Instance of Interest Detection

Fan Yu[1,3], Haonan Wang[1], Tongwei Ren[1,3,*], Jinhui Tang[2], Gangshan Wu[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]School of Computer Science, Nanjing University of Science and Technology, Nanjing, China
[3]Shenzhen Research Institute of Nanjing University, Shenzhen, China
yf@smail.nju.edu.cn,527725618@qq.com,rentw@nju.edu.cn,jinhuitang@njust.edu.cn,gswu@nju.edu.cn

ABSTRACT

In this paper, we propose a novel task named Instance of Interest Detection (IOID) to provide instance-level user interest modeling for image semantic description. IOID focuses on extracting the instances which are beneficial to represent image content, while other related tasks such as saliency analysis, attention model and instance segmentation extract the regions attracting visual attention or with a predefined category. To this end, we propose a Cross-influential Network for IOID, which integrates both visual saliency and semantic context. Moreover, we contribute the first dataset IOID evaluation, which consists of 45,000 images from MSCOCO with manually annotated instances of interest. Our method outperforms the state-of-the-art baselines on this dataset.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Instance of interest; instance of interest detection; instance of interest annotation; instance extraction; interest estimation

ACM Reference Format:
Fan Yu[1,3], Haonan Wang[1], Tongwei Ren[1,3,*], Jinhui Tang[2], Gangshan Wu[1]. 2019. Instance of Interest Detection. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350931

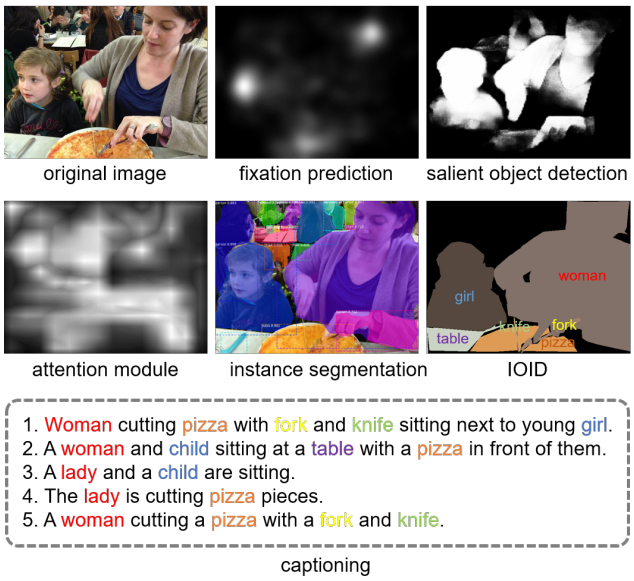

Figure 1: Comparison of IOID with the relevant tasks. The top-left illustrates the original image, and the rest illustrates the results of fixation prediction, salient object detection, attention module, instance segmentation, IOID and captioning, respectively.

1 INTRODUCTION

As the English idiom goes, A picture is worth a thousand words. It is true in real life as pictures demonstrate their superiority over mere descriptions in their richness in information. The same could be said in our research. As shown in Figure 1, the original image illustrated in the top-left contains a women, a girl, a pizza, two forks, two knives, a table, a watch, a ring, and even more than ten persons with four cups and a book in the background. But the above description rarely appears in our daily life because it is flooded with details, even those of little interest to others. Not all the instances share the same importance in representing the content of the picture (as shown in the captioning results in Figure 1). Instances with greater appeal are called "*Instance of Interest*" (IOI). IOI can be treated as a specific kind of Region of Interest (ROI) from a broader sense. From a narrower perspective, the definition we adopt in this paper are: "region" in ROI is comparable to "instance", covering both *thing* and *stuff*, and the "interest" is restricted to the appeal when describing images. Obviously, there is little room for doubt that IOIs are the fundamental elements for image understanding. Their interactions and themselves form the backbone of image description, which can be applied in various image understanding applications, such as image captioning [6, 10, 34, 36, 37] and visual question answering [2, 21].

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350931

A similar concept to IOI is *saliency*. Saliency analysis starts from fixation prediction, which aims to predict the regions people may pay attention to in viewing a picture. Later, salient object detection is proposed, which aims to generate the saliency maps or even extract the whole salient objects [5, 31, 38]. In a nutshell, both fixation prediction and salient object detection focus on exploring what people may *watch* in viewing images rather than *concern* in describing images. Hence, most saliency analysis methods derive information from the characteristics of human vision system with little consideration for semantic cues.

