [52], [53]. Co-saliency

detection aims at detecting the common perceptual saliency consistency from multiple images, while our approach mainly resorts the semantic saliency to coarsely locate those salient objects.

C. SOD via Siamese Network

Siamese network [54] is a network with two identical network branches and a loss module. The two branches share weights during training. Pairs of images and labels are the input of the network, yielding two outputs which are passed to the loss module. The gradients of the loss function with respect to all model parameters are computed by back propagation and updated with the stochastic gradient method. We shall list several most representative models which have adopted the siamese network.

Zhou et al. [55] proposed a novel multi-stage siamese network, where the siamese network is built to aggregate low-level and high-level features, and parallel estimate both edge saliency and region saliency. The predicted regions become more accurate after being enhanced by edges' responses, and the predicted edges can also be improved by suppressing the false positives in background. Lu et al. [56] introduced a novel network, called co-attention siamese network, to address the unsupervised video object segmentation task from a holistic view. The co-attention layers in this model provide efficient and competent stages for capturing global correlations and scene contexts by jointly computing and appending co-attention responses into a joint feature space. Ji et al [57] presented a novel cross-attention based encoder-decoder model, where the siamese framework is applied as the blender for the video salient object detection task. Different from all works mentioned above, Fu et al. [58] designed a novel joint learning and densely cooperative fusion siamese architecture to learn from both RGB and deep channels.

In a word, the siamese network adopted in our salient object detection community is mainly aimed at exploiting saliency-aware cross-modal commonality and complementarity instead of matching or measuring distance.

III. The Proposed Method

Our method follows a coarse-to-fine manner, where the object-level semantic ranking part aims to coarsely locate those semantically salient regions via rectangular boxes, and we shall introduce this part in advance as follows.

A. Object Proposal Preliminary

Since our network attempts to learn the "object-level" semantic saliency, we use the off-the-shelf GARPN [59] to decompose the given image into 10 rectangular object proposals (it can be more than 10 proposals, but we empirically choose 10 for efficiency). In order to filter those overlapped object proposals, we compute the **i**ntersection **o**ver **u**nion (IOU) rate [60] between each 2 out of 10 object proposals and then exclude those overlapped object proposals with low confidence degree (a probability to measure how well a proposal warps an object, and it is provided by GARPN directly). For example, as can be seen in Fig. 3, the object proposals A and B in IMG1 are overlapped partially (i.e., IOU rate>0.4), and only the object proposal B will be retained because its confidence rate is higher than that of the object proposal A. Meanwhile, if an extremely large object proposal has fully wrapped a smaller one completely, we will drop the smaller one to keep high Recall rate when localizing salient objects, where the decreased Precision rate would be compensated via the pixel-wise refinement.

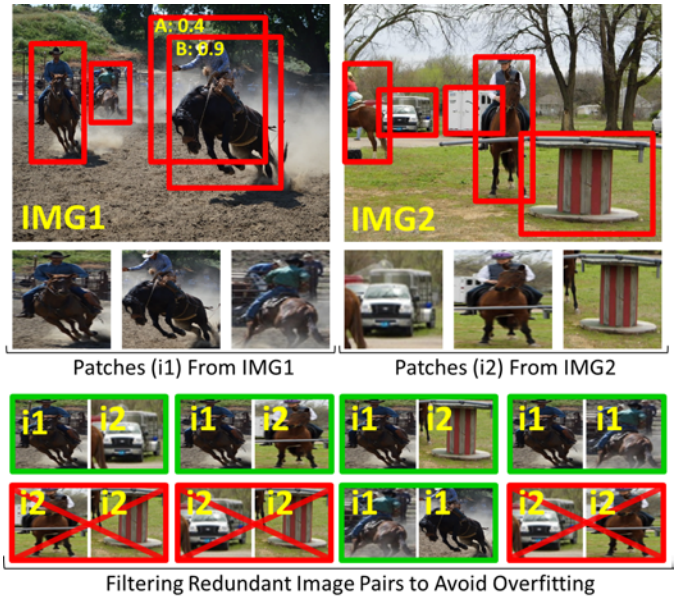

Fig. 3: Pictorial illustration regarding how object pairs are generated. We use "A: 0.4" to denote that the confidence degree of the object proposal A is 0.4; IMG1: a training image; IMG2: one of the corresponding retrieved images; red "×" in bottom row: object proposal pairs needed to be filtered.

To avoid overfitting, we also filter those object proposal pairs whose two object proposals are all obtained from the retrieved image (or other retrieved images), because there will be massive duplicate object pairs if we do not. For example, as shown in Fig. 3, for an image IMG1 and its retrieved image IMG2, we only consider the $\{i1, i2\}$ or $\{i1, i1\}$ cases, and all $\{i2, i2\}$ cases will be filtered, where $i1$ or $i2$ presents a proposal belonging to IMG1 or IMG2.

By using the steps mentioned above, each object in the given image can be efficiently and accurately warped by a single rectangle box tightly, and the rectangular proposal will be resized to a fixed size (224×224) to fit the adopted feature backbone (VGG16).

B. Object-level Semantic Deep Feature

The overall network architecture of our method is quite simple, which simultaneously takes 2 resized rectangular object proposals as input. For each input object proposal with size $w \times h$, we use the off-the-shelf backbone network (VGG16) to compute its high dimensional semantic deep feature (4096).

To ensure a high discriminative deep feature space, for each object proposal (P), we further consider its enlarged version

($\hat{P}$) with size $\{1.5 \times w\} \times \{1.5 \times h\}$. Both P and $\hat{P}$ are fed into the backbone network ($Backbone$) to obtain semantic-aware deep features (4096 dimensional respectively), which are later combined as the {4096+4096} multi-scale (Local + Mid, Fig. 2-B) perceptual contrast (Eq. 1).

$$f \leftarrow Backbone(P) \otimes Backbone(\hat{P}). \quad (1)$$

C. Network Architecture

Since our network aims to coarsely locate those semantically salient regions, there are two feasible choices to design our semantic sub-net:
(I) a typical choice is to directly feed the object-level semantic deep features into a full connected sub-net (we have tested the FC6) to make a binary prediction on if the input object proposal is semantically salient or not;
(II) alternatively, we can follow the siamese structure to receive two semantic deep features (from different object proposals) at each time, aiming to rank the inter-object semantic saliency to predict which one is semantically more salient.

In fact, it is quite ubiquitous that two semantically similar objects have totally different perceptual saliency, and this fact makes the aforementioned choice (I) to face learning ambiguities frequently. Thus, the learning objective of choice (I) seems to be problematically formulated. In sharp contrast, the choice (II) has reformulated the semantic saliency learning problem to a semantic saliency re-ranking problem, and thus it is more reasonable. Also, the learning objective of choice (II) is quite simple and intuitive, which aims to enlarge the semantic deep feature distance between two objects even in the case that these two objects are semantically similar. Moreover, the quantitative comparisons (Table. II) between choice (I) and (II) also suggest adopting the latter option. Therefore, we implement our semantic sub-net by using the choice (II), and the network architecture overview can be found in Fig. 2-B.

