# CapSal: Leveraging Captioning to Boost Semantics for Salient Object Detection

Lu Zhang[1], Jianming Zhang[2], Zhe Lin[2], Huchuan Lu[1], You He[3]
[1]Dalian University of Technology, China
[2]Adobe Research, USA
[3]Naval Aviation University, China
luzhang_dut@mail.dlut.edu.cn, {jianmzha, zlin}@adobe.com, lhchuan@dlut.edu.cn, heyou_f@126.com

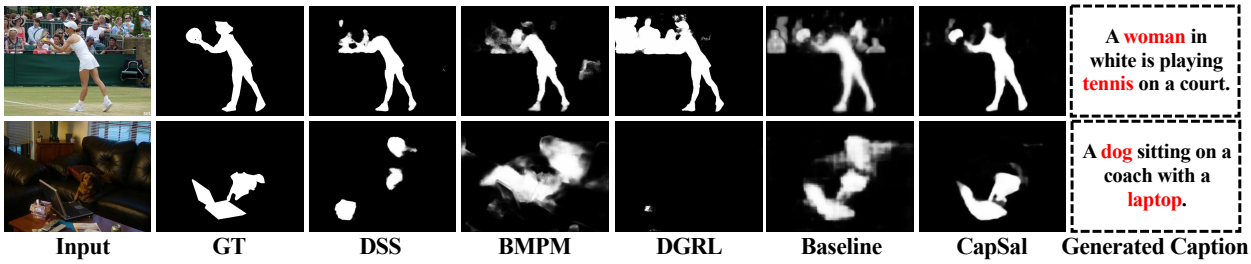



Figure 1: Visual comparison with other CNN-based methods. From left to right: input image, ground truth, saliency maps of DSS [11], BMPM [41], DGRL [35], our baseline model, CapSal model and generated caption by ICN. The words highlighted in red gain higher attention scores. Our CapSal network is trained to leverage the semantics in captioning task for salient object detection, which can precisely localize the salient regions from cluttered background.

## Abstract

*Detecting salient objects in cluttered scenes is a big challenge. To address this problem, we argue that the model needs to learn discriminative semantic features for salient objects. To this end, we propose to leverage captioning as an auxiliary semantic task to boost salient object detection in complex scenarios. Specifically, we develop a CapSal model which consists of two sub-networks, the Image Captioning Network (ICN) and the Local-Global Perception Network (LGPN). ICN encodes the embedding of a generated caption to capture the semantic information of major objects in the scene, while LGPN incorporates the captioning embedding with local-global visual contexts for predicting the saliency map. ICN and LGPN are jointly trained to model high-level semantics as well as visual saliency. Extensive experiments demonstrate the effectiveness of image captioning in boosting the performance of salient object detection. In particular, our model performs significantly better than the state-of-the-art methods on several challenging datasets of complex scenarios.*

## 1. Introduction

Salient object detection is a fundamental problem in computer vision, aiming to localize and segment the most conspicuous regions in an image. In recent years, it has achieved much attention due to its usefulness to many computer vision applications [12, 36, 44].

Although significant progresses have been made in this area thanks to the deep learning technology, it still remains a big challenge to accurately detect salient objects in cluttered scenes (see Fig.1). To address this problem, we argue that the model needs to learn discriminative semantic features for salient objects, such as object categories, attributes and the semantic context. However, existing salient object detection networks are only trained on pixel-level mask annotations, with no supervision on higher-level semantics.

In this work, we propose to use image captioning [37, 26, 30] as an auxiliary task to boost the semantics for salient object detection. The connection between image captioning and saliency detection has already been explored in the image captioning domain. Some works [1, 28] utilize saliency detection to make the network attend to relevant regions for captioning. These works assume that the objects being mentioned in caption are largely consistent and correlated with the salient objects [40]. Based on the same assumption, we believe that the captioning task can provide rich semantic supervision for salient object detection. For example, from the caption "A woman in white is playing tennis on a court", we can obtain the overall knowledge about the category, attribute and motion of salient object (see Fig.1).

To this end, we propose CapSal, a salient object detection framework that exploits image captioning to pro-

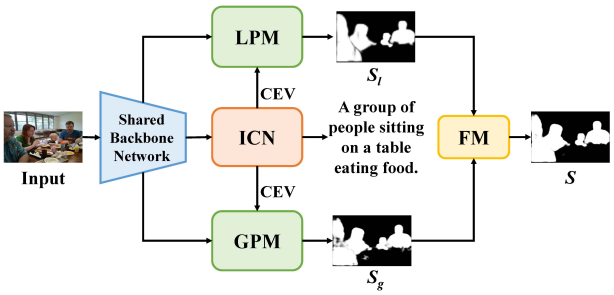



Figure 2: An overview of the proposed CapSal model. For an input image, a shared backbone network is employed to extract multi-level features. Then a RNN-based captioning model is used to encode each word in the caption. The latent features for each word are then pooled using an attention mechanism to obtain a Caption Embedded Vector (CEV). The caption embedded vector and multi-level features are incorporated into local-global perception network, which predicts salient objects in both local and global view. At the end, we obtain the final saliency map $S$ by fusing the saliency maps, $S_l$ and $S_g$, from the local and the global perception modules.

mote the semantic feature learning for salient object detection. The CapSal model consists of two sub-networks with a shared backbone, which are Image Captioning Network (ICN) and Local-Global Perception Network (LGPN), for caption generation and saliency prediction, respectively. The framework of our CapSal model is shown in Fig.2. The ICN is a CNN+LSTM architecture, which takes an image as input and generates a caption. In order to capture the object-level semantic knowledge from caption, we use the hidden vectors of LSTM to represent the encoding feature of each generated word. Considering that not every word in the caption is relevant for describing the salient object, we propose a textual attention mechanism to weight the importance of each word. Then a caption embedded feature vector can be obtained via weighted pooling of the LSTM hidden vectors.

