Pattern Recognition 000 (2019) 106955

Contents lists available at ScienceDirect

Pattern Recognition

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

Language-aware weak supervision for salient object detection

Mingyang Qian, Jinqing Qi, Lihe Zhang, Mengyang Feng, Huchuan Lu*

School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116024, China

ARTICLE INFO

Article history:
Received 20 March 2019
Revised 26 May 2019
Accepted 26 June 2019
Available online xxx

Keywords:
Saliency detection
Natural language
Textual-visual pairwise
Self-supervision

ABSTRACT

Natural Language Processing has achieved remarkable performance in multitudinous computer tasks, but the potential capability of textual information has not been completely explored in visual saliency detection. In this paper, we learn to detect salient object from natural language by addressing the two essential issues: finding a semantic content matching the corresponding linguistic concept and recovering fine details without any pixel-level annotations. We first propose the Feature Matching Network (FMN) to explore the internal relation between the linguistic concept and visual image in the semantic space. The FMN simultaneously establishes the textual-visual pairwise affinities and generates a language-aware coarse saliency map. to refine the coarse map, the Recurrent Fine-tune Network (RFN) is proposed to enhance its predicted performance progressively by self-supervision. Our approach only leverages the caption to provide important cues of salient object, but generates a fine-detailed foreground map at a detecting speed of 72 FPS without any post-processing. Extensive experiments demonstrate that our method takes full advantage of textual information of natural language in saliency detection, and performs favorably against state-of-the-art approaches on the most existing datasets.

1. Introduction

Saliency detection [1,2], which aims to capture the important instance or region in the image, has received much attention in recent years driven by the deep neural networks [3]. Many supervised saliency methods can efficiently highlight a distinct object with accurate boundaries using pixel-level ground truth. However, the work for annotating each pixel is time-consuming and arduous, which needs a great deal of vigor and labor to create a large-scale dataset. To alleviate this situation, there has been a recent keen interest in weakly supervision using image-level tags, like labels or phrases. Most existing weakly detection methods consider the high-level convolutional features as the important saliency detectors, and integrate the semantic feature maps to extract class-aware visual representations. Using these class-aware representations that distill information down to the salient objects is one of the effective solution in saliency detection [4]. However, these tags are limited in the amount of information and they have to depend on the agnostic semantic meanings learned from DNNs, resulting in uncontrollable object prediction and incomplete coverage of foreground areas.

Despite that image-level tags indicate the presence or absence of object categories in the image level, they cannot effectively offer assistance for the network to better predict full-extent saliency map in details. Rather than being fixed on the image categories, natural language for image (i.e. captions) is a high-level global concept and provides rich saliency cues, including location and appearance. Some deep captioning models [5] have succeeded in learning visual representations to translate the input image into natural language, but they do not further discover the potential relation with visual saliency. Therefore, inspired by the weakly supervised structures using tag knowledge in dealing with pixel-level information, we want to utilize the potential contextual information from the natural language to measure the dominant visual contents in image, supervising the detection network for better performance. Although Ramanishka et al. [6] have already explored the caption guided saliency detection, they make a pioneer work in video saliency detection with an end-to-end model, but only produce spatial or spatiotemporal heat maps for each input caption. Furthermore, we want to extract a highlighted salient region with fine-detailed boundaries, by exploring the potential relationship of each feature representation from static visual image and corresponding natural language (shown in Fig. 1).

To bridging cross-model gap of different modalities, previous approaches try to find a good metric that accurately measures the representations from different modalities with low dimensional vectors, and their distances/similarities reflect their semantic relations. Sound source localization is handled via learning the correspondence between visual scene and the sound, while cross-modal

* Corresponding author.
E-mail address: lhchuan@dlut.edu.cn (H. Lu).

https://doi.org/10.1016/j.patcog.2019.06.021
0031-3203/

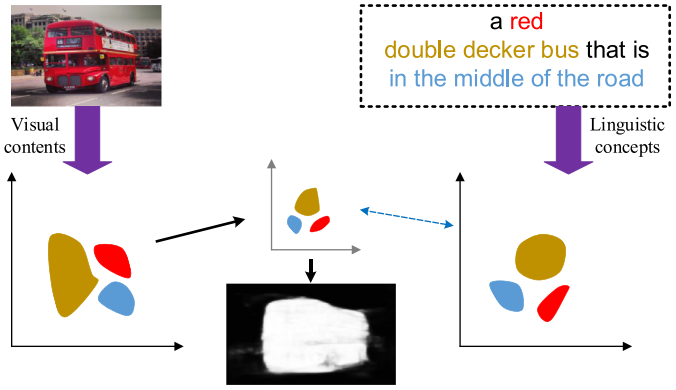


Fig. 1. The illustrative framework of our approach: the semantic visual contents and the contextual linguistic concept share similar feature distribution, which guides the detection for salient object.

retrieval finds a low-dimensional latent common space where multi-modal data can be compared directly. Although they show similar views in aligning features of two modalities, they sometimes dominate the global representation to describe the superficial level information of an image about where and what the document or source contains. More importantly, our method goes a further step that we learn to detect salient objects in the case of limited textual information and generate the finer foreground saliency map with detailed edges. Instead of addressing this difficult task in a simple feedforward network, our approach uses a steady strategy that finding a matching visual content from linguistic descriptions and then refines it by local contextual information. The proposed approach contains two sub-networks: Feature Matching Network (FMN) and Recurrent Fine-tune Network (RFN). By transforming the input image and the corresponding caption into a latent feature space, the FMN is proposed to discover a semantic matching to establish the textual-visual pairwise affinities. This pairwise matching is measured by an objective function that visual feature and linguistic feature belonging to the same specific identity should have a similar feature distribution, thus yielding an initial estimation of the saliency map. In the feedforward processing, the coarse map has already succeeded in locating the corresponding objects described in the language sentence, but fails to preserve enough low-level boundary or texture information. Instead of using common post-processing or handcrafted optimization, we construct a recurrent structure RFN to recover more details of the estimated map, which uses a refinement module to learn by self-supervision. We compare our approach with most existing unsupervised and supervised saliency methods on the large-scale datasets, and the results indicate that our approach captures relatively more accurate regions and detailed boundaries at a faster speed of 72 FPS. The flexibility of our framework also make it possible to be transformed into dense models and achieve better performance in the future.

The contributions of this paper are summarized as follows:

- We first design a language-aware saliency detection framework and clearly demonstrate that with the textual information from the natural language, the network can also be robust and accurate to describe the visual object and generate a fine-detailed saliency map without any pixel-level annotations.
- We propose a novel Feature Matching Network to establish the textual-visual pairwise affinities for explaining the internal relation between language and image, which provides an important saliency prior for detection.
- We leverage a self-supervision mechanism to refine the fine-tune network progressively and the results demonstrate strong competitiveness against existing supervised methods.