Another related task to IOI detection is *instance segmentation*, which aims to segment the instances belonging to the specific categories. The instance category set should be predefined and remain the same, and it will not be suitable for the images containing objects of extremely different types. In IOID, the "interest" of instances with the same category may vary greatly in different images, even in the same image. Hence, instance segmentation cannot be used to detect IOIs. Some relevant techniques, such as semantic segmentation [4, 25, 39, 40] and scene parsing, focus on pixel-level semantic labeling rather than modeling user interest, and they cannot provide instance-level decomposition results.

Moreover, to modeling user interest, attention mechanism is widely used in deep networks for image understanding applications, *e.g.*, image captioning and visual question answering. In attention module, some pixels will affect the caption words or answers of the question. Although semantically-influential pixels would be picked out by attention module, the results are often so vague and broad that the area may include more than one object.

Based on the above observation, we propose a novel task named *Instance of Interest Detection* (IOID) to extract the IOIs from the given images. Figure 1 shows the comparison between IOID and the relevant tasks. We can see that IOID has the following characteristics: 1) IOID represents the detected IOIs at instance-level, which is different from fixation prediction, salient object detection and attention module; 2) Contrary to instance segmentation, IOID only retains the instances concerned in image descriptions; 3) IOID provides the semantic label together with the corresponding segmentation region for each IOI, which is different from captioning.

We propose a IOID method based on the Residual Network (ResNet) and Feature Pyramid Network (FPN). The feature maps are then used for panoptic segmentation and object of interest selection. Since the objects of interest aim to convey the content of an image, the segmentation branch needs to export both things and stuff in the image. On the level of instance, the objects influence each other in a RNN and those of higher significance are picked out.

There is no existing dataset for IOID as of today. Hence, we construct a IOID dataset on the basis of MSCOCO dataset [29]. We labeled the images based on caption and panoptic segmentation annotations and removed some images that are hard to be annotated accurately.

Our contributions mainly include:

- Proposal of a novel IOID task that aims to detect the instances of interest in describing a given image by providing their segmentations and semantic labels.
- Proposal of an IOID method which detects the IOIs through a Cross-influential Network composed of a detection and segmentation module, an interest estimation module and an instance of interest selection module.
- The first IOID evaluation dataset consisting of 45,000 images with manually labeled IOIs.

The rest of the paper is organized as follows. In Section 2, we survey the related works, including visual saliency, visual attention mechanism, object detection and image segmentation, and visual relation detection. In Section 3, we introduce the first dataset for the IOID task. Next, we explain the details of the proposed methods in Section 4 and present the results and analysis of our experiments in Section 5. Section 6 is the conclusion of our research.

2 RELATED WORKS

Visual Saliency. Visual saliency has received attention for many years. Some researchers have contributed to saliency prediction, under the principle to form a saliency map to represent the salient intensity of different pixels. Some have devoted their efforts to salient object detection, with the principle of highlighting the salient object regions in images.

Though the origin of saliency prediction was much earlier than salient object detection, they share the same starting point-analysis of low-level features (such as color, texture, localization and shape [13, 16]), which provide valuable information to human visual system. Meanwhile, fixation prediction [8, 12] which aims at identifying human's fixation points plays an important role during the development of visual saliency. Especially after the proposal of deep learning solutions, the fixation of human viewers often act as the ground truth of saliency prediction [24]. An example is the SALICON dataset [17]–whose images come from MSCOCO [29]–which provides the fixation data of each image, and is now widely used in research on visual saliency. At the same time, deep learning has significantly improved the performance of visual saliency [33]. In recent years, with the development of image semantic segmentation and object proposal, the solutions of salient object detection gradually transfers from pixel-level to instance-level [9, 26].

Image Captioning. In the field of visual information understanding, image captioning and image visual question answering are two important tasks. With the development of neural network, image captioning solutions draw lessons from machine translation [7] and translate the image from visual representation to language [34]. Image question answering bears resemblance to image captioning, the former, however, differs in that the question is given and the answer is inferred by learning the question text and relevant visual information. As mentioned in the saliency part, the MSCOCO dataset also provides data about image captioning. Meanwhile, an

important dataset about image question answering named VQA [2] is constructed on the basis of MSCOCO.