The other sub-network, LGPN, is designed for integrating the caption embedded vector with multi-contextual visual features for identifying salient object. It consists of three components: a Local Perception Module (LPM), a Global Perception Module (GPM) and a Fusion Module (FM). In LPM, caption embedded vector is aggregated with visual features in a local view to capture fine details of objects. While the GPM utilizes context in a more global view to give a holistic estimation of salient regions. LPM and GPM are complementary in detecting objects of various sizes, and their saliency maps are fused by FM to generate the final saliency map. Both LPM and GPM leverage the caption embedded feature generated by ICN to capture higher-level semantic information of the scene. ICN and LGPN are jointly optimized during training using the captioning and saliency supervision respectively.

To train and evaluate the proposed method, we build a new saliency dataset, COCO-CapSal, which contains the ground truth saliency map as well as the corresponding captions for each image. Images in the dataset are from the MSCOCO [22] dataset and have multiple salient objects from 80 categories with cluttered background. Our experiments validate the effectiveness of image captioning in boosting the performance of salient object detection. In particular, our model significantly outperforms the state-of-the-art methods on several challenging datasets, such as our COCO-CapSal test set, PASCAL-S [19] and a recent dataset SOC [6] focused on cluttered scenes.

Our contributions are summarized as follows.

- To the best of our knowledge, this is the first work to explore the usefulness of captioning for salient object detection. And we establish a new dataset, which provides the annotations of salient regions and corresponding captions.
- We propose a new deep neural network model, CapSal, to leverage the captioning information together with the local and global visual contexts for predicting salient regions.
- Extensive experimental results have demonstrated that captioning is indeed effective at promoting the performance of salient object detection, especially in some complicated scenarios.

## 2. Related Work

**Salient Object Detection.** Driven by the remarkable success of CNN, many deep learning models have been proposed for salient object detection. Early methods [31, 16, 15] utilize the CNN features and fully connected layers to predict the saliency scores of image patches. For example, Li *et al.* [17] propose to extract multi-scale contextual CNN features for each superpixel to formulate its saliency probability. In [31], Wang *et al.* put forward two networks for locally estimating salient superpixels and globally searching salient proposals. These methods significantly break the bottleneck of traditional saliency approaches [3, 45, 38, 39]. However, the fully connected layers in the network largely drop the computational efficiency. To address this problem, many attempts [24, 4, 34] have been made to use FCN [25] for generating pixel-wise saliency prediction. Wang *et al.* [33] take the saliency prior to recurrently guide the generation of the final saliency map. In [23], Liu *et al.* first produce a coarse global saliency prediction and refine it by progressively incorporating fine details from lower-level features. In [41], Zhang *et al.* build a bi-directional message passing model for integrating multi-level CNN features. Although impressive results have been achieved, the networks trained only on saliency annotations may not learn sufficient semantic knowledge for handling

extra complicated scenes. To address this problem, we propose a CapSal model, which leverages the high-level object knowledge from captioning to boost the semantic feature learning for salient object detection.

**Image Captioning.** Image captioning aims to generate a syntactically reasonable sentence for describing the image content. Most existing image caption models are beneficial from the CNN+RNN architecture [37, 26, 30, 8], in which the CNN is used to encode the information of image content and RNN is exploited to translate them into caption. Based on the CNN+RNN architecture, top-down visual attention mechanism is introduced to image captioning, which encourages models to selectively focus on the relevant regions described in caption. In [37], Xu *et al.* utilize the LSTM hidden state from last time to formulate the spatial attention, which would be further used for the next word prediction. Recently, a few works [1, 28] have attempted to utilize visual saliency to improve the performance of image captioning models. They use saliency prediction to assist the model to better concentrate on objects of interest. Inspired by the success of these works, we propose to leverage image captioning as an auxiliary task to promote saliency prediction in complicated scenarios.

## 3. Dataset Construction

To train and evaluate our proposed model, we establish a COCO-CapSal dataset, which provides ground truth masks of salient objects and the corresponding image captions. We exploit annotations from two existing datasets, MSCOCO [22] and SALICON [13], to build our dataset. MSCOCO is a challenging real-world dataset, which provides both image captions and instance-level annotations for objects in 80 categories. We take it as a source benchmark to collect images, captions and salient object masks for our COCO-CapSal dataset. SALICON utilizes the mouse clicks to approximate the eye gaze data and provides the human gaze annotations for 15k images of MSCOCO. We conduct two-stage work to build our dataset, which are image selection and saliency ground truth generation. We utilize the human gaze annotation from SALICON to indicate the rough localization of salient regions (see examples in Fig.3 (b)). In the first stage, the image would be selected if (1) its caption descriptions are consistent with salient regions, and (2) the categories of the salient objects are contained in 80 classes of MSCOCO. After this stage, we collect 5265 images for training and 1459 ones for testing. In the second stage, we aim to generate the salient object annotations for the collected images. The image selection strategy in the first stage ensures that the mask of salient object could be collected from MSCOCO dataset. This motivates us to directly use the instance-level annotations from MSCOCO to generate saliency ground truth. For each image, we obtain its object instance masks from MSCOCO and calculate their IoUs with the corresponding human gaze annotation from SALICON. The instances whose IoUs are larger than the mean IoU of the image by 1.5 times would be selected as salient objects. Then their corresponding instance masks are merged to generated the saliency ground truth. With the above-mentioned two stages, we build our COCO-CapSal dataset, which contains 6724 challenging images from real world with well-defined saliency ground truth and caption expressions (see examples in Fig.3).