2. Related work

2.1. Saliency detection

Detection research has been going on for many years, some traditional algorithms [7,8] are successfully applied to detect generic salient objects. However, the breakthrough of improving saliency performance occurs after widely employing the deep learning models. Early methods like MDF [9], MCDL [10], LEGS [2] and so on, mainly focus on aggregating low-level localize features with high-level semantic meanings to maintain visible improvement. They act on small patches and incorporate multiply CNN features to enhance the spatial coherence of the saliency results with a high computational overhead. Recently, Hou et al. [11] propose to introduce short connections into the skip-layer structures within a hierarchical architecture, while Li et al. [12] integrate multi-scale combinatorial grouping by dense connections between three collateral deep networks, generating very promising object instance results. Other approaches like [13,14] use attention mechanism to select informative contents or connectivity to preserve additional attentive details in saliency, generating more accurate and meticulous saliency maps. Specially, interactive learning between different visual tasks provides abundant ideas for saliency detection. For exploring RGB-D saliency detection, Hao et al. [15] propose a novel multi-scale multi-path fusion network with cross-modal interactions, which enables sufficient and efficient fusion with RGB and depth by combining their deep representations in a late stage with only one path. Ji et al. [16] propose a bottom-up framework for salient object detection by both considering objectness and low-level features under the Graph-based Manifold Ranking framework.

However, supervised saliency detection utilizes pixel-level ground truth to guarantee the optimization for each deep neural layers, which may result in a great deal of computation. For reducing training complexity and correcting predicted errors, Wang et al. [17] incorporate a coarse prediction as input and refine the generated predictions recurrently with a succinct and efficient encoder-decoder network. It provides an inspiration for exploring an adaptive learning to replace redundant network architecture, and we extend this idea in our detection processing with language sentence.

2.2. Weakly supervised learning

Fully supervised learning paradigm largely depends on the object annotations and is accompanied with heavy computational redundancy on regions feature extraction. Moreover, the supervised learning is quite domain specific and it fails to achieve satisfactory results when we do not have sufficient labeled data for a new task to train a reliable detection model. Compared with fully supervision methods using pixel-level ground truth, semi-supervision [18] or weakly supervision [19] only needs fewer annotations and limited labels. To leverage these weakly labeled data, Zhou et al. [20] utilize class-aware Peak Response Maps to provide an accurate location and fine-detailed instance-level representation, which extracts instance masks even without external information. Fang et al. [21] explore to localize image regions from specific tags with a weakly supervised attention learning mechanism for textual phrase grounding. In [22], the latent SVM framework with a single integrated objective function is adopted to handle two problems: modeling saliency labels of superpixels as hidden variables and involving in a classification term conditioned to the salient object existence variable. These weakly methods share similar strategy that

they attempt to treat the given annotation as an important guidance to measure each semantic content, so as to extract and purify more useful visual contents. Our method shows similar spirit but we explore the important salient prior by establishing the text-image matching in a shared latent space. However, the challenge is that weakly learning is an indirect supervision and can hardly generate a full-extend object region. Typically, supervised approaches can use conditional random fields (CRF) or edge-aware boundary detection methods to fine-tune the output performance, whereas weakly supervised works usually involve time-consuming training strategies, e.g., repeatedly model learning or online proposal selection. In contrast, our approach constructs a self-supervision mechanism that we use a refined saliency map as ground truth back to fine-tune the network, enhancing the detection capability of the network without extra effort of requiring any explicit models. The flexibility of our approach make it possible to extend adaptively in weakly supervised learning.

2.3. Natural language processing

Vision and language are two important aspects in understanding the world. Research endeavor is motivated by breaking the boundaries between the two in image-sentence retrieval [23,24], image captioning [25] and detection. The key to bridging vision-language gap is to learn similarity or the mapping relation that accurately measures the semantic image-sentence similarity, and based on which the semantically similar images and sentences can be properly associated. In [23], image-sentence retrieval is investigated by a collaboration of global representation learning and co-attention learning, which discovers the correlative components and rectifying inappropriate component-level correlation to produce more accurate sentence-level ranking results. To solve cross-modal retrieval, Wu et al. [24] propose an online similarity function learning framework to learn the metric (e.g., modality specific kernels and multiple kernel learning) that can well reflect the cross-modal semantic relation. In the textual-visual matching task, Li et al. [26] propose a latent co-attention mechanism associating the relevant visual regions to each affinitive word, aligning different sentences to make the matching results more robust to sentence structure variations. Other methods use dense captioning models [27] to both localize and describe objects or regions referring to the input language sentence. Although the purposes of these tasks are quite different, most of these methods segment images into numerous patches and utilize attention mechanism or feature similarity to select the tightly-coupled matching pairs between patches and words. However, unlike LSTM decoders for language processing [5] whose each step corresponds to a specific inner word, natural language for saliency is a holistic concept that semantic analysis should be focused on the most discriminative object, such that the text information should come into contact with the global image content. We leverage this visual-textual matching to serve the object detection and the details are discussed in the following.

3. Proposed method

Language-based saliency detection is a high-level matching problem that two essential issues should be taken into consideration: finding a semantic content matching the corresponding linguistic concept and the way to recover fine details without any pixel-level annotations. In this work, we address the two questions by proposing a weakly supervised approach, which leverages cross-model textual-visual matching structure to describe the salient objects, and enhance the prediction accuracy of each pixel by a recurrent structure progressively. In the first stage concentrating on the location information, Feature Matching Network (FMN) in Section 3.1 is proposed to leverage the contextual language feature from the caption, guiding the CNN network to predict a coarse saliency map. Then in the second stage, Recurrent Fine-tune Network (RFN) in Section 3.2, which shares parameters with the FMN, tries to learn fine details assisted with a refinement module through a flexible encoder-decoder structure. More details of network and training are provided as follows.

3.1. Feature matching network

The structure of FMN is illustrated in Fig. 2. We can see that the FMN is composed of two components: a language LSTM branch and a visual CNN branch. The language branch on the right integrates the contextual language feature from the caption by a LSTM-based network, and the visual branch on the left is trained with fully convolutional network (FCN) and inference layers, which takes input image of arbitrary sizes and generate associated visual features. The FMN is trained to map the input image and reference caption into a joint feature space, such that the representations belonging to visual concepts should have the similar characteristic distribution with that of linguistic features.

3.1.1. Language feature integration

Classification-based detection methods attempt to compute the prediction scores of each global categories, but the limitation for each nerve cell to learn the specific object category makes the valid classification points concentrate on the most discriminative part of the visual contents. These works have weakness in capturing complicated matching relation between image and sentence. We want to explore the full-extent of salient object not just the discriminative features in the image, thus LSTM-based language processing method is introduced to analyze what details the reference sentence can convey about the image. For the input caption, each word is encoded into a one-hot vector by an embedding dictionary matrix, and we use a pre-trained bi-directional LSTM (Bi-LSTM) network to integrate all the textual information, which is shown in Fig. 2 right. At each time step, the Bi-LSTM network scans through the embedded word vector in sequence and outputs each hidden state. After seeing the whole caption sequence, the hidden states of both forward and backward directions will be concatenated together, feeding into a max pooling layer to obtain the encoded textual representations $\boldsymbol{h}$. The elements in the $\boldsymbol{h}$ represent the visual or implicit contents in the language sentence, which is considered as a standard characteristic distribution for the visual feature extraction. The intuition is that if the language branch can provide an exhaustive semantic concept, the encoder CNN can be able to describe the representations of image contents with similar characteristic distribution.