Attention Mechanism. Attention Mechanism was proposed long time ago, but it started to become popular after [35] was published. In this paper, they add the attention mechanism to a recurrent model to classify images. After that, attention mechanism is used widely in natural language processing to perform translation and alignment simultaneously and helps to overcome difficulty in translation of long sentences. Inspired by machine translation, researchers applied attention mechanism in computer vision again. In [37], attention mechanism is used to add visual information to sequence generation in image captioning. The attention module introduces visual features to language parsing rather than the entire image embedding from the fully connected layer of a deep CNN.

Object Detection and Image Segmentation. Object detection aims to identify different things in the image. The most popular method to detect objects is to generate a large number of region proposals and predict the category probability of each region. In further research, instance segmentation was proposed on the basis of object detection [14] [15].

The target of semantic segmentation is to align a category label to each pixels in the image. A typical method uses a fully connection layer after a series of convolutions. [18] proposed fully convolutional networks for semantic segmentation to improve the performance of solution to this task.

Semantic segmentation and instance segmentation have drawn attention for some time now, however, panoptic segmentation remains yet to be studied. Panoptic segmentation is a joint task of thing and stuff segmentation [23]. Many good solutions were published in the 2018 COCO and Mapillary Recognition Challenge [22, 27]. A heap of solutions handle this problem using separate networks for instance and semantic segmentation, and some other solutions design end-to-end framework to share computation. After that, researchers still make efforts to panoptic segmentation [3]. Instance segmentation and panoptic segmentation both depend on object detection.

3 DATASET

We construct the first dataset for IOID based on the training set of MSCOCO 2017 [29], which contains 123,287 images with manually labelled captions and panoptic segmentation. According to its definition, IOI is the instance of interest in the description of a given image. Hence, we may annotate the IOIs in a given image by labelling the instances appear in its captions and selecting the corresponding regions from its panoptic segmentation.

In each image, we extract the nouns in the captions of each image by NLTK library, and automatically check the possibility for each noun of having one or more corresponding instance candidates in the panoptic segmentation of the image. For each noun, if there is one or more instances with the categories it belongs to, *e.g.*, "man" belongs to "person", the noun is considered to have one or more instance candidates.

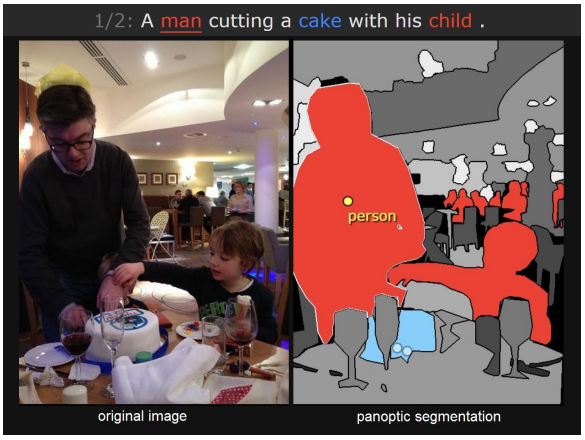

Figure 2: An example of IOI annotation. The annotator selects an instance region from the panoptic segmentation (bottom-right) for a noun in the caption (top) while viewing the original image (bottom-left) as a reference.

We filter the captions in which the number of nouns without any instance candidate is more than 20% of that of all the nouns, and retain the images with at least one remaining caption.

Figure 2 shows an example of IOI annotation. In the annotation system interface, a caption is shown on the top, and the original image and its panoptic segmentation are shown in the bottom. In the caption, the nouns with one or more instance candidates are labelled in different colors while the other words are illustrated in white. Specifically, if a noun has only one instance candidate, the noun and its instance candidate in the panoptic segmentation are labelled in blue, where the annotators are only required to confirm the matching. If a noun has more than one instance candidates, the noun and its instance candidates are labelled in red, in which case the annotators also need to select the matching instance from all the candidates. In addition, the currently annotated noun is underlined for emphasis. Such a strategy may improve annotation accuracy and speed effectively. In our annotation, it only requires 20 seconds to annotate each IOI in average. Furthermore, in the cases where the instance segmentation is not accurate enough or it is hard to pick out the corresponding instance for a noun, we encourage the annotators to discard the images. Finally, we annotate 45,000 images with 205,711 IOIs.