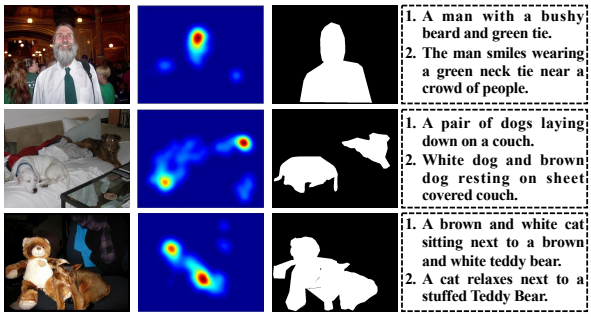



Figure 3: Examples of proposed COCO-CapSal dataset. From left to right: (a) input image, (b) human gaze ground truth from SALICON [13], (c) salient object ground truth of our COCO-CapSal dataset, (d) corresponding image captions from MSCOCO [22].

## 4. CapSal Model

In this paper, we propose a CapSal model, which takes high-level captioning information to bootstrap the learning of semantics for salient object detection. Our CapSal network consists of three components, shared backbone network, Image Captioning Network (ICN) and Local-Global Perception Network (LGPN). The detailed architectures of three sub-networks are shown in Fig.4.

### 4.1. Shared Backbone Network

We use Resnet101 [10] as feature extractor, and remove the last average pooling and fully connected layers to make it fit our task. For an input image $I$ with size $W \times H$, we use the revised Resnet101 to extract features from Res2_x to Res5_x, which are represented as $\mathbf{F} = \{\mathbf{f}_i\}_{i=2}^{5}$ with size $\frac{W}{2^i} \times \frac{H}{2^i}$. The multi-level features contain various information about the salient objects. Features from deeper layers could capture some high-level semantic knowledge, which are beneficial for identifying salient regions. And shallower layers can provide more spatial details about the object boundary. To effectively exploit the multi-level features, we propose to integrate them in a top-down manner:

$$\mathbf{P}_i = \begin{cases} \text{ReLU}(\mathbf{W}_{\mathbf{f},i}\mathbf{f}_i + \mathbf{b}_{\mathbf{f},i}) + \text{Up}(\mathbf{P}_{i+1}), i = 2, 3, 4 \\ \text{ReLU}(\mathbf{W}_{\mathbf{f},i}\mathbf{f}_i + \mathbf{b}_{\mathbf{f},i}), i = 5 \end{cases} \quad (1)$$

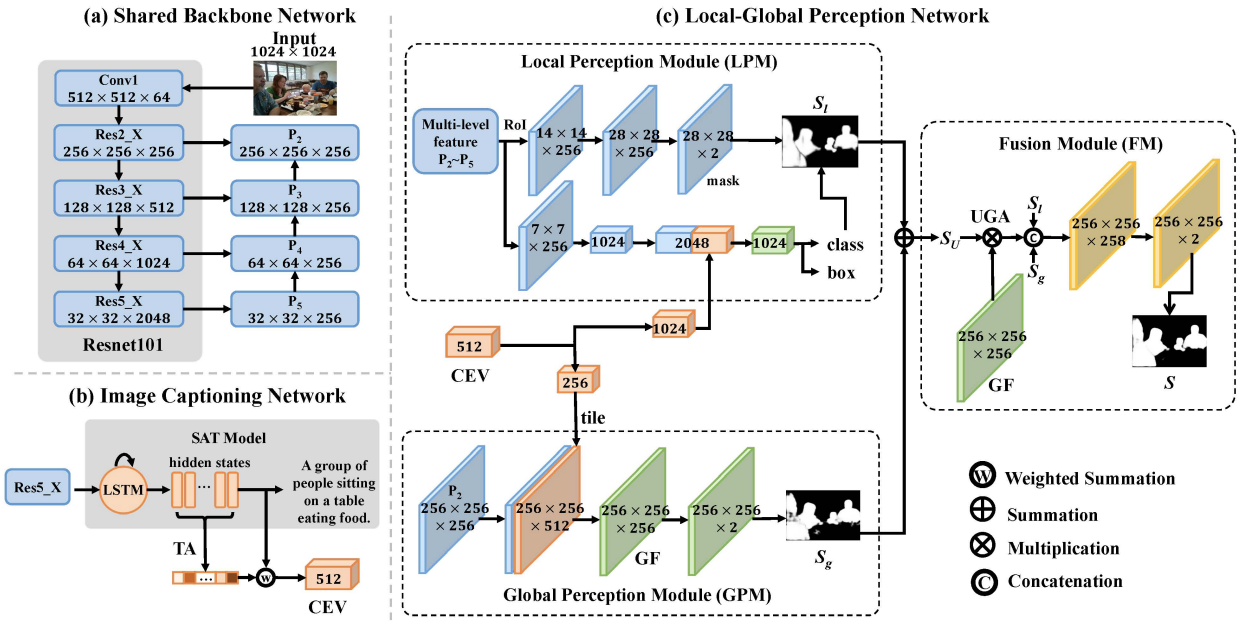



Figure 4: Network Overview. (a) Details of the shared backbone network. (b) Details of Image Captioning Network (ICN). SAT model is built in the same architecture with [37]. CEV and TA represent Caption Embedded Vector and Textual Attention. (c) Details of the three modules in Local-Global Perception Network (LGPN). GF is Global Feature map in GPM, UGA means Union Guided Attention in FM. $\mathbf{S}_l$, $\mathbf{S}_g$, $\mathbf{S}_U$ and $\mathbf{S}$ are local perception saliency map, global perception saliency map, union saliency map and final saliency map, respectively.

where $\mathbf{W}_{\mathbf{f},i}$ and $\mathbf{b}_{\mathbf{f},i}$ are parameters of the convolutional layer. Up() denotes the up-sampling operation. Features in the shared backbone network will be used in the subsequent ICN and LGPN for image captioning generation and salient object detection, respectively.