D. Object-level Training Data

Since our method aims to rank the inter-object semantic saliency, the learning scope should comprise the following two aspects: 1) semantically "different" object pairs; 2) semantically "similar" object pairs.

We formulate the training set to include both intra-image and inter-image object pairs (see Fig. 2-A and Fig. 3). By using the aforementioned object proposals extraction scheme (Sec. III-A), we obtain multiple non-overlapped object proposals from each image in our retrieval image pool (THUR15K[61]+MSRA10K[19]+DUTSTE5K[62]). For each two objects in an identical image, it will be composed as an intra-image training instance. Meanwhile, for each image in our training set DUTSTR10K, we use it to retrieve K similar images from the retrieval image pool as the retrieved sub-group images, where two object proposals, one from the training image and the other from the retrieved sub-group images, will be composed as an inter-image training instance.

Also, for images with either similar semantical information or similar overall scene layout, their object-level semantic categories tend to reach stable co-occurrence status, e.g., *car* frequently accompanies with *road*. Thus, we adopt the hybrid image retrieval scheme (see Fig. 4), i.e., the retrieved sub-group K images are composed by top-K_s semantically similar images (measured by image-level semantical deep features provided by VGG16) and another top-K_l images with similar scenes (measured by GIST [63] & HOG [64] feature distance), where $K = K_s + K_l$. In our implementation, we assign $K = 5$, $K_s = 2$, $K_l = 3$ as the optimal choice according to our component evaluation.

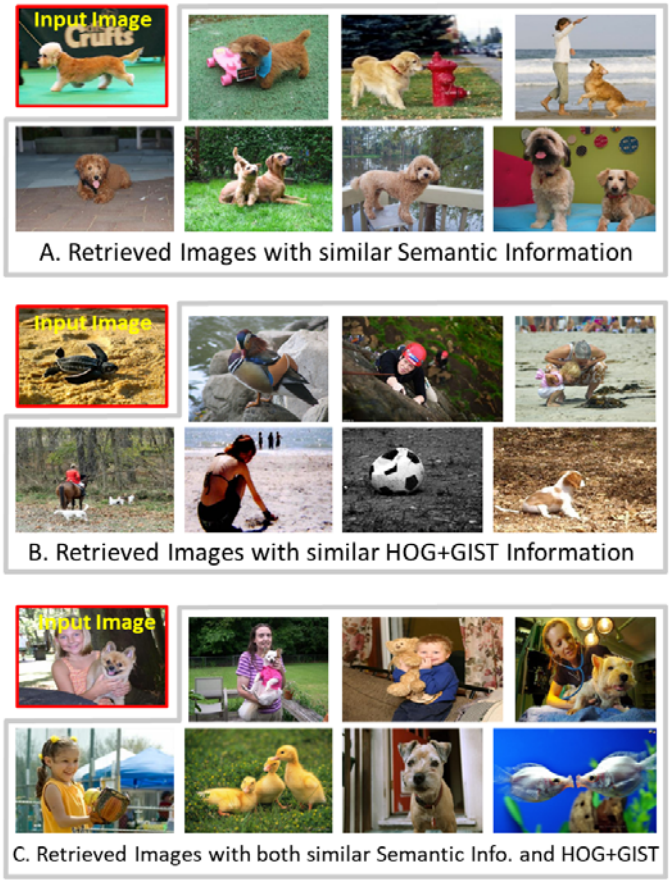

Fig. 4: Illustration of different retrieval schemes. The hybrid image retrieval scheme (C) ensures a stable co-occurrence state for the object semantic categories of the retrieved images, which is clearly more suitable for our object-level semantic saliency re-ranking task.

E. Semantic Saliency Pseudo-GT

In previous subsections, we have converted the original training set into object pairs. For each object pair ($\{P_1, P_2\}$), our network aims to make binary decision on whether the object P_1 is semantically more salient than its competitor P_2.

To achieve this goal, for each object P in the training set (DUTSTR10K), we assign its perceptual saliency label (PSL) = 1 if the rectangular overlapping rate between this object and its corresponding perceptual saliency GT map (provided by the training set) $> 70\%$, otherwise, we assign $PSL = -1$.

Thus, for an object pair ($\{P_1, P_2\}$) (both objects belong to the training set DUTSTR10K) with different perceptual

saliency labels, we formulate its semantic saliency pseudo-GT (pGT) as Eq. 2.

$$pGT_{\{P_1,P_2\}} = \begin{cases} 1 & if\ PSL_{P_1} > PSL_{P_2} \\ -1 & if\ PSL_{P_1} < PSL_{P_2} \end{cases}. \quad (2)$$

For cases with object pair ($\{P_1, P_2\}$) with identical perceptual saliency labels (both objects belong to the training set DUTSTE10K) or one of its objects belongs to the retrieval image pool, we use five off-the-shelf deep saliency models, which are all pre-trained using the same training set as our method (i.e., the DUTSTE10K only), to formulate the semantic saliency pseudo-GT (pGT) as Eq. 3.

$$pGT_{\{P_1,P_2\}} = \begin{cases} 1 & if\ PSL5_{P_1} > PSL5_{P_2} \\ -1 & otherwise \end{cases}, \quad (3)$$

where $PSL5_{P_1} = ||pat(S, P_1)||_1/(w_1 \times h_1)$; w_1, h_1 respectively denote the width and height of the object proposal P_1; the $S = \sum Sal_i, i \in \{1, ..., 5\}$, where Sal_i denotes the saliency map predicted by the i-th pre-trained saliency model; $pat(S, P_1)$ crops a patch from S, where the cropping position is provided by the object proposal P_1; $||\cdot||_1$ denotes the L_1-norm.

Our rationale is to assign a binary GT for each object pair no matter whether its two objects are semantically similar or not, because an object with higher perceptual saliency value (predicted by the five pre-trained deep models) could have large potential to be more salient in its semantic aspect.

F. Loss Function

We represent the object-level training set as $X = \{f_1, f_2\}, Y = pGT$, where X and Y respectively denote the training instance (i.e., object pair deep feature) and its binary label (Eq. 2 and Eq. 3). Since the binary label pGT is formulated in a weakly supervised manner, it may occasionally conflict with the real semantic rank between f_1 and f_2. Thus, we choose to use the hinge loss (Eq. 4).

$$L(f_1, f_2, \theta) = max\Big\{0, pGT \cdot \big(FC_2(f_2) - FC_1(f_1)\big) - \rho\Big\}, \quad (4)$$

where θ denotes the network parameters, FC_1 and FC_2 are two full connection sub branches (with size 8192-1024-2048-2048-1024-1024-2) in our siamese network (see Fig. 2); ρ is the hinge loss margin (we empirically set it to 10) that vanishes the gradient to alleviate the learning ambiguity in the case of two objects with almost the same semantic saliency. The hinge loss is able to focus the back-propagated gradients on the semantic saliency ranking problem, which is more suitable than the widely used entropy loss, and the quantitative evidence can be found in Table. IV.