3.1.2. Visual feature extraction

Accordingly, deep CNN extracts image features through a hierarchy of visual contents by arranging stacked convolution. The convolutional filter performs as a pattern detector, e.g., lower-layer filters detect low-level visual cues like edges and corners, while higher-level ones detect high-level semantic patterns like parts and objects. However, concatenating multi-layered CNN features for fusing different levels of visual information may lead to limited performance improvement in learning cross-modal correlation due to the disturbance from uninformative feature levels and redundancy of dimensionality [24]. To match the language feature into visual saliency detection, we consider a discriminative feature extraction architecture in Fig. 2 left. Compared with the deep classification network (e.g., VGG), which used to incorporate a series of convolutional (and pooling) layers and fully connection layers to obtain a spatial vector of predicted classification scores. We use the fully convolutional network [28] as our base model for dense

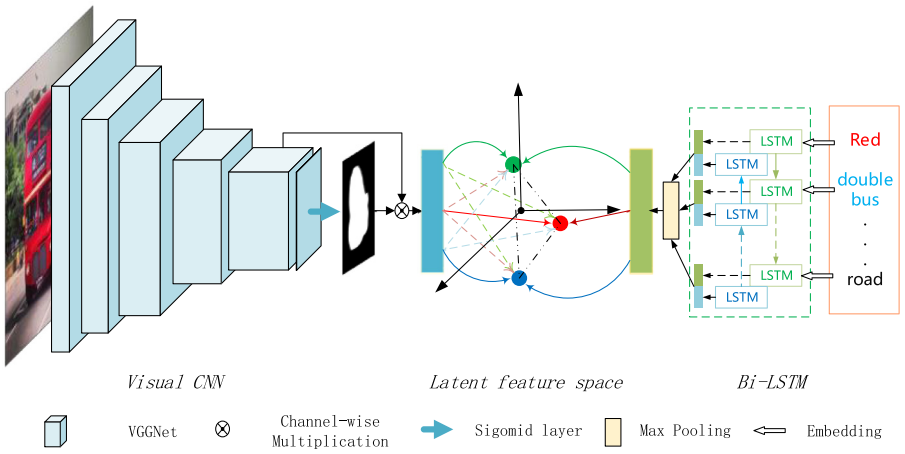


Fig. 2. Overview of the Feature Matching Network (FMN). A latent feature space is set up in the higher semantic level that the visual feature extracted from stacked convolutional layers is tend to match the language feature integrated from the Bi-LSTM. The pivotal layers are shown respectively.

inference with a spatial coordinate. Given an input image $\boldsymbol{I}$ with $W \times H \times 3$, we first feed it into five stacked convolutional blocks, extracting deep semantic meanings with high responses. The input image is finally down-sampled by a factor of 1/32 into $w \times h$ spatial feature maps with D channels, and all the output feature maps are then fed into a convolutional layer to predict an initialized saliency map $\boldsymbol{S}$, by applying a 3×3 kernel with sigmoid activation function. The coarse map $\boldsymbol{S}$ representing global semantic information can be used as a location prior for saliency, but it filters out much low-level visual cues. Before matching the corresponding representations of all the textual concepts, semantic analysis should be done that we use a channel-wise multiplication between those semantic feature maps and coarse saliency map to obtain a series of processed feature maps.

The discriminative parts in these processed maps are weighted with high responses to emphasize the contact with the salient object, and the fully connection layer just integrates all the processed maps into a fixed vector representations $\boldsymbol{v}$, thus extracting the visual feature. The Visual Feature Extraction branch can be formulated as

$$\boldsymbol{S}_i = \boldsymbol{f}_s(\boldsymbol{F}(\boldsymbol{I};\theta);\varphi), \tag{1}$$

$$\boldsymbol{v}_i = \boldsymbol{f}_n(\boldsymbol{S}_i \otimes \boldsymbol{F}(\boldsymbol{I};\theta)), \tag{2}$$

where $\boldsymbol{S}_i$ and $\boldsymbol{v}_i$ are the initial saliency map and the visual feature vector for ith input pair of image and caption; $\boldsymbol{F}(.;\theta)$ denotes the stacked convolutional blocks of FCN parameterized by θ acting on input image $\boldsymbol{I}$; $\boldsymbol{f}_s(.;\varphi)$ denotes the convolutional layer with sigmoid function parameterized by φ that generates the coarse saliency map $\boldsymbol{S}_i$, and $\boldsymbol{f}_n$ denotes the finial fully connected layer with softmax.

3.1.3. Training details

Given ith input image and caption pair, the visual feature $\boldsymbol{v}_i$ and the language feature $\boldsymbol{h}_i$ generated by the two branches are now stored into a shared latent space, each of which enables efficient calculation of textual-visual affinities between the same sampled identity features. We put the two branches into an end-to-end network and train the FMN on the Microsoft COCO 2017 caption evaluation dataset [29], containing 118 k training samples and 5 captions for each image. For each input image, most captioning methods usually average the language feature vectors across all the five captions, such that the information in the captions will be leveraged completely. In contrast, we consider that it is hard to explain the internal processing to predict saliency map with promiscuous contextual information and duplicate defined attributes. Thus, the Bi-LSTM network choose one of the five pairs by the means of random perturbation, output a fixed language vector to train our Feature Matching Network. All the input captions are tailored to 20 words for reducing conditional complexity, where the missing word vectors are filled with 0 and the extra ones are directly ignored. We train the FMN by minimizing the following matching function:

$$\boldsymbol{L}_{fmn} = \frac{1}{N}\sum_{i=1}^{N} -log\frac{exp\left(\boldsymbol{v}_{i,j}.\bar{\boldsymbol{h}}_{i,j}\right)}{\sum_{j=1}^{N} exp\left(\boldsymbol{v}_{i,j}\bar{\boldsymbol{h}}_{i,j} + \epsilon\right)}, \tag{3}$$

where ϵ is a small number to avoid numerical problems, $\bar{\boldsymbol{h}}_{i,j}$ and $\boldsymbol{v}_{i,j}$ denote each element of the normalized linguistic feature and visual feature, and N is a flexible dimension of matching. Inspired by cross-entropy loss, the matching loss enforces the consistence of the distributions of linguistic concepts and visual contents. We solve the loss function using min-batch Stochastic Gradient Descent (SGD) and the batch size is set as 32, and the learning rate is initialized as 1e-2 and decreased by 0.1 for every 20 epochs. For reducing memory consumption, all the input images are first down-sampled to a fixed resolution of 256^2. In each training iteration, the parameters of visual CNN are updated via back propagation, while the Bi-LSTM network is fixed to make sure the language feature value is standard in each iteration. Since the accuracy of features from the language branch may influence the result of visual feature extraction, the Bi-LSTM network is pre-trained on the text-image matching task followed Zhang and Lu [30]. The feature vectors store all feature representations of different modalities, and the matching vector dimension is set to 128 as to enlarge the receptive field of the both visual and language feature. By minimizing the matching entropy, the visual feature has a more similar distribution to linguistic features, thus the FMN can approximately describe what the caption concentrates on and output a language-aware saliency map.