### 4.2. Image Captioning Network

We leverage the recent advances in image captioning task to embed the object-level information from caption. We exploit a CNN+LSTM captioning network, the SAT [37] model, to generate the Caption Embedded Vector (CEV) from input image. We take Res5_x (*i.e.,* $\mathbf{f}_5$) as the input of our image captioning network and use the hidden vectors of LSTM at $T$ steps $\{\mathbf{h}_t\}_{t=1}^{T}$, $\mathbf{h}_i \in \mathbb{R}^n$ to represent the embedding features of the generated words. Considering that not every word in the caption is equally important for describing the objects, we propose a Textual Attention mechanism (TA) to distill the caption to obtain more essential information. Specifically, we use two fully connected layers to compute the attention scores $\{\boldsymbol{\alpha}_t\}_{t=1}^{T}$ for $T$ generated words:

$$\mathbf{u}_t = \mathbf{W}_{\mathbf{u}}(\tanh(\mathbf{W}_{\mathbf{h}}\mathbf{h}_t)) + \mathbf{b}_{\mathbf{u}} \tag{2}$$

$$\boldsymbol{\alpha}_t = \frac{\exp(\mathbf{u}_t)}{\exp(\sum_{t=1}^{T}\mathbf{u}_t)} \tag{3}$$

where $\mathbf{W}_{\mathbf{u}}$, $\mathbf{W}_{\mathbf{h}}$ and $\mathbf{b}_{\mathbf{u}}$ are the parameters of the fully connected layers and $\sum_{t=1}^{T}\boldsymbol{\alpha}_t = 1$. The attention score $\boldsymbol{\alpha}_t$ reflects the importance of the corresponding word $t$, and the caption embedded vector is obtained via a weighted sum of the hidden states:

$$\mathbf{c} = \sum_{t=1}^{T}\boldsymbol{\alpha}_t\mathbf{h}_t \tag{4}$$

The caption embedded feature vector $\mathbf{c} \in \mathbb{R}^n$ is able to encode the overall semantic knowledge of the salient object. We exploit it to boost the semantics of visual features for localizing salient object from complex clutters.

### 4.3. Local-Global Perception Network

Contextual information has shown its effectiveness in salient object detection [31, 18, 17]. Larger context could capture the global structure of object and provide holistic estimation of the salient regions. While smaller context focuses on the local part of the object and is capable of retaining more spatial details. We propose a Local-Global Perception Network (LGPN), which incorporates the caption embedded vector with multi-contextual visual features for saliency prediction. The LGPN contains three components, Local Perception Module (LPM), Global Perception Module (GPM) and Fusion Module (FM). The detailed architecture of each module is shown in Fig.4.

**Local Perception Module.** Previous works [31, 18, 17] tend to exploit superpixels to extract the local information of salient objects. While these segments destroy the spatial consistence of salient regions and make the models fail to

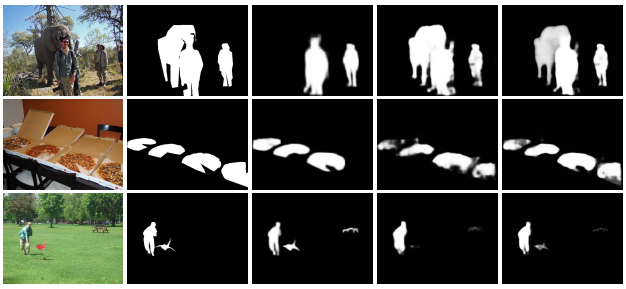
uniformly highlight the object interior. To avoid this problem, we propose a Local Perception Module (LPM), which exploits bounding box to capture the local context for localizing and segmenting salient regions. We employ the Mask-RCNN [9] to implement our LPM. Given an image, the Mask-RCNN first uses Region Proposal Network (RPN) to produce a set of candidate RoIs (*i.e.,* bounding boxes). Then two parallel networks are designed for bounding box recognition (denoted as $\phi_{recog}$ ) and object mask segmentation (denoted as $\phi_{mask}$).

We exploit the Mask RCNN to generate the saliency probability and object mask for each candidate box. To effectively utilize the multi-level CNN feature, we build the Mask-RCNN on top of Feature Pyramid Network (FPN) [21]. Specifically, we apply the RPN and RoIAlign on the integrated multi-level features $\{\mathbf{P}_i\}_{i=2}^{5}$ to produce candidate boxes $\{B_i\}_{i=1}^{N_B}$ and their corresponding feature maps $\{\mathbf{f}_{B,i}\}_{i=1}^{N_B}$. The feature vector before the final classification layer is used to represent the local context of each bounding box (defined as $\tilde{\mathbf{f}}_{B,i}$). In LPM, we utilize the high-level semantic information from captioning to boost the classification of bounding box. We integrate the caption embedded vector $\mathbf{c}$ with bounding box context $\tilde{\mathbf{f}}_{B,i}$ as follows:

$$\mathbf{l}_i = \text{ReLU}(\mathbf{W_{B,c}}(\text{Cat}(\tilde{\mathbf{f}}_{B,i}, \mathbf{c}_{\uparrow}) + \mathbf{b_{B,c}}) \tag{5}$$

where $\mathbf{W_{B,c}}$ and $\mathbf{b_{B,c}}$ are convolutional parameters, Cat() represents cross-channel concatenation operation. $\mathbf{c}_{\uparrow}$ is the CEV after dimension augmentation via $1 \times 1$ convolution. The feature vector $\mathbf{l}_i$, which incorporates both caption semantics and visual cues, is further processed with two fully connected layers to produce candidate saliency probability and box regression. Then the bounding boxes whose class probabilities are larger than a fixed threshold $\theta_T$ would be chosen as salient candidates. Their corresponding object masks are mapped to the original location in the image to generate the local perception saliency map $\mathbf{S}_l$. In order to obtain the saliency probability map, we here do not binarize the object mask like [9]. By capturing local appearance of object with bounding box, the LPM is capable of uniformly highlighting the interior of the salient objects and retaining some fine details (as shown in Fig.5 (c)). However, the LPM may bring some mistaken detection results due to the lack of enough global contextual information. Thus we also propose a global perception module to distinguish salient regions by considering more global contexts.