G. Semantically Salient Object Localization

Given a testing image I, we aim to coarsely locate all those objects that are supposed to be semantically salient. As such, for each object (P_i) in I, we define its semantic saliency $Score$ as Eq. 5.

$$Score_i = \sum_{P_j \in I^+ or\ I} \Big|\Big| \max\big\{0, FC_1(f_i) - FC_2(f_j)\big\} \Big|\Big|_0, \quad (5)$$

where $||\cdot||_0$ is the L_0-norm; f_i denotes the deep feature of the i-th object in image I; FC_1 and FC_2 are the learned sub branches (shared weights) in our siamese network, which are identical to Eq. 4; I^+ denotes the retrieved sub group images with K similar images to I (Sec. III-D).

By using Eq. 5, those semantically salient objects in I will be assigned with large $Score$. However, we do not know the exact number of how many objects in I are the semantically salient ones, which hinders us to directly use $Score$ value for the localization of semantically salient objects.

Thus, we select the top-q objects with the largest $Score$s in I as the semantically salient ones, where the value of q is adaptively determined by Eq. 6.

$$q = arg\max_i \{\frac{\partial \xi}{\partial i}\}, \quad \xi \leftarrow des\Big([Score_1, Score_2, ...]\Big), \quad (6)$$

where ∂ denotes the partial derivative; $des(\cdot)$ ranks its input according to the $Score$ values in descending order. The rationale of Eq. 6 is that those semantically salient objects in I should simultaneously have the following attributes: 1) with an extremely large $Score$; 2) and its $Score$ value should also be significantly larger than those of the rest.

The salient objects determined by our scheme are usually more accurate than that determined by the conventional models (see T able. V). Moreover, our method can allow "multiple" objects to go the saliency refinements if their semantic scores are similarly high (see Eq. 5 and Eq. 6). Thus, our method can perform well for datasets with multiple salient objects belonging to different categories.

H. Pixel-wise Saliency Refinement

Thus far, we have coarsely located q semantically salient objects in the given image I (see the white rectangle in Fig. 2-C, and we denote it as I_c).

To complete I_c with more spatial details, we use five off-the-shelf deep saliency models, all trained on the same training set as our method, to perform the pixel-wise saliency refinement as a post-processing procedure. We use $Sal_i, i \in \{1, 2, 3, 4, 5\}$ to represent their saliency predictions.

Generally speaking, the saliency predictions respectively predicted by these five off-the-shelf models might vary from each other, thus we should not directly fuse these five saliency maps into the I_c, which easily lead to suboptimal result. Since I_c has already told us the coarse location of the salient objects, we use the overlap rate between these I_c and Sal_i to present the model-level confidence ($Conf$) that reflects the quality of its saliency map. Thus, the quality of the i-th deep saliency model for the j-th object box (P_j) in I can be formulated as Eq. 7:

$$Conf(i, j) = \underbrace{\Big|\Big|\frac{pat(Sal_i, P_j)}{pat(I_c, P_j)}\Big|\Big|_1}_{local\ similarity} + \lambda \cdot \underbrace{\Big|\Big|\frac{Sal_i \odot I_c + C}{Sal_i + I_c + C}\Big|\Big|_1}_{global\ similarity}, \quad (7)$$

where C is a small constant value to avoid division by zero; $\odot$ denotes the element-wise Hadamard product; function $pat(Sal_i, P_j)$ uses P_j to crop a patch from Sal_i; λ is a balance factor between local and global similarity, and we empirically assign it to 0.5.

The final saliency map ($FinalS$) can be computed by adaptively fusing all these saliency maps by Eq. 8.

$$FinalS = \frac{\sum_{i=1}^{5}\sum_{j=1}^{q} Conf(i,j)\cdot\left\{M(I_c,P_j)\odot Sal_i\right\}}{\sum_{i=1}^{5}\sum_{j=1}^{q} Conf(i,j)}, \quad (8)$$

where q is the number of semantically salient objects in the given image; function $M(I_c, P_j)$ returns a binary mask matrix $Mask \in \{0, 1\}$ (with identical size to I_c), in which only those pixels in object box P_j are non-zero elements. The saliency map quality computed by our adaptive fusion scheme can significantly outperform other conventional fusion schemes (see Table V and Table VI).

I. Why Does Our Method Is A Weakly-supervised One?

The main reason is that we do not directly use the saliency outputs of the adopted five pre-trained models in our training process. Instead, our pseudo-GT is just the binary ranks, which are newly formulated by the novel scheme mentioned in Sec. III-E.

The novelty of our weakly supervised learning can be confirmed by the following two facts. First, the main purpose is different. Our method aims to investigate the object-level semantic ranks between similar images, whereas the previous work were mainly designed for pixel-wise saliency regression in single image. Second, the key methodology is different. We have adopted the object-level IOU rate to formulate the semantic pseudo-GT, making the key idea of using semantic ranking for the SOD task being feasible.

J. Major Differences to the SOTA Models

The previous work (e.g., CPD [65]) has treated the SOD task as a multi-task problem which aims to conduct pixel-wise saliency regression and segmentation-like saliency refinement simultaneously, making their networks difficult to reach convergence. As a result, even at the expense of decreasing the use of semantic information, the existing models must update their feature backbones to ensure the learning convergence.

In sharp contrast, our method has divided the SOD problem into two sequential tasks: localization first and refinement later. Since our localization has a relatively small problem domain, it can be fully realized from the semantic perspective by using "fixed" backbones with shallower FC layers. Moreover, different from the conventional methods which have focused on learning pixel-wise saliency in single image using perceptual saliency cues solely, our method has investigated the object-level semantic ranks between multiple images, which is more consistent with both human vision system and human brain.

IV. Experiments and Results

We have conducted massive quantitative experiments to validate the effectiveness of our method, where we have compared our model with 11 SOTA models over 5 public available datasets to demonstrate the advantages of our method. The testing sets adopted in our quantitative evaluation include DUTS-TE [62], ECSSD [20], HKU-IS [45], PASCAL-S [72] and SOD [73].

A. Evaluation Metrics

We adopt the F-measure and the mean absolute error (MAE) to evaluate the performance of our method. As the recall rate is inversely proportional to the precision, the tendency of the trade-off between precision and recall can truly indicate the overall saliency detection performance. Thus, we utilize the F-measure (β^2=0.3) to evaluate such trade-off. Moreover, since both metrics of MAE and F-measure are based on pixel-wise errors and often ignore the structural similarities, we also adopt the structure measure S-measure [74] and enhanced-measure E-measure [75] to conduct quantitative evaluation. The evaluation source code is provided by Fan et al. [1], [74], [75]. In addition, we have also resorted the **n**ormalized **s**canpath **s**aliency (NSS) [76] metric, a simple correspondence measure between saliency maps and GT computed as the average normalized saliency at fixated locations.