We construct our dataset by representing each IOI with its category and its region in the panoptic segmentation. There are 133 categories of IOIs in total, the same as those provided by the panoptic segmentation in MSCOCO 2017, containing 80 thing categories, such as person, ball and cow, and 53 stuff categories, such as wall, tree and mountain. We further divide the datasets into the training set and the test set, which contain 36,000 images with 165,094 IOIs and 9,000 images with 40,617 IOIs, respectively.

4 METHOD

4.1 Overview

IOI represents the user interest in image description. It is influenced by many factors as to whether an instance is an IOI, see the description of the image it belongs to. For example, the instances with salient visual characteristics, such as large size, distinctive color and central position, are likely to be mentioned, such as the women and the girl in Figure 1. Meanwhile, the instances which have interactions with the visually salient instances are also mentioned in image description, such as the knife and the fork in Figure 1 are used in the hands of the woman, though there are another fork and knife with larger sizes beside them. Moreover, the instances about the surroundings may also be mentioned, such as the table in Figure 1 shows where the woman and the girl sit. Hence, an effective IOID method should be able to integrate different influences rather than emphasizing a specific characteristic, which is commonly used in existing tasks, such as visual saliency in salient object detection and object category in object detection.

To address the problem of IOID, we propose a novel end-to-end Cross-Influential Network (CIN), to handle the subtasks in IOID, *i.e.*, instance-level image decomposition, instance recognition, and instance interest estimation. Figure 3 shows an overview of the proposed CIN. It consists of three key modules: instance extraction, interest estimation and IOI selection. Specifically, both instance extraction module and interest estimation module use the feature maps extracted from each layer of a five-layer ResNet101 [20] as their inputs, and their results are used as inputs for the IOI selection module.

4.2 Instance Extraction

Because IOID represents the detected IOI at instance-level, it is necessary to decompose the original images into instances, including both things and stuff. Meanwhile, IOID also requires provision of the semantic labels of IOIs.

Instance segmentation methods, such as Mask R-CNN [14], apply object detection by setting numbers of candidate boxes at the beginning. While stuff are usually distributed in a large area discontinuously. It is therefore difficult to cover the entire stuff while leaving out certain things. For this reason, we decide to combine instance segmentation and semantic segmentation to tackle the subtasks of IOID, instance-level image decomposition and instance recognition, together in the instance extraction module.

Inspired by [22, 27], we use the FPN [28] as the backbone to obtain the multi-level feature maps, and feed them into the two branches of thing and stuff instance extractions respectively. FPN takes the top four output feature maps from ResNet as input, and adds a light top-down pathway with lateral connections. The feature maps from higher pyramid levels are spatially coarser, but semantically stronger. FPN upsamples those feature maps, and merges those of the same spatial size from the bottom-up pathway and the top-down pathway to obtain more accurate location:

$$p_k^o = \phi(f_k) + \chi(p_{k+1}^o), \tag{1}$$

where p_k^o is the output of the kth layer of FPN; f_k is the output of the kth layer of ResNet, $k \in \{1, 2, 3, 4\}$; $\chi(.)$ and $\phi(.)$ denote the upsampling and convolution operations, respectively.

In the thing instance extraction branch, we follow the procedure of Mask R-CNN [14]. The RPN is used to take each output feature map of the FPN as an input, and it performs ROI pooling. The results of each level in FPN are integrated and imported into a proposal layer to extract candidate ROIs. The ROIs are refined in a detection layer and imported into the classifier and mask heads with the top four feature maps of the FPN.

In the stuff instance extraction branch, similar to [22], the outputs of the top four layers from the FPN are convoluted, so that channel number is equal to the number of classes. And then they are upsampled to the same size and summarized together:

$$s = \sum_{k=1}^{4} \chi(\phi(p_k^o)), \tag{2}$$

where s is the result with 134 channels implying classification of each pixel; p_k^o is the output of the kth layer of FPN; $\chi(.)$ and $\phi(.)$ denote the upsampling and convolution operations, respectively. In this way, all of the feature maps are transformed into the size of the last layer. And then the outputs from all levels of the original pyramid are merged into a single output by an integrated convolution and finally it is upsampled to generate the required semantic segmentation. To separate stuff from things, we omit the regions of things in the semantic segmentation, which comes from the stuff instance extraction branch, and extract the bounding box and mask of each stuff to ensure consistency with the thing instance extraction branch.