**Global Perception Module.** Global context is an effective cue to give a convincing estimation of the salient regions. We propose a Global Perception Module (GPM), which incorporates the caption embedded vector with global visual context, to give a precise localization of salient object. We take feature map $\mathbf{P}_2$ with resolution $\frac{W}{2^2} \times \frac{H}{2^2}$ as visual representation of GPM. Different from LPM, which uses caption embedding to assist the classification of bounding box, GPM incorporates the caption embedded vector $\mathbf{c}$ with visual feature $\mathbf{P}_2$ in a pixel-wise manner,



**(a)** **(b)** **(c)** **(d)** **(e)**

Figure 5: Saliency maps generated in LGPN. (a) Input image, (b) ground truth, (c)-(e) saliency map generated by LPM, GPM and FM.

$$\mathbf{g} = \text{ReLU}(\mathbf{W_{p,c}}(\text{Cat}(\mathbf{P}_2, \text{tile}(\mathbf{c}_{\downarrow})) + \mathbf{b_{p,c}}) \tag{6}$$

where $\mathbf{W_{p,c}}$ and $\mathbf{b_{p,c}}$ are weight and bias of convolutional layer. $\mathbf{c}_{\downarrow}$ is the 256-dimensional CEV. tile() is to tile the caption vector $\mathbf{c}_{\downarrow}$ into feature map of size $\frac{W}{2^2} \times \frac{H}{2^2} \times 256$. The obtained Global Feature map (GF) (denoted as $\mathbf{g}$) is capable of integrating both semantic knowledge from captioning and visual information. We process the global feature map with a convolutional layer and sigmoid function to generate the saliency probability for each pixel,

$$\mathbf{S}_g = \text{Sigmoid}(\mathbf{W_g g} + \mathbf{b_g}) \tag{7}$$

where $\mathbf{W_g}$ and $\mathbf{b_g}$ are parameter of convolutional layer for predicting global perception saliency map $\mathbf{S}_g$.

**Fusion Module.** As above-mentioned, we propose the LPM and GPM to integrate high-level caption embedding with visual features for saliency inference. By capturing the local appearance of salient object, the LPM is capable of uniformly detecting the interior of salient object and retaining some fine details. On the other hand, the GPM is able to give a promising estimation of the saliency localization by considering more global contexts. Saliency maps generated by LPM and GPM are complementary (see Fig.5). We propose to combine them to produce the final saliency map. An intuitive fusion method is to concatenate two saliency maps and use a convolutional layer to learn their combination weight. However, without the prior information about salient object, some common mistakes in both saliency maps may not be avoided. To address this problem, we propose an effective Fusion Module (FM) which utilizes CNN feature map as prior information for facilitating the learning of combination weight. To strengthen the feature of salient regions, we first propose a Union Guided Attention mechanism (UGA), in which the union of two saliency maps is exploited as spatial attention map. Then the feature map after UGA is concatenated with local, global perception saliency maps for producing the final result. Specifically, we take the global feature map $\mathbf{g}$ from GPM as input

and the fusion process is conducted by:

$$\mathbf{S} = \text{Sigmoid}(\mathbf{W}_s(\text{Cat}(\mathbf{S}_l, \mathbf{S}_g, (\mathbf{g} \odot \mathbf{S}_U))) + \mathbf{b}_s) \quad (8)$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are the parameters of the combination convolutional layer. $\odot$ represents element-wise multiplication. $\mathbf{S}_U = \mathbf{S}_l + \mathbf{S}_g$ is the union of local and global perception saliency maps. And $\mathbf{S}$ represents the final saliency map.

In the training process, we propose a multi-task loss for jointly optimizing ICN and LGPN:

$$L = L_L + L_G + L_F + \lambda L_C \quad (9)$$

$L_L$ is the loss of LPM, which has the same definition as Mask RCNN [9]. $L_G$ indicates the loss for GPM, which is defined as the cross entropy loss between the global perception saliency map $\mathbf{S}_g$ and ground truth. Similarly, $L_F$ is formulated as the cross entropy loss between final saliency map $\mathbf{S}$ and ground truth. $L_C$ is the loss for ICN with the same definition of SAT model [37] and $\lambda$ represents the trade-off between losses for ICN and LGPN. During inference, our CapSal model can simultaneously produce the caption as well as saliency map for each input image.

# 5. Experiment

## 5.1. Experimental Setup

**Dataset.** We utilize our proposed COCO-CapSal dataset as well as other five saliency datasets to evaluate the performance of our model. The **COCO-CapSal** dataset presented in Sec.3 has 5265 images for training and 1459 ones for testing. The **PASCAL-S** dataset [19] contains 850 challenging images selected from the PASCAL VOC 2009 segmentation dataset. **DUTS** [32] is a large-scale dataset, containing 10553 images for training and 5019 images for testing. Salient objects in DUTS always have various locations and scales. **HKU-IS** [17] has 4447 images with multiple salient objects and low-color contrast. **THUR** [2] includes 6,232 images with categories of butterfly, coffee, dog, giraffe and plane. **SOC** [6] is a new-built dataset, which includes 3000 images selected from MSCOCO [22] and 3000 ones with non-salient object. We exploit the images with salient objects from the SOC validation set to evaluate our method.

**Evaluation Criteria.** To evaluate our CapSal model as well as other state-of-the-arts, we use four common metrics in salient object detection, including Precision-Recall (PR) curves, F-measure, S-measure [7] and Mean Absolute Error (MAE). By binarizing the predicted saliency map with thresholds in [0,255], a sequence of precision and recall pairs are calculated for each image of the dataset. The PR curve is plotted using the average precision and recall of the dataset at different thresholds. We also use F-measure to obtain an overall performance evaluation. It is computed as $F_\beta = \frac{(1+\beta^2)\times Precision \times Recall}{\beta^2 \times Precision + Recall}$, where $\beta^2$ is 0.3 to weight precision more than recall [39]. We report the $F_\beta$ which is calculated by thresholding the saliency map with its twice mean saliency score. Except for PR curve and F-measure, we also report the MAE and S-measure [7] to provide an overall evaluation.