B. Training and Implementation Details

We implement the proposed approach in Python with the Pytorch toolbox. We run our approach on a computer with 4 Tesla P100 GPU with 64G memory. The retrieval image pool consists of almost 30K images, i.e., THUR15K[61]+MSRA10K[19]+DUTSTE5K[62]. It should be noted that we do not use any perceptual saliency ground truth of these 30K images.

Our training set only comprises 10K images (DUTSTR10K [62]), and we have applied its perceptual saliency GT to facilitate the pseudo-GT formulation for the intra-image instances which take about 30% of the total. As for the rest of 70% object pairs (i.e., the inter-image instances), we weakly formulate their pseudo-GT by using the saliency predictions of five adopted saliency models.

To avoid data breach, all these 5 adopted saliency models are pre-trained using the same training set as our method (i.e., DUTSTR10K), and the selected 5 methods include CPD [65], BASNet [70], MWS [8], PFA [71] and BRN [77].

C. Comparison with Other SOTA Models

We have compared our method with other 11 most recent SOTA deep models, which are EGNet [66], BANet [67], AFNet [68], CPD [65], MLM [69], BASNet [70], MWS [8], PFA [71], BRN [77], RFCN [78] and RADF [12].

For an objective comparison, all quantitative evaluations are conducted using the saliency maps either provided by the authors or obtained by the runable codes with parameters unchanged. We demonstrate the detailed S-measure [74],

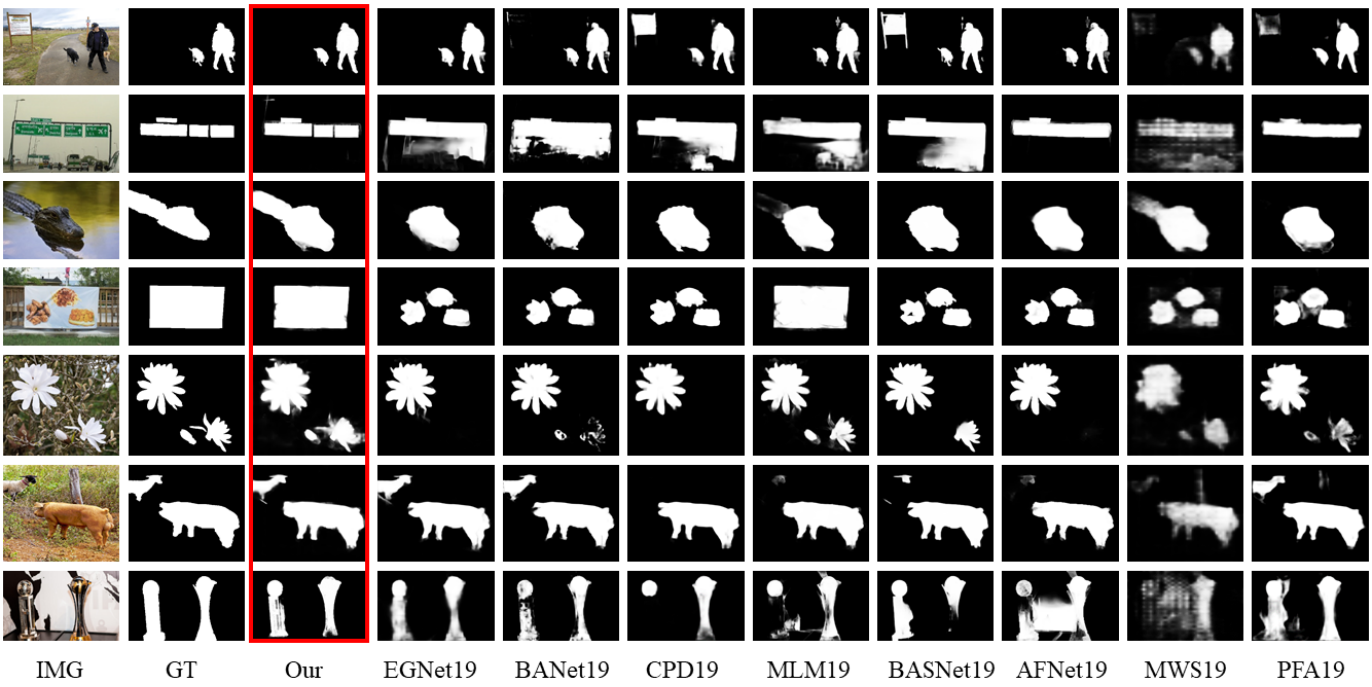

Fig. 5: Qualitative comparisons between our method and several most recent deep models, including EGNet [66], BANet [67], AFNet [68], CPD [65], MLM [69], BASNet [70], MWS [8] and PFA [71].

MAE, F-measure, E-measure [75] with different thresholds in Table. I, including NSS as well.

For those images with multiple salient objects (Fig. 5), the SOTA models tend to assign large saliency values to those non-salient objects due to their strong perceptual saliency. The SOTA models also tend to miss-detect some salient objects occasionally, because these multiple salient object cases are really challenge for their single-image based methodology. However, benefited from the proposed semantic saliency, our model can produce better detection results.

Also, in the view of quantitative comparisons (Table I), our model outperforms other competitors by large margin on 4 out of 5 sets excepting the ECCSD set. The main reason is that the semantic categories of ECCSD set are quite different to those of our retrieval image pool. Even so, the overall performance of our model on ECCSD is still comparable to other SOTA models. We believe this issue can be alleviated if we increase the retrieval image pool size or diversity.

In addition, our model outperforms the EGNet on the DUTS-TE set, where the EGNet has performed the best among all tested SOTA models. In this case, our model improves the SOTA performance significantly, e.g., the maxFm value has been improved from 0.866 to 0.910. Similar tendencies take place on HKU-IS and PASCAL-S sets. For example, the maxFm metric on the PASCAL-S set has been improved from 0.858 (the second best, achieved by PFA) to 0.902. The main reason for such significant improvement is that our model is able to locate salient object more accurate than other SOTA competitors. Thus, the maxFm metric, a valuable indicator to show if salient objects are located accurately, clearly shows the effectiveness of the proposed methodology, where our method frees the segmentation task out its main learning objective.

Notice that our model cannot perform the best in the MAE metric. This is because the MAE metric is mainly used to evaluate deep models in the view of tiny saliency details. In a word, a deep model would exhibit small MEA value if this model is strong in retaining spatial details (e.g., sharp object boundary). For example, the EGNet has adopted multiple edge losses for preserving sharp object boundaries, thus this model can exhibit the best MAE results for 4 out of 5 tested datasets. Since the saliency detail aspect is slightly beyond the main topic of this paper, it is quite reasonable for our model being not very good in MAE metric.