3.2. Recurrent fine-tune network

The language-aware saliency map generated by the FMN can already highlight the discriminative regions in the image. In this stage, attention must be turned to a local view to recover more details, and delineate a full-extent and compact saliency map without any pixel-level ground truth. We take a long-term strategy that a recurrent network RFN (shown in Fig. 3) is designed for continuous optimization without any pixel-level annotations. The RFN is an encoder-decoder network that the encoder part share parameters

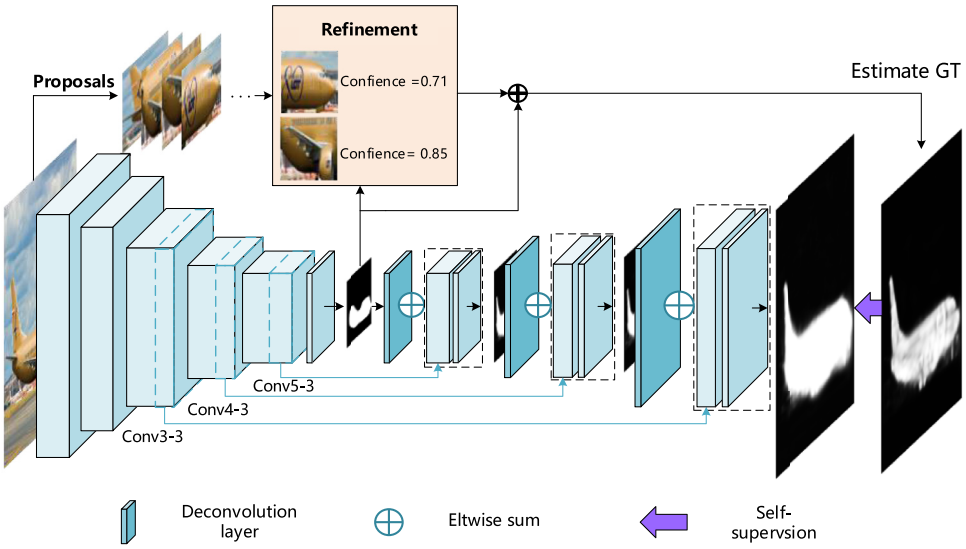


Fig. 3. The overall architecture of Recurrent Fine-tune Network (RFN). The estimated saliency map refined by proposals serves as ground truth back to fine-tune the network by self-supervision.

with the visual CNN from the FMN to capture the global semantic information. The decoder part is an upsampling structure that rich detailed information is transferred from the shallower layers to the top layers, learning an elaborate saliency map. Inspired by Zhang et al. [31], FCNs methods directly integrate multi-level features indiscriminately, and are defective due to the redundant details and distractions from background. To combine the multi-level features, we use a skip layer to transferred the shallow-level feature maps from encoder part to the high layers, and an additional convolution layer to predict an integrated feature map, receiving more underlying details. Then, instead of using concat layers to associating contextual semantic information, we use an eltwise layer to directly combine the integrated feature map with the upsampled coarse semantic feature map, and generate a finer foreground map. Three deconvolutional layers are added to expand the feature map to the size of the original image with high resolution. After the multi-level feature fusion, we refine the language-aware saliency map in a coarse-to-fine manner, and the global semantic information is adaptively applied to fuse together with shallow details to generate more effective saliency map.

Since there is no other pixel-level ground truth, a refinement module is designed to recover the coarse map by supplying a series of low-level information at the beginning of the training, generating a finer map and back to fine-tune the network. Moreover, saliency is closely correlated with the object-level concepts, by considering color, contrast and texture, objectness proposal information within a neighborhood plays an important role to retain local salient features. Therefore, given the input image $\boldsymbol{I}$, we first segment it into a set of object candidates using geodesic object proposal method [32]. Each individual object candidate concentrates on a small patches with a high response and projects low-level distinct texture and edge details, thus the refinement module takes in the object candidates and the coarse saliency map $\boldsymbol{S}$, yielding a set of object candidate masks $\{\boldsymbol{O}_j\}_N$. The refinement module focuses on the relevant confidence that proposal can provide valid supplement for the salient object, so two measurements characterized by accuracy $\mathbf{Ps}_j$ and coverage $\mathbf{Rc}_j$ are considered to describe $\boldsymbol{O}_j$, which are defined by

$$\mathbf{Ps}_j = \frac{\sum_{x,y} \mathbf{O}_j(x,y) \times \mathbf{S}(x,y)}{\sum_{x,y} \mathbf{O}_j(x,y)}, \tag{4}$$

$$\mathbf{Rc}_j = \frac{\sum_{x,y} \mathbf{O}_j(x,y) \times \mathbf{S}(x,y)}{\sum_{x,y} \mathbf{S}(x,y)}, \tag{5}$$

$\mathbf{O}_j \epsilon \{0, 1\}$ indicates the pixel at (x, y) belongs to the jth object candidate, and the $\mathbf{S}(x, y)$ represents the saliency value of the same coordinate point predicted by sigmoid function. As a result, the accuracy $\mathbf{Ps}_j$ and coverage $\mathbf{Rc}_j$ respectively measure the average local saliency value and the covered proportion of salient area in the jth object candidate, which functions as Precision and Recall. We define the final confidence for the jth object candidate by considering both the accuracy score and coverage score as $\mathbf{Ps}_j \times \mathbf{Rc}_j$. We sort all the candidates by their confidences and the refined saliency map $\boldsymbol{R}$ is generated by averaging the top 20 candidate regions with the coarse saliency map. Since there is no other ground truth, we considered as the refined map $\boldsymbol{R}$ as the training ground truth back to fine-tune the network. The RFN outputs the finer saliency map $\boldsymbol{F}$ and the loss function of Recurrent Fine-tune Network is followed as:

$$\begin{aligned} \boldsymbol{L}_{rfn} = &- \sum_i [\eta r_i + (1-\eta)\delta] \mathbf{log}(f_i) \\ &+ [\eta(1-r_i) + (1-\eta)\delta] \mathbf{log}(1-f_i), \end{aligned} \tag{6}$$

where r_i and f_i are the salient value at ith pixel index of the training ground truth $\boldsymbol{R}$ and output saliency map $\boldsymbol{F}$, respectively; the monomial δ is the indicator function: $\delta = \mathbf{1}$ if $f_i > 0.5$, which performs as a threshold of saliency prediction; η is a weight parameter which fixed to 0.95 for a robust cross-entropy loss. We solve the loss function using mini-batch SGD, with a batch size of 32. The learning rates of the RFN is initialized as 1e-3, and decreased by 0.1 for every 10 epochs.