**Implementation Details.** We utilize the training set of COCO-CapSal dataset to train our proposed model on a PC with GTX 1080Ti GPU. The parameters of the shared backbone network is initialized by Resnet101 pretrained on MSCOCO [22]. Our LPM shares the same architecture and parameter settings with Mask RCNN [9]. We design a two-stage training strategy to facilitate the convergence of our CapSal model. First, we train the ICN using caption data of COCO-CapSal. In this stage, the shared backbone network and LGPN are fixed without training. We utilize SGD optimizer with learning rate 0.001 to train ICN until it converges. In the second stage, two sub-networks and shared backbone are jointly optimized using the multi-task loss defined in Eq.9. The trade-off $\lambda$ is set to 0.1. The SGD optimizer with learning rate 0.0001 is exploited for the training of both LGPN and ICN. The weight decay and momentum are set to 1e-4 and 0.9 in both stages. In our experiment, the input images are resized and padded into $1024 \times 1024$. During inference, the threshold $\theta_T$ in LPM is set to 0.8.

## 5.2. Comparison with State-of-the-arts

We compare the proposed CapSal model with 11 deep learning methods, including LEGS [31], MDF [17], RFCN [33], DCL [18], DHS [23], NLDF [27], DSS [11], Amulet [42], UCF [43]), BMPM [41] and DGRL [35]. The saliency maps of different methods are published by the authors or achieved by running available codes.

**Quantitative Evaluation.** We compare our CapSal model with other 11 methods in terms of PR curves, F-measure and MAE. The comparison results in Fig.6 and Tab.1 consistently demonstrate our model largely outperforms other approaches on challenging COCO-CapSal, PASCAL-S [19] and SOC [6] datasets, and performs comparably on DUTS-test [32], THUR [2] and HKU-IS [17] datasets. We also provide the S-measure results of three datasets in Tab. 2, which also verify the effectiveness of our model. The PR curve on THUR dataset is provided in supplementary material. The MDF [17] use HKU-IS dataset for training, we do not report its result on this dataset.

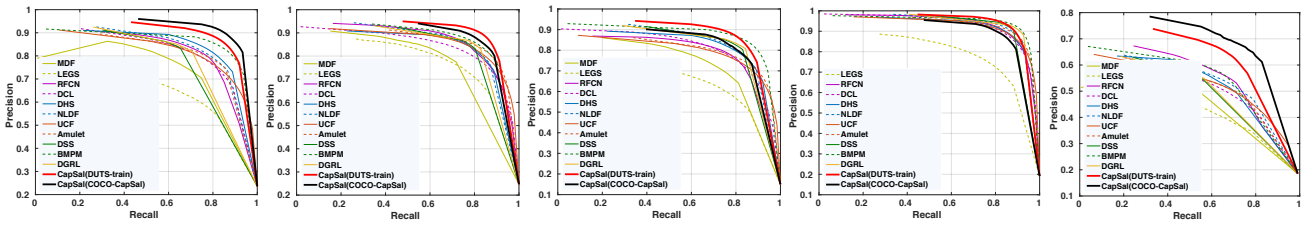
**Qualitative Evaluation.** To qualitatively estimate the performance of our CapSal model, we show some visual examples generated by our method and other 11 approaches in Fig.7. We can observe that our method can accurately detect salient objects from complicated background.

## 5.3. Ablation Study

In this section, we analyze the contribution of each component in our CapSal model. The results on COCO-CapSal

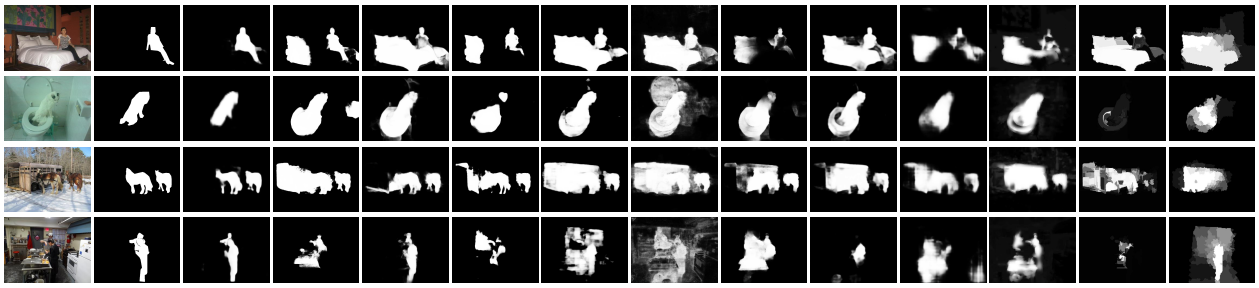
Table 1: Quantitative comparisons with other state-of-the-arts in term of F-measure and MAE on six datasets. The best three results are shown in red, green and blue. "CapSal(DUTS-train)" and "CapSal(COCO-CapSal)" represent the result of our CapSal model trained on DUTS-train [32] and COCO-CapSal datasets.