TABLE II: Quantitative comparisons between the single-scale deep feature (i.e., 4096) and multi-scale deep feature (i.e., 4096+4096). "SS": Single-scale, "MS": Multi-scale, GH: the classic GIST+HOG image retrieval, HB: the hybrid image retrieval (3 images using SE, and 2 images using GH); "DUT.": DUTS-TE [62], "ECS.": ECSSD [20], "HKU.": HKU-IS [45], "PAS.": PASCAL-S [72], SOD [73]; "Sme.": Smeasure, "ad.E": adpEmeasure, "me.E": meanEmeasure, "ma.E": maxEmeasure, "ad.F": adpFmeasure, "me.F": meanFmeasure, "ma.F": maxFmeasure.

Sets		Sme.	MAE	ad.E	me.E	ma.E	ad.F	me.F	ma.F
DUT.	SS	.780	.616	.850	.857	.863	.691	.702	.711
	MS	**.909**	**.039**	**.928**	**.894**	**.962**	**.839**	**.841**	**.910**
ECS.	SS	.699	.110	.761	.766	.781	.801	.819	.823
	MS	**.836**	**.044**	**.934**	**.918**	**.972**	**.909**	**.899**	**.955**
HKU.	SS	.724	.111	.766	.791	.801	.752	.756	.795
	MS	**.930**	**.037**	**.962**	**.916**	**.974**	**.912**	**.884**	**.943**
PAS.	SS	.757	.129	.692	.700	.702	.681	.690	.714
	MS	**.894**	**.065**	**.894**	**.879**	**.943**	**.846**	**.845**	**.902**
SOD	SS	.745	.161	.742	.751	.761	.777	.781	.788
	MS	**.806**	**.112**	**.836**	**.779**	**.912**	**.825**	**.777**	**.878**

TABLE I: Quantitative comparisons between our method and SOTA methods in metrics including maximum F-measure (larger is better), MAE (smaller is better) , E-measure (larger is better), adaptive F-measure (larger is better), adaptive E-measure (larger is better) and NSS (larger is better). The top three results are highlighted in red, green, and blue, respectively.

	Metric	OUR	EGNet 19 [66]	BANet 19 [67]	AFNet 19 [68]	CPD 19 [65]	MLM 19 [69]	BASNet 19 [70]	MWS 19 [8]	PFA 19 [71]	BRN 18 [77]	RFCN 18 [78]	RADF 18 [12]
DUTS-TE [62]	Smeasure↑	.909	.887	.879	.867	.869	.862	.866	.759	.874	.851	.792	.824
	MAE↓	.039	.039	.040	.046	.043	.049	.048	.091	.041	.054	.074	.072
	adpEm↑	.928	.891	.892	.879	.886	.860	.884	.814	.877	.860	.837	.819
	meanEm↑	.894	.907	.913	.893	.898	.883	.895	.743	.910	.876	.826	.840
	maxEm↑	.962	.927	.927	.910	.914	.907	.903	.833	.933	.899	.854	.874
	adpFm↑	.839	.815	.815	.792	.805	.745	.791	.684	.784	.755	.709	.700
	meanFm↑	.841	.839	.835	.812	.821	.792	.822	.648	.815	.780	.713	.737
	maxFm↑	.910	.866	.858	.838	.840	.827	.838	.720	.852	.811	.736	.786
	NSS↑	2.92	2.89	2.76	2.71	2.73	2.71	2.68	2.44	2.58	2.61	2.29	2.47
ECSSD [20]	Smeasure↑	.836	.925	.924	.913	.910	.911	.917	.828	.903	.909	.869	.895
	MAE↓	.044	.037	.035	.042	.040	.045	.037	.096	.046	.046	.067	.060
	adpEm↑	.934	.927	.928	.918	.922	.914	.922	.885	.910	.915	.907	.908
	meanEm↑	.918	.943	.948	.935	.938	.927	.944	.792	.936	.930	.897	.907
	maxEm↑	.972	.955	.958	.947	.944	.945	.952	.910	.948	.947	.921	.933
	adpFm↑	.909	.920	.923	.908	.915	.869	.882	.840	.886	.893	.871	.872
	meanFm↑	.899	.918	.923	.905	.912	.890	.917	.763	.889	.893	.866	.893
	maxFm↑	.955	.936	.939	.924	.923	.918	.931	.859	.913	.917	.885	.905
	NSS↑	2.10	2.01	2.00	1.97	1.96	1.95	1.97	1.81	1.80	1.94	1.86	1.93
HKU-IS [45]	Smeasure↑	.930	.918	.913	.905	.905	.906	.908	.818	.913	.901	.862	.888
	MAE↓	.037	.031	.032	.036	.034	.039	.032	.084	.033	.041	.054	.050
	adpEm↑	.962	.950	.950	.942	.944	.937	.945	.895	.946	.939	.924	.919
	meanEm ↑	.916	.944	.946	.934	.938	.930	.943	.787	.948	.929	.897	.909
	maxEm ↑	.974	.958	.958	.949	.950	.950	.951	.908	.962	.949	.931	.938
	adpFm ↑	.912	.901	.900	.888	.890	.871	.895	.814	.883	.875	.858	.851
	meanFm ↑	.884	.902	.903	.888	.891	.878	.902	.734	.891	.875	.848	.856
	maxFm ↑	.943	.924	.923	.910	.910	.910	.918	.835	.918	.903	.872	.895
	NSS↑	2.26	2.25	2.23	2.22	2.22	2.23	2.22	2.02	2.00	2.20	2.10	2.19
PASCAL-S [72]	Smeasure↑	.894	.849	.849	.849	.846	.841	.835	.765	.859	.841	.796	.815
	MAE↓	.065	.076	.072	.070	.072	.076	.078	.135	.066	.079	.105	.101
	adpEm↑	.894	.852	.857	.851	.853	.840	.850	.789	.840	.838	.830	.821
	meanEm↑	.879	.879	.889	.883	.880	.873	.876	.732	.894	.871	.828	.832
	maxEm↑	.943	.889	.897	.895	.889	.889	.883	.830	.917	.888	.847	.854
	adpFm↑	.846	.821	.828	.822	.825	.762	.773	.716	.818	.798	.772	.765
	meanFm↑	.845	.826	.834	.826	.823	.805	.820	.670	.823	.807	.772	.775
	maxFm↑	.902	.844	.852	.845	.837	.833	.837	.756	.858	.833	.788	.808
	NSS↑	2.11	2.03	2.02	2.00	2.00	1.98	1.94	1.73	2.00	1.95	1.81	1.87
SOD [73]	Smeasure↑	.806	.802	.788	/	.767	.786	.769	.700	/	.779	.719	.765
	MAE↓	.112	.099	.107	/	.112	.108	.113	.167	/	.109	.146	.133
	adpEm↑	.836	.818	.812	/	.793	.800	.777	.775	/	.801	.791	.801
	meanEm↑	.779	.818	.825	/	.778	.799	.798	.657	/	.798	.745	.783
	maxEm↑	.912	.868	.860	/	.848	.844	.829	.821	/	.843	.820	.832
	adpFm↑	.825	.840	.831	/	.810	.764	.746	.738	/	.794	.768	.776
	meanFm↑	.777	.819	.822	/	.770	.779	.790	.631	/	.782	.737	.766
	maxFm↑	.878	.845	.838	/	.814	.806	.805	.772	/	.807	.777	.798
	NSS↑	1.70	1.70	1.66	/	1.61	1.63	1.55	1.48	/	1.60	1.46	1.59