The generated object candidates encode informative shape and boundary cues, and serve as an over-complete coverage of the object in an image. By taking in a subset of proposals with high confidence, the refined saliency map $\boldsymbol{R}$ is integrated by local estimation and generic proposals as a complementary process, but it may be not robust enough to distinguish the input noise and blurry boundary cues. Motivated by Wang et al. [17] that they use recurrent structure to fine-tune the network and correct prediction errors, we leverage the output map $\boldsymbol{F}$ to replace $\boldsymbol{S}$ and later refine it by proposals as training truth to update $\boldsymbol{R}$, constituting a training cycle. We repeat the cycle for three times, and the loss

function eliminates approximately the distance between the output map and training ground truth, enforcing the RFN to learn finer saliency map by self-supervision. For keeping the size consistence, the map $\boldsymbol{S}$ in the first recurrent stage is upsampled to the same size of input image while the later map $\boldsymbol{F}$ remains unchanged before performing refinement. In practice, to further accelerate the process of learning and decrease the number of recurrent iterations, those proposals with high confidences are further refined by CRF for better predicting.

4. Experimental results

We train the FMN on the Microsoft COCO caption evaluation dataset, and only captions are leveraged as the supervised truth. To reduce the complexity of the training data in COCO, we also train the RFN using the images on the DUTS-TR dataset [4] without any ground truth. In the test stage, for fair comparison, we generate the saliency maps from the refined RFN without post-processing. All the proposed algorithms are implemented in Caffe and MATLAB (Caffe is used for network training of FMN and RFN, while MATLAB is used for generating training ground truth with gop method and inference), and the average detecting speed per image is 72 FPS with one 3.4 GHz CPU and one 1080 TI GPU.

We compare the proposed approach with a part of state-of-the-art saliency methods, including three unsupervised methods: DRFI [40], DSR [39], BSCA [38] and one weakly supervised method (to our knowledge): WSS [4]. Some representative supervised methods are also presented by year for deep discussion, like LEGS [2], MDF [9], DCL [33], MCDL [10], ELD [34], RFCN [17], DS [35], DLS [36] and UCF [37].

The test benchmark datasets for evaluation are formed by DUT-OMRON [41], ECSSD [42], THUR [43], PASCAL-S [44], DUTS-TE [4] and HKU-IS [9]. Three metrics are utilized to measure the performance, including precision-recall (PR) curves, F-measure and mean absolute error (MAE). The Precision and Recall are computed by segmenting a salient region with a threshold, and comparing the binary map with the ground truth. The PR curves demonstrate the mean Precision and Recall of saliency maps at different thresholds. The F-measure is calculated by

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall}, \tag{7}$$

where β^2 is set to 0.3 as suggested in [45]. The MAE is defined as the average pixel-wise absolute difference between the binary ground truth and the saliency map [46].

4.1. Performance comparison

We use either the implementations or the saliency maps provided by the authors for fair comparison, the result is shown in Table 1 and Table 2. Both the F-measure and MAE evaluated on the benchmark datasets show that our method performs favorably against the representative unsupervised methods, and demonstrate strong competitiveness against the weakly supervised methods [4] using classification annotations. However, in terms of the F-measure and the PR cures in Fig. 5, there is still a gap with the latest approaches using dense structures supervised by pixel-level ground truth, but our approach shows the relatively high accuracy with low error rate than most existing supervised methods. We also provide the quantitative and qualitative analysis in Fig. 4, showing the results of different saliency detection methods for comparison. We can see that both positioning accuracy and the saliency boundary of our method are effective compared to many other methods, such as the leg of the man in first row and the green car in fifth row. However, our approach sometimes fails to uniformly delineate the object regions in the complex scenarios, like the woman on the bench.

Table 1
Part 1: the F-measure and MAE of different saliency detection methods on ECSSD, DUTS and DUT-OMRON. The top nine are fully supervised methods, the middle three are unsupervised methods and the final is weakly supervised methods including ours. The best three results for different detecting methods are shown in **bold**, ***bold italic*** and *italic* fonts respectively.

	ECSSD		DUTS		DUT-OMRON	
	F_β	MAE	F_β	MAE	F_β	MAE
LEGS [2]	0.789	0.119	0.585	0.138	0.592	0.133
MDF [9]	0.805	0.108	0.673	0.100	***0.644***	0.157
MCDL [10]	0.796	0.101	0.594	0.106	0.625	**0.089**
DCL [33]	0.827	0.152	**0.714**	0.149	**0.684**	0.157
ELD [34]	0.810	***0.082***	0.628	0.093	0.611	***0.092***
RFCN [17]	***0.834***	0.109	***0.712***	***0.090***	0.627	0.111
DS [35]	0.826	0.122	0.632	*0.091*	0.602	0.120
DLS [36]	0.766	0.090	–	–	0.591	0.093
UCF [37]	**0.840**	**0.079**	0.629	0.117	0.613	0.132
BSCA [38]	0.707	0.183	0.500	0.196	0.509	0.190
DSR [39]	0.664	0.178	0.518	0.145	0.524	0.139
DRFI [40]	0.734	0.166	0.541	0.174	0.551	0.133
WSS [4]	0.823	0.106	0.657	0.100	0.602	0.110
Ours	*0.831*	*0.088*	*0.682*	**0.084**	*0.634*	*0.093*

Table 2
Part 2: the F-measure and MAE of different saliency detection methods on PASCAL-S, THUR and HKU-IS. The best three results for different detecting methods are shown in **bold**, ***bold italic*** and *italic* fonts respectively.

	PASCAL-S		THUR		HKU-IS	
	F_β	MAE	F_β	MAE	F_β	MAE
LEGS [2]	0.695	0.155	0.607	0.124	0.723	0.119
MDF [9]	0.708	0.146	0.636	0.108	0.784	0.129
MCDL [10]	0.691	0.145	0.620	0.103	0.757	0.092
DCL [33]	0.714	0.181	*0.676*	0.160	**0.853**	*0.072*
ELD [34]	0.718	***0.123***	0.634	*0.098*	0.769	0.074
RFCN [17]	**0.751**	0.132	**0.695**	0.100	***0.835***	0.079
DS [35]	0.659	0.176	0.626	0.116	0.788	0.080
DLS [36]	0.651	0.136	0.621	0.099	0.748	*0.072*
UCF [37]	0.706	*0.126*	0.645	0.111	0.808	0.074
BSCA [38]	0.601	0.223	0.536	0.181	0.654	0.175
DSR [39]	0.557	0.215	0.541	0.141	0.677	0.142
DRFI [40]	0.618	0.206	0.576	0.149	0.722	0.145
WSS [4]	*0.724*	0.141	0.663	***0.096***	*0.822*	0.079
Ours	***0.741***	**0.119**	***0.684***	**0.088**	0.821	**0.067**

Note that some supervised methods use multi-scale or hierarchical structures to detect saliency, our method outperforms them by a stepwise encoder-decoder framework. We decompose the process of saliency detection that the encoder network captures high semantic visual concepts measured by linguistic concepts, while the decoder part optimizes local regions progressively with a self-supervision. The final detecting framework is also highly efficient with a average detecting speed of 72 FPS, which is actually faster than most of the detection methods. We perform two additional evaluations to verify the generalization ability of our method in the Ablation Studies, and the superior performance confirms that our method is flexible and effective for language-based saliency learning without requiring redundant computation overhead.