In the end, all instances extracted from the thing branch and the stuff branch are integrated together with their category, bounding box and segmentation mini mask.

We use the pretrained model of Mask R-CNN. Firstly we retrain the stuff branch, keeping the weights of ResNet101 and FPN unchanged. And then we finetune all the weights, and final loss combines the losses of thing branch and stuff branch.

4.3 Interest Estimation

This module estimates pixel-interest directly according to features extracted from the backbone, with little consideration for object segmentation and category, and contributes to the IOI selection module. The ground truth used in this procedure is the binary segmentation of IOIs. Inspired by [30], we use a Contextual Attention Network (CAN), which shares the ResNet101 as the backbone with the instance extraction module.

CAN uses dual attention mechanism, global attention and local attention, to collect contextual information in different scales. Specifically, the feature maps of the top two layers in

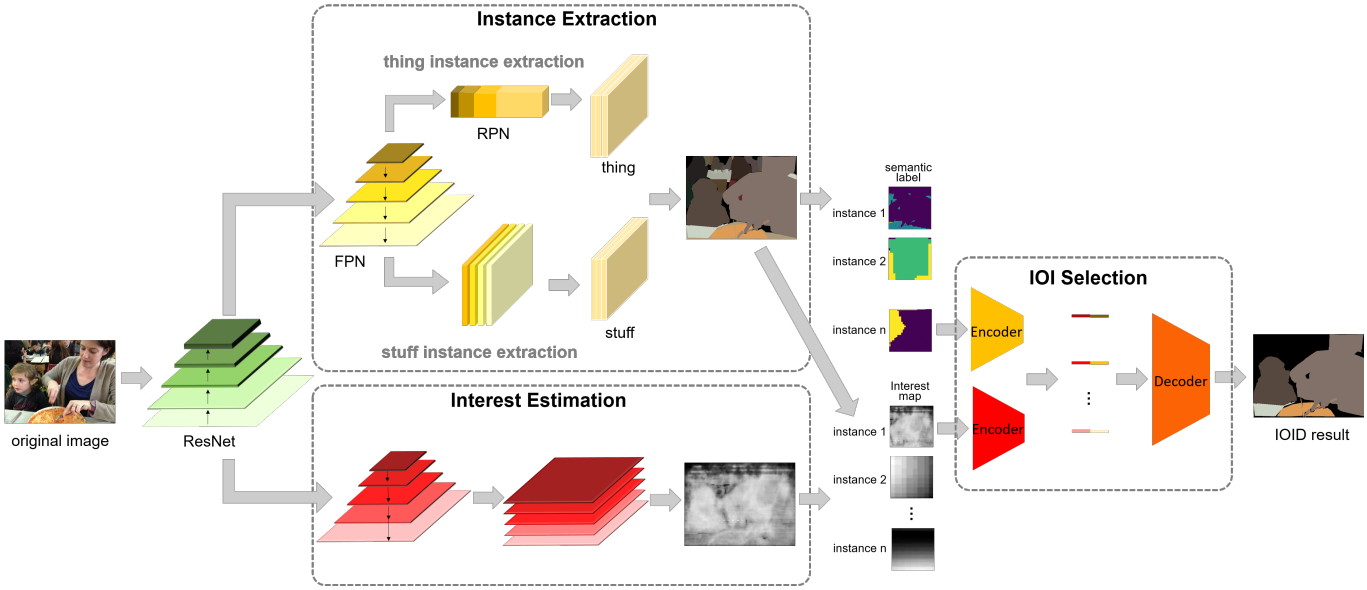

Figure 3: An overview of the proposed IOID method.

ResNet101 are used to generate global attention, while those of the next two layers are applied in local attention, which works on a local region centered at a position.