D. Component Evaluations

1) *The Effectiveness of the Proposed Perceptual Contrast*

As we have mentioned in Sec. III-B, we combine the local deep features with those of its enlarged version, obtaining a {4096+4096} dimensional semantic-aware deep feature. Compared with using the single-scale 4096 dimensional deep features, the proposed perceptual contrast is capable of alleviating the learning ambiguity when performing semantical saliency ranking for two object proposals with similar semantic information. As shown in Table II, the multi-scale perceptual contrast scheme (MS) outperforms the single-scale scheme (SS) persistently and significantly for all tested sets. We take the HKU-IS set for instance, where the MS scheme has boosted the S-measure metric value from 0.724 to 0.930.

2) *The Effectiveness of the Proposed Hybrid Retrieval Scheme*

As we have mentioned in Sec. III-D, the exact choice of image retrieval scheme influence the overall performance directly. In this quantitative verification, we have tested three different retrieval schemes, including (I) retrieval using semantical information solely, (II) retrieval using {GIST+HOG} features and (III) retrieval using both semantical information and {GIST+HOG} features, where the corresponding pictorial

TABLE III: Quantitative comparisons between different image retrieval choices (we empirically retrieval 5 images), i.e., SE: the semantic information based image retrieval, GH: the classic GIST+HOG image retrieval, HB: the hybrid image retrieval (3 images using SE, and 2 images using GH); "DUT.": DUTS-TE [62], "ECS.": ECSSD [20], "HKU.": HKU-IS [45], "PAS.": PASCAL-S [72], SOD [73]; "Sme.": Smeasure, "ad.E": adpEmeasure, "me.E": meanEmeasure, "ma.E": maxEmeasure, "ad.F": adpFmeasure, "me.F": meanFmeasure, "ma.F": maxFmeasure.

		Sme.	MAE	ad.E	me.E	ma.E	ad.F	me.F	ma.F
DUT.	HB	**.909**	.039	**.928**	**.894**	**.962**	.839	**.841**	**.910**
	SE	.889	**.038**	.909	.882	.918	.820	.839	.847
	GH	.864	.039	.903	.889	.921	**.852**	.839	.847
ECS.	HB	**.936**	**.044**	**.934**	.918	**.972**	.909	.899	**.955**
	SE	.912	.045	.923	**.940**	.951	**.928**	**.921**	.926
	GH	.869	.056	.861	.873	.892	.847	.825	.853
HKU.	HB	**.930**	**.037**	**.962**	.916	**.974**	.912	.884	**.943**
	SE	.905	.039	.947	**.944**	.952	**.915**	**.909**	.915
	GH	.911	.045	.928	.917	.943	.891	.862	.888
PAS.	HB	**.894**	**.065**	**.894**	.879	**.943**	**.846**	**.845**	**.902**
	SE	.867	.068	.860	**.888**	.912	.827	.835	.867
	GH	.861	.066	.850	.874	.901	.821	.817	.849
SOD	HB	**.806**	.112	**.836**	**.779**	**.912**	**.825**	**.777**	**.878**
	SE	.758	**.110**	.759	.772	.814	.778	.765	.802
	GH	.748	.113	.738	.761	.790	.756	.746	.764

demonstrations were previously provided in Fig. 4.

As shown in Table III, the hybrid scheme (HB) outperforms other two as expected. The main reason is also clear that the hybrid scheme is more likely to provide a stable object-level co-occurrence status, making our learning objective more practical.

TABLE IV: Quantitative comparisons between different network architectures (see details in Sec. III-C) using the conventional entropy loss and our hinge loss respectively. "EL": Entropy Loss, "HL": Hinge Loss, "DUT.": DUTS-TE [62], "ECS.": ECSSD [20], "HKU.": HKU-IS [45], "PAS.": PASCAL-S [72], SOD [73]; "Sme.": Smeasure, "ad.E": adpEmeasure, "me.E": meanEmeasure, "ma.E": maxEmeasure, "ad.F": adpFmeasure, "me.F": meanFmeasure, "ma.F": maxFmeasure.

		Sme.	MAE	ad.E	me.E	ma.E	ad.F	me.F	ma.F
DUT.	EL	.904	.042	.910	.888	.956	**.869**	**.850**	**.915**
	HL	**.909**	**.039**	**.928**	**.894**	**.962**	.839	.841	.910
ECS.	EL	**.905**	.062	.909	.890	.948	.868	.846	.917
	HL	.836	**.044**	**.934**	**.918**	**.972**	**.909**	**.899**	**.955**
HKU.	EL	.928	.039	.949	.912	.970	.897	.867	.942
	HL	**.930**	**.037**	**.962**	**.916**	**.974**	**.912**	**.884**	**.943**
PAS.	EL	.877	.079	.855	.844	.917	.796	.788	.879
	HL	**.894**	**.065**	**.894**	**.879**	**.943**	**.846**	**.845**	**.902**
SOD	EL	.786	.124	.811	.763	.881	.801	.757	.849
	HL	**.806**	**.112**	**.836**	**.779**	**.912**	**.825**	**.777**	**.878**

3) *Why Do We Choose to Learn the Semantic Saliency Relationship between Objects?*

Actually, this issue has been fully discussed in both Sec. III-C and Sec. III-J, where, as one of the major advantages, the proposed object-level semantical ranking scheme is more consistent with the real human visual system and human brain, making the SOD task more easier to be realized in practice.

Meanwhile, from the technical implementation perspective, the major difference between the proposed object-level semantical ranking scheme and the conventional single-object based regression is the usage of different loss functions, where we use the **h**inge **l**oss (HL) in our object-level semantical ranking scheme, while the single-object based regression scheme adopts the conventional **e**ntropy **l**oss (EL).

The quantitative results regarding these two different schemes have been provided in Table IV. Not surprisingly, the HL outperforms the EL significantly for all tested sets.