4.2. Ablation studies

To analyze the relative contributions of different components of our methods, we evaluate some variants of the proposed method with different settings. We first evaluate the performance of the caption compared with label by directly replacing the language branch in the FMN with softmax layer for classification. We de-

Fig. 4. Results of different saliency detection methods.

note it as Our-label network and train the network on the ImageNet object detection dataset [47] with only image-level tags of training images, individually. Then, the specific effect of the RFN on different datasets are demonstrated in Fig. 6 for three different fine-tune stages.

Weakly supervision with caption vs label. For fair comparison, we just train the Our-label variant on the visual branch of FMN, and the output saliency map is transformed uniformly by preserving pixels whose saliency score is greater than the average of the whole map. We show the *F*-measure scores from the five large scale datasets learned by caption and label in Fig. 6(left). Although the image-level labels indicate the global category of the salient object, the limited information contents hardly cover all the visible details and locate the important part. Unlike class-aware results, the foreground maps learned by contextual information are more precise and uniform, which improve by 2–7% on the numerical value than that from labels. The results suggest that the combination of natural language processing and DNNs is valuable and worth to further explore for saliency detection.

The recurrent network. To demonstrate the effectiveness of recurrent network, we illustrate the performance of the top three recurrent times in Fig. 6(right). In addition, Table 3 and Table 4 provide a quantitative and accurate numerical analysis, The *F*-measure scores improve by 2–6% and the MAE scores reduce by 1–3% at each recurrent processing on the different public datasets, which indicates that the progressive learning by self-supervision can readily tap the potential of the small network instead of using duplicated parameters. Obviously, post processing (e.g., CRF) can easily improving the effectiveness of significant object regions, however, it could not replace the feature extraction of the convolutional layers. With training samples approaching the ground true more closely using the refinement module, the RFN shows superiority with those dense networks and the final saliency maps achieve surprising performance. However, in our experiments, we observe that the accuracy of the saliency maps almost converges after the third time step, and it spends a lot of computational overhead to get a more improvement on the numerical data. Intuitively, it can be explained that the proposal method searches for a subset of candidates with high responses, and each recurrent stage tries hard to digest more information from the sharp boundaries and texture. However, it is not an impeccable process for our refinement to screen out elaborate details after the multiple recurrent training, and the limitation of the network is just another obstacle to take.

Weakness. Since our method is based on language-aware saliency detection, it sometimes fails to delineate the correlative regions when the corresponding captions describe some confusing content (shown in Fig 7) in the complex scenarios. We hope to mitigate this issue by conditionally selecting appropriate training dataset, and explore various structures to extract contextual linguistic information in the future.

5. Conclusion

In this paper, we propose a weakly supervised saliency detection method by constructing the matching relation network between visual image and natural language and preserving more informative details with recurrent network. We demonstrate that the affluent textual information learned from caption has a complete concept to cover the dominant visual attention in high-level semantic patterns, which shows an internal relation between language and image. By establishing the textual-visual pairwise affini-

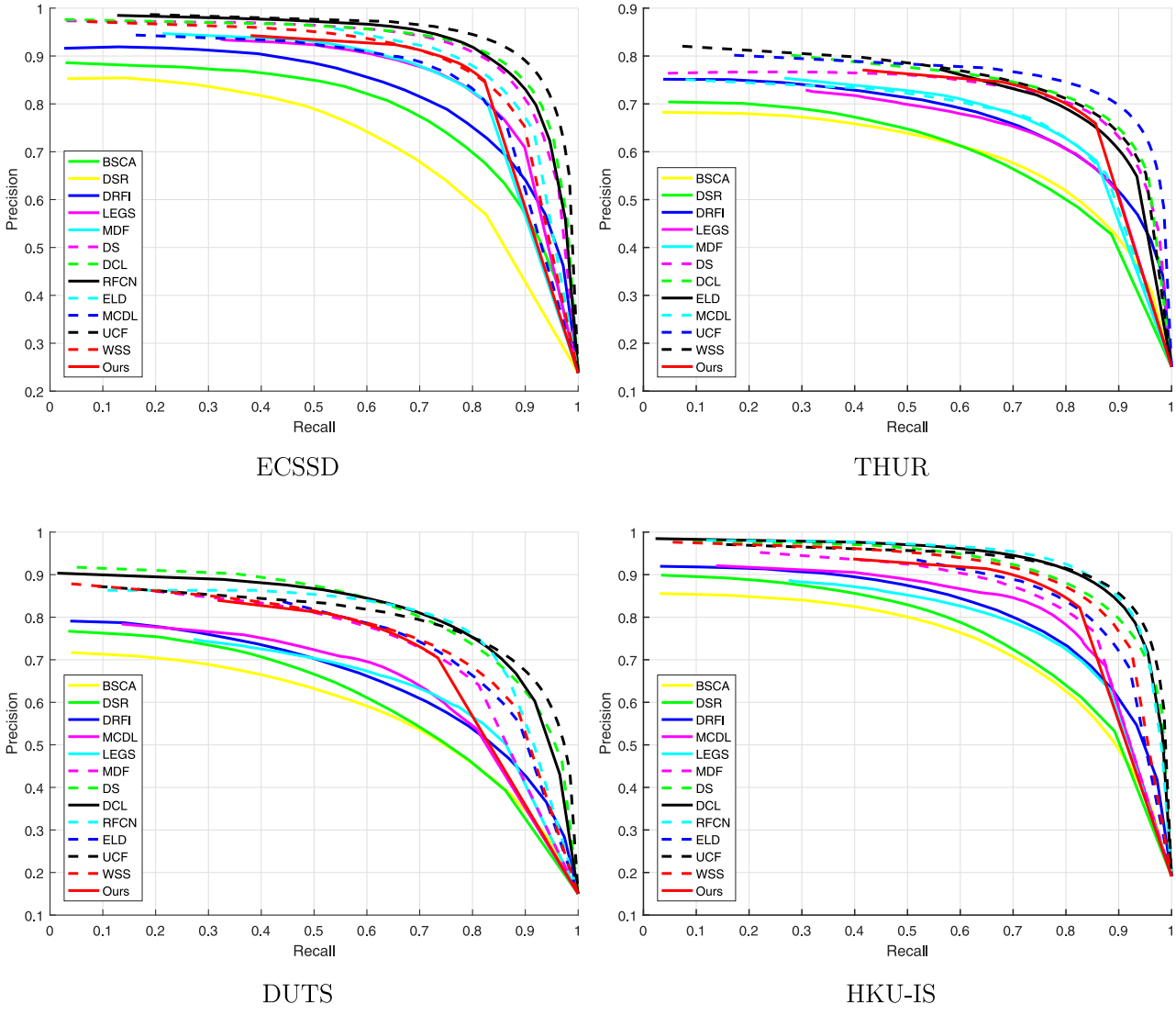


Fig. 5. The PR curves of proposed method with other state-of-the-art methods on ECSSD, THUR, DUTS and HKU-IS dataset.