In the global attention mode, each pixel is swept by two bidirectional LSTM in horizontal and vertical direction. Thus the contexts from four directions can be blended in to propagate the information of each pixel to all other pixels. The output is then transformed into the original size of the input feature map and normalized via a softmax function to generate the attention weight $\alpha^g_{i,j}$:

$$\alpha^g_{i,j} = \frac{\exp(x_{i,j})}{\sum_{i,j}\exp(x_{i,j})}, \tag{3}$$

where (i,j) denotes the position of a pixel, $i \in \{1,...,W\}$, $j \in \{1,...,H\}$, and W and H are the width and height of the input feature map, respectively; $x_{i,j}$ denotes the output at the position (i,j). Finally, the input feature map is summed with weight $\alpha^g_{i,j}$ to generate the global feature F^g:

$$\xi^g_{i,j} = \sum \alpha^g_{i,j} f_{i,j}, \tag{4}$$

where $f_{i,j}$ is the flatten conv-value at (i,j) in the input feature map.

In the local attention mode, attention is generated by performing several convolutions. Each pixel is affected by its neighboring context region with the width of w and height of h. The attention kernel is generated by several convolution from the original feature maps and normalized to a tensor whose channel is $w \times h$. It is also normalized by a softmax function $\alpha^l_{i,j} = \frac{\exp(x_{i,j})}{\sum_{i,j}\exp(x_{i,j})}$, here $i \in \{1,...,w\}$ and $j \in \{1,...,h\}$. The feature maps are unfolded into a tensor whose last dimension is $w \times h$, and summed with the transposed attention weight to generate the local feature $\xi^l_{i,j} = \sum \alpha^l_{i,j} f_{i,j}$.

In this way, we generate four features, two global and two local, respectively. In the training procedure, the loss function of interest estimation is defined as the sum of the loss on these four features:

$$\mathcal{L} = \sum \psi(\xi^* - \widetilde{\xi^*}), \tag{5}$$

where $\mathcal{L}$ represents loss of interest estimation; $\psi(.)$ represents binary cross entropy; ξ^* denotes a global feature ξ^g or a local feature ξ^l, and $\widetilde{\xi^*}$ denotes the corresponding groundtruth ξ^*. The last layer feature is treated as the interest map of the whole image.

4.4 IOI Selection

Though interest estimation module contains semantic information by using high level feature maps, it is still not accurate nor explicit. The IOI selection module acts as an instance selector, which aims to use interest estimation results to pick out the IOIs from the instances proposed from the instance extraction module. To solve the problem of IOI selection, we propose a Cross-influential Encoder-decoder Network (CIEDN).

The interest of one instance is concerned with others interacting with it. CIEDN predicts whether the pair of instances is interesting, instead of the interest of one instance. It is assumed that if the pairs which contain/with a certain instance are interesting in general, the instance is considered to be interesting as well. Hence,if an instance is classified to be uninteresting by pixel-interest estimation, it may be finally judged as IOI with the "help" of other instances. On the contrary, if the pixel-interest estimation module mistakes an instance for an IOI, CIEDN may decide that it is not interesting, even in the pairs which consider it to be interesting.

CIEDN is composed by an encoder and a decoder. As Figure 3 shows, the rectangular area of semantic label and that of interest estimation result surrounding an instance are respectively input into a corresponding sub-encoder, and the output vectors are stitched together. Then, each pair of two instances is evaluated by a decoder. If the average probability of the pairs including a certain instance exceeds 0.5, the instance is determined as an IOI.

5 EXPERIMENTS

5.1 Evaluation Metrics and Experiment Settings

Evaluation metrics. We validate the effectiveness of our method by comparing it with the state-of-the-art baselines on the dataset provided in Section 3. Since IOID can be formulated as a binary classification problem to determine whether an instance is an IOI, we use precision, recall and F-score as the evaluation metrics. Considering a good IOID method needs to provide instance-level detection results rather than pixel-level ones, we refine the definition of precision, recall and F-score on instance level in our evaluation. Here, F-score is calculated as follows:

$$F = \frac{(\beta^2 + 1) precision \cdot recall}{\beta^2 \cdot precision + recall}, \tag{6}$$

where β^2 is the parameter to control weights of precision and recall. Similar to salient object detection [1], we emphasize precision by setting β^2 to 0.3.