TABLE V: Salient object localization comparisons between our method and the conventional perceptual saliency models. The first three are our method with different schemes to assign q. CPD [65], MWS [8], MLM [69]. "P": Precision, "R": Recall, "F": F-measure; "DUT.": DUTS-TE [62], "ECS.": ECSSD [20], "HKU.": HKU-IS [45], "PAS.": PASCAL-S [72], SOD [73].

		q Eq. 6	$q=3$	$q=5$	**CPD**	**MWS**	**MLM**
DUT.	Pre.	**.989**	.645	.399	**.999**	.880	.959
	Rec.	**.760**	.540	.545	.579	.653	**.751**
	Fme.	**.925**	.617	.425	.856	.815	**.901**
ECS.	Pre.	**.843**	.520	.470	.706	.755	**.769**
	Rec.	**.964**	.786	.949	.851	.660	**.916**
	Fme.	**.868**	.564	.532	.735	.730	**.798**
HKU.	Pre.	**.947**	.660	.455	.948	.864	**.970**
	Rec.	**.843**	.313	.800	.553	.668	**.754**
	Fme.	**.921**	.526	.505	.814	.809	**.910**
PAS.	Pre.	**.888**	.788	.694	.871	.652	**.884**
	Rec.	**.949**	.913	.931	.783	**.907**	.891
	Fme.	**.902**	.814	.738	.849	.697	**.885**
SOD	Pre.	**.916**	.446	.303	**.906**	.795	.747
	Rec.	**.942**	.738	.813	**.890**	.875	.864
	Fme.	**.922**	.490	.355	**.902**	.812	.771

4) *W.r.t the Salient Object Localization, If the Proposed Semantical Saliency Outperforms the Conventional Perceptual Saliency?*

We have compared our method with the three most representative perceptual saliency models (i.e., CPD, MWS and MLM) towards the salient object localization task, where the original well annotated saliency GTs have been replaced by rectangular GTs wrapping salient objects tightly. We regard a localization (i.e., the white rectangle in Fig. 2-C) as a successful one if the overlapping rate between this proposal and the corresponding rectangular GT is larger than 70%.

We show the quantitative comparison results in Table V, where we have computed the precision, recall and f-measure based on the aforementioned successful rate. The quantitative results have suggested that our method has achieved almost 8% in f-measure improvement. Meanwhile, this quantitative evaluation has also verified the effectiveness of the proposed adaptive scheme (Eq. 6).

5) *The Effectiveness of the Proposed Refinement Scheme*

We have tried to fuse the saliency maps derived from five adopted saliency models via multiplicative, average, maximizing fusion strategies. Because these simple fusion schemes have failed in making full use of the complementary status

TABLE VI: Quantitative comparisons between our method and several classic fusion strategies. Max5, Mult5, and Mix5 respectively represent the saliency maps by fusing the 5 adopted saliency models using maximizing, multiplicative and average.

		Sme.	MAE	ad.E	me.E	ma.E	ad.F	me.F	ma.F
DUT.	OUR	**.909**	**.039**	**.928**	**.894**	**.962**	**.839**	**.841**	**.910**
	Mix5	.873	.057	.848	.852	.926	.740	.783	.871
	Max5	.820	.076	.782	.840	.899	.684	.721	.824
	Mult5	.854	.056	.831	.879	.917	.732	.782	.843
ECS.	OUR	.836	**.044**	**.934**	.918	**.972**	**.909**	**.899**	**.955**
	Mix5	**.926**	.052	.920	.904	.961	.886	.879	.945
	Max5	.898	.054	.872	.916	.955	.859	.855	.928
	Mult5	.915	.046	.907	**.929**	.958	.888	.884	.932
HKU.	OUR	**.930**	**.037**	**.962**	.916	**.974**	**.912**	**.884**	**.943**
	Mix5	.916	.045	.938	.900	.964	.878	.860	.931
	Max5	.886	.052	.904	.907	.947	.827	.831	.900
	Mult5	.899	.045	.921	**.922**	.957	.850	.856	.916
PAS.	OUR	**.894**	**.065**	**.894**	**.879**	**.943**	**.846**	**.845**	**.902**
	Mix5	.856	.087	.834	.840	.905	.766	.790	.859
	Max5	.812	.100	.759	.834	.896	.749	.753	.844
	Mult5	.846	.082	.813	.865	.907	.784	.796	.861
SOD	OUR	.806	.112	**.836**	.779	**.912**	**.825**	.777	**.878**
	Mix5	.797	.120	.815	.766	.878	.792	.758	.848
	Max5	**.826**	**.097**	.799	**.850**	.876	.798	.805	.836
	Mult5	.822	.104	.812	.831	.874	.814	**.807**	.840

between different saliency maps, their overall performances are quite limited. In sharp contrast, our method is able to selectively fuse these saliency maps, compressing those less trustworthy regions while highlighting the salient objects in a full automatical way, thus, as shown in Table VI, it can outperform other competitors significantly. More qualitative demonstration regarding this quantitative comparison can be found in Fig. 6, where our refinement scheme can produce better saliency maps clearly.

E. Limitation

Our method tends to be time-consuming in general, because our approach is not an end-to-end one, and the object-level multi-scale deep feature computation is the major computational bottle-neck. As shown in Table VII, our method takes about 0.65s for a single image.

TABLE VII: The averaged time consumption (using a GTX1080GPU) towards different components.

Main Steps	Time(seconds)
Object Proposal Preliminaries (Sec. III-A)	0.020
Multi-scale Deep Feature (Sec. III-B)	0.450
Semantically Salient Object Localization (Sec. III-G)	0.150
Pixel-wise Saliency Refinement (Sec. III-H)	0.030
Total	0.650

V. Conclusion

In this paper, we have revisited the problem of image salient object detection from the semantic perspective, i.e., the semantic saliency is more important than the conventional perceptual saliency to drive the attention of the real human visual system. We have provided a novel network to learn the relative semantic saliency degree for two input objects proposals. Also, we have proposed a weakly supervised scheme to train our siamese network for ranking the object-level semantic saliency, where the final saliency map can be computed in the coarse-to-fine manner. One major advantage of the proposed scheme is its strong ability in locating salient objects accurately. As a post-processing plug-in, we have devised a novel saliency refinement scheme to compensate the lost spatial details when performing the single-task salient object localization. Lastly, we have conducted extensive quantitative evaluations to verify the effectiveness of each component in our method.

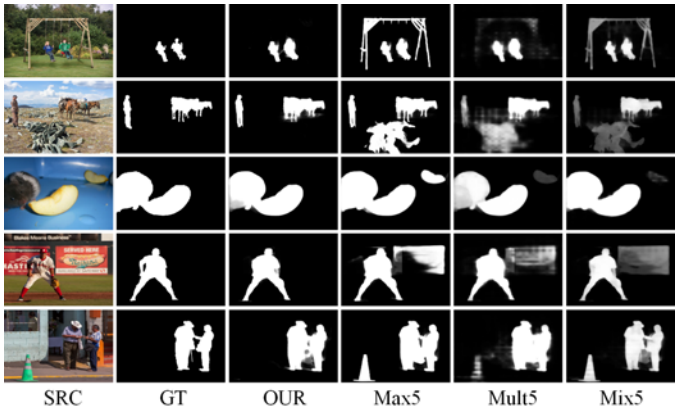

Fig. 6: Visual comparisons between our method and multiple classic fusion schemes; Max5, Mult5, and Mix5 respectively represent saliency maps obtained by simply fusing all outputs of the adopted five deep models via maximizing, multiplicative and average; the adopted five deep models include CPD [65], BASNet [70], MWS [8], PFA [71] and BRN [77].