| Method | COCO-CapSal | | PASCAL-S | | DUTS-test | | HKU-IS | | THUR | | SOC-val | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F_\beta$ | MAE | $F_\beta$ | MAE | $F_\beta$ | MAE | $F_\beta$ | MAE | $F_\beta$ | MAE | $F_\beta$ | MAE |
| LEGS [31] | 0.594 | 0.187 | 0.697 | 0.155 | 0.584 | 0.138 | 0.732 | 0.119 | 0.607 | 0.125 | 0.445 | 0.216 |
| MDF [17] | 0.665 | 0.152 | 0.709 | 0.146 | 0.673 | 0.100 | - | - | 0.636 | 0.109 | 0.409 | 0.168 |
| RFCN [33] | 0.754 | 0.127 | 0.751 | 0.133 | 0.712 | 0.090 | 0.835 | 0.089 | 0.627 | 0.100 | 0.531 | 0.159 |
| DCL [18] | 0.730 | 0.108 | 0.714 | 0.125 | 0.714 | 0.149 | 0.853 | 0.136 | 0.676 | 0.161 | 0.480 | 0.177 |
| DHS [23] | 0.768 | 0.097 | 0.773 | 0.095 | 0.724 | 0.067 | 0.852 | 0.054 | 0.673 | 0.082 | 0.519 | 0.135 |
| UCF [43] | 0.662 | 0.145 | 0.701 | 0.127 | 0.629 | 0.117 | 0.808 | 0.074 | 0.645 | 0.112 | 0.428 | 0.238 |
| Amulet [42] | 0.751 | 0.102 | 0.763 | 0.098 | 0.678 | 0.085 | 0.839 | 0.052 | 0.670 | 0.094 | 0.497 | 0.169 |
| NLDF [27] | 0.754 | 0.107 | 0.779 | 0.099 | 0.743 | 0.066 | 0.874 | 0.048 | 0.700 | 0.080 | 0.500 | 0.158 |
| DSS [11] | 0.742 | 0.133 | 0.804 | 0.096 | 0.791 | 0.057 | 0.895 | 0.041 | 0.731 | 0.074 | 0.493 | 0.151 |
| BMPM [41] | 0.741 | 0.079 | 0.769 | 0.074 | 0.750 | 0.049 | 0.871 | 0.038 | 0.690 | 0.079 | 0.500 | 0.134 |
| DGRL [35] | 0.780 | 0.118 | 0.825 | 0.072 | 0.768 | 0.051 | 0.882 | 0.037 | 0.716 | 0.077 | 0.495 | 0.135 |
| CapSal (DUTS-train) | 0.815 | 0.065 | 0.830 | 0.064 | 0.789 | 0.044 | 0.878 | 0.039 | 0.728 | 0.069 | 0.604 | 0.105 |
| CapSal (COCO-CapSal) | 0.860 | 0.057 | 0.823 | 0.075 | 0.756 | 0.063 | 0.836 | 0.059 | 0.711 | 0.081 | 0.631 | 0.117 |




COCO-CapSal PASCAL-S DUTS-test HKU-IS SOC

Figure 6: Comparisons of the proposed approach and 11 methods on five datasets in terms of PR curves.



Input GT Ours DGRL BMPM DSS Amulet UCF NLDF DHS DCL RFCN MDF LEGS

Figure 7: Qualitative comparisons of the proposed method and state-of-the-art algorithms. Our saliency map is the result of CapSal model trained on COCO-CapSal dataset.

Table 2: The S-measure results on three datasets.

| | COCO-CapSal | PASCAL-S | SOC-val |
|---|---|---|---|
| DSS [11] | 0.726 | 0.797 | 0.602 |
| BMPM [41] | 0.832 | 0.845 | 0.656 |
| DGRL [35] | 0.740 | 0.836 | 0.597 |
| CapSal (DUTS-train) | 0.846 | **0.857** | 0.705 |
| CapSal (COCO-CapSal) | **0.868** | 0.837 | **0.710** |

and DUTS-test [32] datasets are shown in Tab.3.

**Analysis of LGPN.** We take only visual features as the input of LGPN to predict saliency map and regard this model as our baseline network. The comparison results in Tab.3 demonstrate LPM, GPM and FM all contribute to the generation of saliency map. From the visual comparisons in Fig.5, we can see that LPM and GPM are complementary and combining them using FM can achieve a better result.

**Effectiveness of captioning on LGPN.** We investigate the effectiveness of our CapSal model by comparing it with baseline network. The quantitative results in Tab.3 verify the efficacy of captioning in promoting the performance of LPM, GPM and their final fusion. From the visual examples in Fig.8, we can observe that captioning is helpful for accurately localizing the salient regions in some complicated scenarios. To demonstrate the effectiveness of textual attention in ICN, we remove this part and use the last hidden

Table 3: Results of ablation studies of CapSal network on COCO-CapSal and DUTS-test datasets [32]. LGPN: local-global perception network, ICN: image captioning network, LPM: local perception module, GPM: global perception module, TA: textual attention, JT: joint training.

| Model Setting | COCO-CapSal | | DUTS-test | |
|---|---|---|---|---|
| | $F_\beta$ | MAE | $F_\beta$ | MAE |
| Analysis of LGPN (Our baseline) | | | | |
| LPM | 0.813 | 0.067 | 0.698 | 0.081 |
| GPM | 0.786 | 0.071 | 0.701 | 0.074 |
| LPM+GPM | 0.821 | 0.063 | 0.717 | 0.072 |
| Effectiveness of captioning on LGPN | | | | |
| LPM+ICN (w/o JT) | 0.830 | 0.064 | 0.713 | 0.074 |
| GPM+ICN (w/o JT) | 0.821 | 0.064 | 0.713 | 0.070 |
| LGPN+ICN (w/o JT) | 0.843 | 0.062 | 0.720 | 0.069 |
| LPM+ICN (w/o TA) | 0.822 | 0.065 | 0.715 | 0.075 |
| GPM+ICN (w/o TA) | 0.818 | 0.069 | 0.719 | 0.069 |
| LGPN+ICN (w/o TA) | 0.837 | 0.063 | 0.725 | 0.067 |
| LPM+ICN | 0.844 | 0.060 | 0.730 | 0.063 |
| GPM+ICN | 0.834 | 0.060 | 0.731 | 0.063 |
| LGPN+ICN | 0.860 | 0.057 | 0.756 | 0.063 |
| Influence of the captioning accuracy | | | | |
| LPM+GT caption | 0.849 | 0.056 | - | - |
| GPM+GT caption | 0.839 | 0.057 | - | - |
| LGPN+GT caption | 0.866 | 0.055 | - | - |

state of LSTM as caption embedded vector (denoted as "w/o TA"). The results in Tab.3 and Fig.8 show that TA can emphasize the words about salient object and contributes to the final result of LGPN. We also verify the efficacy of our joint training strategy. We use the hidden states of ICN pretrained in the first stage to generate fixed caption embedded vector and only update the LGPN in the second training stage. The results in Tab.3 prove that the joint training of LGPN and ICN can bring a better performance.