To a detected IOI I_i, if an IOI in the corresponding groundtruth $\widetilde{I}_j$ satisfies these two conditions, we consider I_i matches $\widetilde{I}_j$ and treat I_i as a true-positive detection result: 1) the category of I_i should be same to that of $\widetilde{I}_j$; 2) the Intersection of Union (IoU) between I_i and $\widetilde{I}_j$ should be larger than 0.5. If more than one detected IOIs matches the same IOI in groundtruth, we select the detected IOI with the largest IoU to the IOI in groundtruth, *i.e.*, each IOI in groundtruth can match one detected IOI at most.

In our experiments, we found that the inaccuracy of instance extraction limits the performance of IOID seriously. Figure 4 shows the number distribution of the *missed* IOIs in instance extraction, *i.e.*, the IOIs in groundtruth without any extracted instances have the same categories to them and sufficient IoUs (larger than 0.5) with them. We can see that the missed IOIs are pervasive. The number of images without any missed IOI is less than 6%; the largest number of the missed IOIs in an image reaches 43, and the average number of the missed IOIs is 1.78. Thus, to validate the performance of determining whether an instance is interesting, we use additional metrics by removing the missed IOIs in our evaluation. The removal of the missed IOIs affects the metrics of recall and F-score but not precision, so we represent the two new metrics as recall* and F^*.

Experiment settings. In training procedure, if an instance is an IOI, its probability is set to 1; otherwise, it is set to 0. Moreover, the probability of a pair is set to the mean value of the probabilities of the two instances it includes. Hence, if an instance is an IOI, the probability of each pair it belongs to achieves 0.5. In test procedure, we predict the probability of an instance with the mean value of the probabilities of all the pairs it belongs to. We slightly relax the threshold to 0.4, which can obtain the best performance as shown in Table 1.

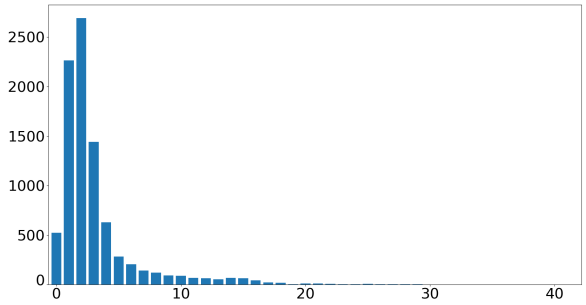

Figure 4: Number distribution of the missed IOIs in instance extraction on the test set.

Table 1: Threshold analysis of IOI probability on the training set.

threshold	0.30	0.35	0.40	0.45	0.50	0.55
precision	41.76	53.99	66.02	75.66	82.17	**86.55**
recall	**49.98**	41.67	33.25	24.65	16.78	10.07
F	43.41	50.54	**53.78**	51.21	43.26	31.45

5.2 Component Analysis

Instance extraction. To verify the importance of both thing and stuff in instance extraction, we use either the extracted things or stuff instead of both of them. We generate two baselines based on the state-of-the-art instance segmentation method, Mask R-CNN [14], and semantic segmentation method, Deeplab [4]. Specifically, for the baseline which only includes things (Thing), we use the instance segmentation results of Mask R-CNN directly; and for the baseline which only includes stuff (Stuff), we remove the instances belonging to things from the segmentation results of Deeplab.

Table 2 shows the performance of our method and these two baselines. We can see that:

1) Both Thing and Stuff obtain relatively low scores on recall and recall* as compared to our method. Specifically, the scores of Thing on recall and recall* drop 20.49% and 23.80%, respectively; and that of Stuff drop 27.56% and 34.76%, respectively. It shows both things and stuff are indispensable in IOID.

2) The scores of Thing on both recall and recall* are higher than that of Stuff. For example, the score of Thing on recall is more than two times higher than that of Stuff. It is because that things usually represent the primary content of images and they are more possible to be IOIs than stuff.

Table 2: Evaluation of our method with different instance extraction modules.