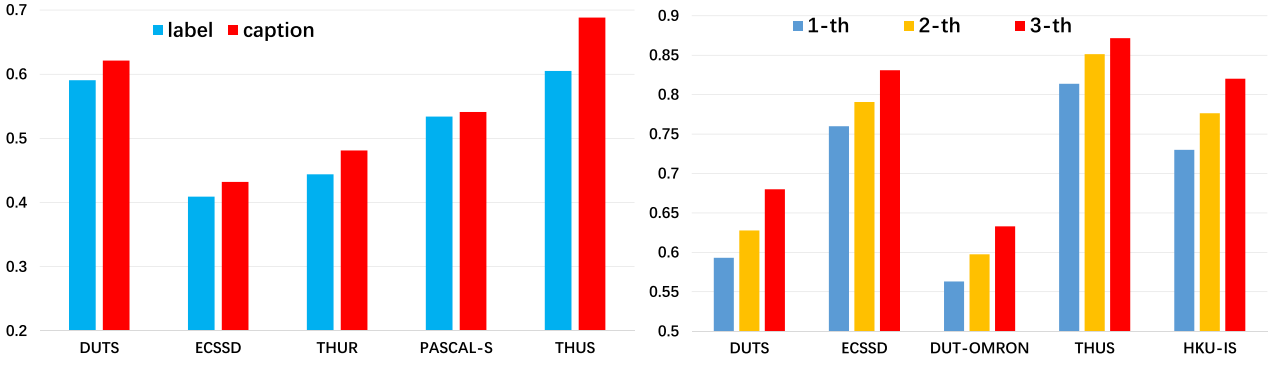


Fig. 6. F_β measure scores on five benchmark datasets. The left is the result of training networks learned by label and caption, the right is the comparisons of scores from the three time steps of training.

Table 3
The Precision, Recall, *F*-measure and MAE of different stages in the training processing on the DUTS and HKU-IS datasets.

	DUTS				HKU-IS			
Stage	*Precis*	*Recall*	F_β	*MAE*	*Precis*	*Recall*	F_β	*MAE*
FMN	0.4376	0.6387	0.4322	0.2577	0.5965	0.6795	0.5858	0.1863
RFN-1st	0.5903	0.7834	0.5932	0.1194	0.7256	0.8371	0.7300	0.1077
RFN-2nd	0.6309	0.7819	0.6179	0.1102	0.7866	0.8406	0.7764	0.0945
RFN-3rd	0.7143	0.7104	0.6821	0.0840	0.8552	0.7843	0.8206	0.0665

Table 4
The Precision, Recall, *F*-measure and MAE of different stages in the training processing on the DUT-OMRON and ECSSD datasets.

	DUT-OMRON				ECSSD			
Stage	*Precis*	*Recall*	F_β	*MAE*	*Precis*	*Recall*	F_β	*MAE*
FMN	0.3616	0.5847	0.3590	0.3149	0.6582	0.6355	0.6212	0.2061
RFN-1st	0.5672	0.7312	0.5631	0.1320	0.7883	0.7901	0.7599	0.1245
RFN-2nd	0.6071	0.7506	0.5975	0.1202	0.8373	0.7903	0.7907	0.1136
RFN-3rd	0.6733	0.6676	0.6344	0.0933	0.8850	0.7623	0.8314	0.0882

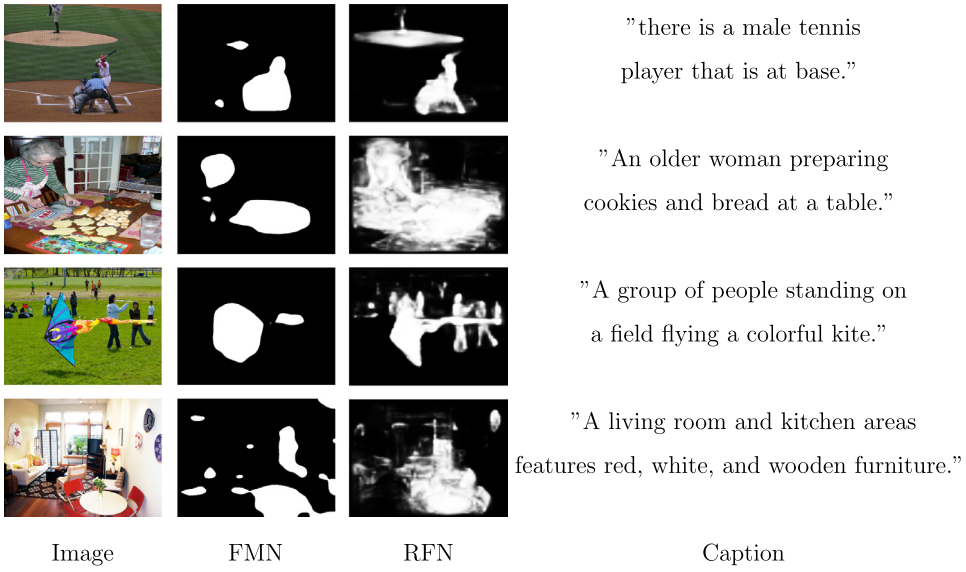


Fig. 7. Some failure cases from out proposed method on the COCO datasets.

ties in the shared latent feature space, our approach provides an accurate information transfer from textual concept of captions to saliency detection than using single labels. Specially, the Recurrent Fine-tune Network considers the predicted map as an important prior and further recovers it by a recurrent self-supervision, enabling our approach to enhance the detection accuracy progressively. Our method is efficient to generate fine-detailed saliency maps at a detecting speed of 72 FPS, extensive evaluations on widely adopted datasets verify the effectiveness and state-of-the-art performance of our method.

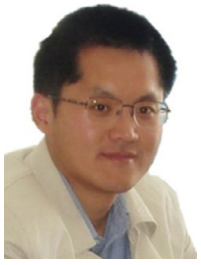
However, our method is trained only on the weak supervision of captions, it fails to precisely locate language-aware object regions in the complex scenarios or distinguish multiple objects like semantic segmentation. In the future, a selective and attentive feature matching algorithm should be considered by carefully discovering the correlative components from different visual CNN layers to produce more accurate language-aware contextual information. On the other hand, experimental results show that the recurrent fine-tune method based on self-supervision can progressively achieve high accuracy on a simple encoder-decoder network, but it is limited and can be rebuilt by dense models like ResNet in the future. This prompts us to consider a more efficient base model to obtain higher saliency detection accuracy from a weak supervision learning perspective.