**Efficacy on other saliency training dataset.** To demonstrate the effectiveness of captioning on other training data, we utilize the caption embedding from the first stage ICN and train our LGPN on DUTS-train dataset [32]. The comparison with 11 state-of-the-art (see "CapSal(DUTS-train)" in Tab.1) verifies the generalization of our CapSal model on other training dataset.

**Influence of the captioning accuracy.** To verify the influence of caption's accuracy on saliency detection, we process the ground truth caption with embedding layer and LSTM for producing caption embedded vector. The results are also listed in Tab.3. Note that other saliency datasets do not contain caption data, this experiment is only conducted on our COCO-CapSal dataset. It can be seen that using caption with higher accuracy could achieve a better saliency detection performance.

Table 4: Image captioning results of ICN on COCO-CapSal testing set. "Baseline" and "Joint Training" indicate the captioning results without/with jointly training with LGPN.

| Method | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|
| Baseline | 0.286 | 0.242 | 0.527 | 0.874 |
| Joint Training | 0.291 | 0.245 | 0.530 | 0.903 |

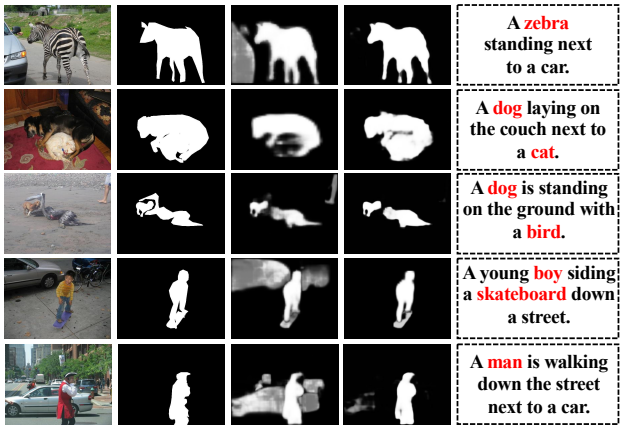



Figure 8: Visual comparison between baseline and CapSal model. (a) Input image, (b) ground truth, (c)-(d) saliency maps of baseline and CapSal model, (e) generated caption by ICN. Word with higher attention score is colored in red.

**The performance of ICN.** We also investigate the performance of ICN for image captioning. We report the results on COCO-CapSal dataset in terms of BLEU-4 [14], METEOR [5], ROUGE-L [20] and CIDEr [29] in Tab.4. We use the pretrained ICN in the first training stage as baseline. The comparison results demonstrate that the performance of ICN could be improved by jointly training with LGPM.

## 6. Conclusion

We propose a CapSal model, which utilizes image captioning to boost the semantic feature learning for salient object detection. We first design an Image Captioning Network (ICN) to embed the semantic knowledge of caption. Then a Local-Global Perception Network (LGPN) is proposed to incorporate caption embedding with local and global contexts for saliency inference. The ICN and LGPN are jointly trained with a multi-task loss. Experiments on six datasets verify the effectiveness of image captioning in promoting salient object detection.

**Acknowledgements.** This work was supported by the Natural Science Foundation of China under Grant 61725202, 61751212 and 61829102 and gifts form Adobe. We thank Xiaohui Shen for the helpful discussion in the early stage of the work.

## References


[1] S. Chen and Q. Zhao. Boosted attention: Leveraging human attention for image captioning. In *Proceedings of European Conference on Computer Vision*, 2018.

[2] M. M. Cheng, N. J. Mitra, X. Huang, and S. M. Hu. Salientshape: group saliency in image collections. *Visual Computer*, 30(4):443–453, 2014.

[3] M. M. Cheng, G. X. Zhang, N. J. Mitra, and X. Huang. Global contrast based salient region detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[4] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng. $R^3$Net: Recurrent residual refinement network for saliency detection. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2018.

[5] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[6] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of European Conference on Computer Vision*, 2018.

[7] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[8] J. Gu, J. Cai, G. Wang, and T. Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In *AAAI Conference on Artificial Intelligence*, 2018.

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[12] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[13] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[14] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 2007.

[15] G. Lee, Y. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[16] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[17] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[18] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[20] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of ACL 2004 Workshop on Text Summarization Branches Out*, 2004.

[21] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision*, 2014.

[23] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[24] N. Liu, J. Han, and M.-H. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[26] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[27] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, , and P. M. Jodoin. Non-local deep features for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[28] H. R. Tavakoliy, R. Shetty, A. Borji, and J. Laaksonen. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[29] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[31] L. Wang, H. Lu, X. Ruan, and M.-H. Yang. Deep networks for saliency detection via local estimation and global search.

In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[32] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[33] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *Proceedings of European Conference on Computer Vision*, 2016.

[34] 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 International Conference on Computer Vision*, 2017.

[35] 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 IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[36] H. Xiao, J. Feng, G. Lin, Y. Liu, and M. Zhang. Monet: Deep motion exploitation for video object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Computer Science*, pages 2048–2057, 2015.

[38] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[39] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[40] K. Yun, Y. Peng, D. Samaras, and G. J. Zelinsky. Studying relationships between human gaze, description, and computer vision. In *Computer Vision and Pattern Recognition*, pages 739–746, 2013.

[41] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang. A bi-directional message passing model for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[42] 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 International Conference on Computer Vision*, 2017.

[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 International Conference on Computer Vision*, 2017.

[44] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[45] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.