Acknowledgments. This research was supported in part by National Key R&D Program of China (No. 2017YFF0106407), National Natural Science Foundation of China (No. 61802215 and No. 61806106), Natural Science Foundation of Shandong Province (No. ZR201807120086) and National Science Foundation of USA (No. IIS-1715985, IIS-0949467, IIS-1047715, and IIS-1049448).

References

[1] D. Fan, M. Cheng, J. Liu, S. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 186–202.

[2] Y. Fang, C. Zhang, H. Huang, and J. Lei, "Visual attention prediction for stereoscopic video by multi-module fully convolutional network," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5253–5265, 2019.

[3] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3156–3170, 2017.

[4] C. Chen, G. Wang, C. Peng, X. Zhang, and H. Qin, "Improved robust video saliency detection based on long-term spatial-temporal information," *IEEE Transactions on Image Processing*, vol. 29, pp. 1090–1100, 2019.

[5] G. Ma, C. Chen, S. Li, C. Peng, H. Qin, and A. Hao, "Salient object detection via multiple instance joint re-learning," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 324–336, 2020.

[6] C. Peng, C. Chen, Z. Kang, J. Li, and Q. Cheng, "RES-PCA: A scalable approach to recovering low-rank matrices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7317–7325.

[7] Y. Zhuge, Y. Zeng, and H. Lu, "Deep embedding features for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 9340–9347.

[8] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6074–6083.

[9] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4296–4307, 2020.

[10] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2019, pp. 7274–7283.

[11] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin, "Accurate and robust video saliency detection via self-paced diffusion," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1153–1167, 2019.

[12] X. Hu, L. Zhu, J. Qin, C. Fu, and P. Heng, "Recurrently aggregating deep features for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[13] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 815–828, 2019.

[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[15] T. Wang, L. Zhang, and H. Lu, "Kernelized subspace ranking for saliency detection," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 450–466.

[16] C. Chen, S. Li, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2303–2316, 2015.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2006, pp. 545–552.

[18] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1597–1604.

[19] M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[20] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1155–1162.

[21] Z. Liu, O. LeMeur, S. Luo, and L. Shen, "Saliency detection using regional histograms," *Optics Letters*, vol. 38, no. 5, pp. 700–702, 2013.

[22] L. Olivier, "Predicting saliency using two contextual priors: The dominant depth and the horizon line," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6.

[23] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 438–445.

[24] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2013, pp. 1761–1768.

[25] L. Maczyta, P. Bouthemy, and O. LeMeur, "Unsupervised motion saliency map estimation based on optical flow inpainting," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4469–4473.

[26] O. LeMeur and P. Fons, "Predicting image influence on visual saliency distribution: the focal and ambient dichotomy," in *Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.

[27] D. Fan, W. Wang, M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8554–8564.

[28] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin, "A plug-and-play scheme to adapt image saliency deep model for video data," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2020.

[29] C. Chen, S. Li, H. Qin, Z. Pan, and G. Yang, "Bi-level feature learning for video saliency detection," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3324–3336, 2018.

[30] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4819–4831, 2019.

[31] K. Fu, D. Fan, G. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3052–3062.

[32] D. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 275–292.

[33] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[34] H. Li, G. Li, B. Yang, G. Chen, L. Lin, and Y. Yu, "Depthwise nonlocal module for fast salient object detection using a single thread," *IEEE Transactions on Cybernetics*, pp. 1–12, 2020.

[35] X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, and H. Qin, "Data-level recombination and lightweight fusionscheme for RGB-D salient object detection," *IEEE Transactions on Image Processing*, pp. 1–1, 2021.

[36] B. Follet, O. LeMeur, and T. Baccino, "New insights into ambient and focal visual fixations using an automatic classification algorithm," *i-Perception*, vol. 2, no. 6, pp. 592–610, 2011.

[37] L. Maczyta, P. Bouthemy, and O. LeMeur, "CNN-based temporal detection of motion saliency in videos," *Pattern Recognition Letters*, vol. 128, pp. 298–305, 2019.

[38] A. Nebout, W. Wei, Z. Liu, L. Huang, and O. LeMeur, "Predicting saliency maps for ASD people," in *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2019, pp. 629–632.

[39] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[40] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2017, pp. 202–211.

[41] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2017, pp. 4019–4028.

[42] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6609–6617.

[43] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2017, pp. 212–221.

[44] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detectio," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3051–3060.

[45] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5012–5024, 2016.

[46] L. Wang, H. Lu, X. Ruan, and M. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3183–3192.

[47] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2017.

[48] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, 2013.

[49] K. Zhang, T. Li, B. Liu, and Q. Liu, "Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3095–3104.

[50] C. Tsai, W. Li, K. Hsu, X. Qian, and Y. Lin, "Image co-saliency detection and co-segmentation via progressive joint optimization," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 56–71, 2018.

[51] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 568–579, 2018.

[52] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1639–1651, 2017.

[53] D. Fan, T. Li, Z. Lin, G. Ji, D. Zhang, M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *arXiv preprint arXiv:2007.03380*, 2020.

[54] C. Sumit, H. Raia, and L. Yann, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 539–546.

[55] S. Zhou, J. Zhang, J. Wang, F. Wang, and D. Huang, "SE2Net: Siamese edge-enhancement network for salient object detection," *arXiv preprint arXiv:1904.00048*, 2019.

[56] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3623–3632.

[57] Y. Ji, H. Zhang, Z. Jie, L. Ma, and J. Wu, "CASNet: A cross-attention siamese network for video salient object detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2020.

[58] K. Fu, D. Fan, G. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *arXiv preprint arXiv:2008.12134*, 2020.

[59] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2965–2974.

[60] M. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *International Symposium on Visual Computing*, 2016, pp. 234–244.

[61] M. Cheng, N. Mitra, X. Huang, and S. Hu, "Salientshape: Group saliency in image collections," *The Visual Computer*, vol. 30, no. 4, pp. 443–453, 2014.

[62] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1265–1274.

[63] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[64] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2003, pp. 273–280.

[65] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3907–3916.

[66] J. Zhao, J. Liu, D. Fan, Y. Cao, J. Yang, and M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2019, pp. 8779–8788.

[67] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2019, pp. 3799–3808.

[68] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1623–1632.

[69] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8150–8159.

[70] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7479–7489.

[71] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[72] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 280–287.

[73] V. Movahedi and J. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 49–56.

[74] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2017, pp. 4548–4557.

[75] D. Fan, C. Gong, Y. Cao, B. Ren, M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proceedings of the International Joint Conference on Artificial Intelligence (AAAI)*, 2018, pp. 698–704.

[76] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[77] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3127–3135.

[78] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1734–1746, 2018.