References

[1] Y. Ji, H. Zhang, Q.M.J. Wu, Saliency detection via conditional adversarial image-to-image network, Neurocomputing 316 (2018) 357–368.
[2] L. Wang, H. Lu, X. Ruan, 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, pp. 3183–3192.
[3] P. Li, D. Wang, L. Wang, L. Huchuan, Deep visual tracking: review and experimental comparison, Pattern Recognit. 76 (2018) 323–338.
[4] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan, Learning to detect salient objects with image-level supervision, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 136–145.
[5] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.

[6] V. Ramanishka, A. Das, J. Zhang, K. Saenko, Top-down visual saliency guided by captions, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3135–3144.
[7] L. Zhang, C. Yang, H. Lu, X. Ruan, M.-H. Yang, Ranking saliency, IEEE Trans. Pattern Anal. Mach.Intell. 39 (9) (2017) 1892–1904.
[8] L. Zhang, J. Ai, B. Jiang, H. Lu, X. Li, Saliency detection via absorbing Markov chain with learnt transition probability, IEEE Trans. Image Process. 27 (2) (2018) 987–998.
[9] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5455–5463.
[10] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1265–1274.
[11] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5300–5309.
[12] G. Li, Y. Xie, L. Lin, Y. Yu, Instance-level salient object segmentation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 247–256.
[13] Y. Ji, H. Zhang, Q.M.J. Wu, Salient object detection via multi-scale attention CNN, Neurocomputing 322 (2018) 130–140.
[14] Y. Yan, J. Ren, G. Sun, H. Zhao, J. Han, X. Li, S. Marshall, J. Zhan, Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement, Pattern Recognit. 79 (2018) 65–78.
[15] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi–path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 86 (2019) 376–385.
[16] Y. Ji, H. Zhang, K. Tseng, T.W.S. Chow, Q.M.J. Wu, Graph model-based salient object detection using objectness and multiple saliency cues, Neurocomputing 323 (2019) 188–202.
[17] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Saliency detection with recurrent fully convolutional networks, in: Proceedings of European Conference on Computer Vision, 2016, pp. 825–841.
[18] S. Wang, Q. Huang, S. Jiang, T. Qi, (SMKL)-M-3: scalable semi-supervised multiple kernel learning for real-world image applications, IEEE Trans. Multimed. 14 (4) (2012) 1259–1274.
[19] Z. Shi, Y. Yang, T.M. Hospedales, T. Xiang, Weakly-supervised image annotation and segmentation with objects and attributes, IEEE Trans. Pattern Anal. Mach.Intell. 39 (12) (2017) 2525–2538.
[20] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, J. Jiao, Weakly supervised instance segmentation using class peak response, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3791–3800.
[21] Z. Fang, S. Kong, T. Yu, Y. Yang, Weakly supervised attention learning for textual phrases grounding, CoRR (2018). arXiv: 1805.00545.
[22] H. Jiang, Weakly supervised learning for salient object detection, CoRR (2015). arXiv: 1501.07492.
[23] S. Wang, Y. Chen, J. Zhuo, Q. Huang, Q. Tian, Joint global and co-attentive representation learning for image-sentence retrieval, in: ACM Multimedia Conference, 2018, pp. 1398–1406.
[24] Y. Wu, S. Wang, G. Song, Q. Huang, Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval, IEEE Trans. Image Process. (2019). 1–1
[25] J. Mao, J. Huang, A. Toshev, O. Camburu, A.L. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 11–20.
[26] S. Li, T. Xiao, H. Li, W. Yang, X. Wang, Identity-aware textual-visual matching with latent co-attention, in: Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 1908–1917.
[27] J. Johnson, A. Karpathy, L. Fei-Fei, DenseCap: fully convolutional localization networks for dense captioning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4565–4574.
[28] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[29] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollar, C.L. Zitnick, Microsoft COCO captions: data collection and evaluation server, CoRR (2015). arXiv: 1504.00325.
[30] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of European Conference on Computer Vision, 2018, pp. 686–701.
[31] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 714–722.
[32] P. Krähenbühl, V. Koltun, Geodesic object proposals, in: Proceedings of European Conference on Computer Vision, 2014, pp. 725–739.
[33] G. Li, Y. Yu, Deep contrast learning for salient object detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 478–487.
[34] G. Lee, Y.W. Tai, 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, pp. 660–668.
[35] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, J. Wang, DeepSaliency: multi-task deep neural network model for salient object detection, IEEE Trans. Image Process. 25 (8) (2016) 3919–3930.
[36] P. Hu, B. Shuai, J. Liu, G. Wang, Deep level sets for salient object detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 540–549.
[37] P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, in: Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 212–221.
[38] Y. Qin, H. Lu, Y. Xu, H. Wang, Saliency detection via cellular automata, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 110–119.
[39] X. Li, H. Lu, L. Zhang, X. Ruan, M.-H. Yang, Saliency detection via dense and sparse reconstruction, in: Proceedings of IEEE International Conference on Computer Vision, 2013, pp. 2976–2983.
[40] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: a discriminative regional feature integration approach, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2083–2090.
[41] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.
[42] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1155–1162.
[43] M.-M. Cheng, N.J. Mitra, X. Huang, S.-M. Hu, SalientShape: group saliency in image collections, Vis. Comput. 30 (4) (2014) 443–453.
[44] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.
[45] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.
[46] F. Perazzi, P. Krahenbuhl, Y. Pritch, A. Hornung, Saliency filters: contrast based filtering for salient region detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 733–740.
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

Mingyang Qian received her B.E. degree in electrical and information engineering, Dalian University of Technology (DUT), China, in 2018. He is currently a master student in Signal and Information Processing, Dalian University of Technology (DUT). His research interest is in saliency detection and video object segmentation.



Mengyang Feng received the B.E. degree in electrical and information engineering from the Dalian University of Technology in 2015, where he is currently pursuing the Ph.D. degree under the supervision of Prof. H. Lu.

Lihe Zhang received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2004. He is currently an Associate Professor with the School of Information and Communication Engineering, Dalian University of Technology. His research interests include pattern recognition and computer vision.

Jinqing Qi received the Ph.D. degree in communication and integrated system from the University of Tokyo Institute of Technology, Tokyo, Japan, in 2004. He is currently an Associate Professor of Information and Communication Engineering at University of DUT, Dalian, China. His recent research interests focus on computer vision, pattern recognition and machine learning. He is a member of IEEE.

Huchuan Lu received the M.S. degree from the Department of Electrical Engineering, Dalian University of Technology (DUT), China in 1998 and his Ph.D. degree of System Engineering from DUT in 2008, respectively. From 1998 to now, he is a faculty of School of Electronic and Information Engineering of DUT. He has been associate professor since2006. He has visited Ritsumeikan University from Oct. 2007 to Jan. 2008. His recent research interests focus on computer vision, artificial intelligence, pattern recognition and machine learning. He is a member of IEEE and IEIC.