Method	precision	recall	F	recall*	F^*
Thing [14]	**87.06**	9.66	30.56	26.00	56.47
Stuff [4]	19.91	2.59	7.82	15.04	18.52
Our	68.47	**30.15**	**52.95**	**49.80**	**63.02**

3) Thing obtains much higher score on precision than Stuff and our method. A possible reason is that the removal of stuff increases the accuracy of IOI selection. Because stuff is more ambiguous to be IOIs or not than things, its ambiguity increases when determining the probability of an instance to be an IOI based on the pairs including it and stuff as compared to those only referring to things. However, as mentioned in the first point, the cost of only retaining things is poor scores on recall and F. In contrast, our method that retains both things and stuff achieves a balance between high precision and high recall, and obtains the best performance on F.

4) The scores of all the three methods on recall* are obviously higher than that on recall, from 12.45% to 19.65%. It also leads to more than 10% increase in the scores on F^* as compared to those on F. It shows that instance extraction is still one of the limitations to obtain satisfactory IOID results.

Interest estimation. Since there is no existing instance interest estimation method, we generate six baselines based on five state-of-the-art salient object detection methods, namely DSS [32], MSRNet [11], NLDF [31], PiCANet [30] and SalGAN [19], and an attention model, SAT [37], to illustrate the effectiveness of our interest estimation module. In the generation of these baselines, we directly replace the interest estimation module of our method with these salient object detection methods and the attention model, and treat their outputs as the interest maps.

Table 3 shows the performance of our method and these baselines, in which we represent the baselines with the names of the salient object detection methods and attention model. We have:

1) The baselines using salient object detection methods usually perform well on precision, but they are weak on recall. For example, DSS obtains the highest precision score, which is slightly higher than that of our method, but it performs the worst on recall and recall*. It is because that salient object detection methods focus on salient objects, which are usually IOIs, but ignore the IOIs without salient appearances. Thus, these baselines achieve worse performance than our method on F and F^*. Among all these baselines, the baseline using MSRNet has similar performance to our method. It also obtains a balance between precision and recall, and it is slightly worse than ours on all metrics.

2) The baseline using attention model SAT obtains higher recall than our method, but it performs the worst on precision as compared to all other baselines. Similarly, it obtains lower scores than our method on F and F^*.

Table 3: Evaluation of our method with different interest estimation modules.

Method	precision	recall	F	recall*	F^*
DSS [32]	**68.78**	15.24	37.99	25.18	49.14
MSRNet [11]	63.87	29.92	50.62	49.42	59.83
NLDF [31]	67.33	23.18	46.77	38.28	57.30
PiCANet [30]	67.63	24.36	47.97	40.24	58.45
SalGAN [19]	60.31	23.66	44.43	39.09	53.59
SAT [37]	52.09	**30.73**	44.89	**50.76**	51.78
Our	68.47	30.15	**52.95**	49.80	**63.02**

Table 4: Evaluation of our method with different IOI selection modules.

Method	precision	recall	F	recall*	F^*
Binary	40.93	35.71	39.59	58.98	44.04
RNN	46.57	**49.10**	47.13	**81.12**	51.64
Our	**68.47**	30.15	**52.95**	49.80	**63.02**

3) By comparing Table 2 and 3, the baselines of replacing interest estimation module perform better than those of replacing instance extraction module. It means that the existing salient object detection methods and attention models can partly solve the problem of interest estimation, such as extracting the instances with salient appearances by salient object detection methods and the ones attracting attention by attention models.

IOI selection. To validate the effectiveness of the proposed IOI selection module, we generate two baselines. One baseline (Binary) binarizes the interest maps by setting the pixels with high interest, top 25% in our experiments, to 1, and setting the rest pixels to 0. If more than half of the pixels within an extracted instance are set to 1, the instance is selected as an IOI. The other baseline (RNN) uses an RNN to predict the probabilities of all the extracted instance, which uses a vector consisting of the categories and the maximum interest values of all the extracted instances as the input. Here, we use RNN rather than a classifier to explore the cross-influence between instances.

Table 4 shows the performance of our method and the two baselines. We can see that:

1) Binary is too simple to provide satisfactory IOI selection results. Though it obtains higher scores than our method on recall and recall*, its scores on precision and F are low. If we adjust the threshold in interest map binarization for higher precision score, it will inevitably cause the diminution of recall score.

2) RNN provides better results than Binary, *i.e.*, RNN improves the performance on all the metrics as compared to Binary. Specifically, RNN obtains much higher scores on recall and recall* than our method. However, if the recall score goes too high, it may detect many false-positive IOIs which limits its performance on precision and F.

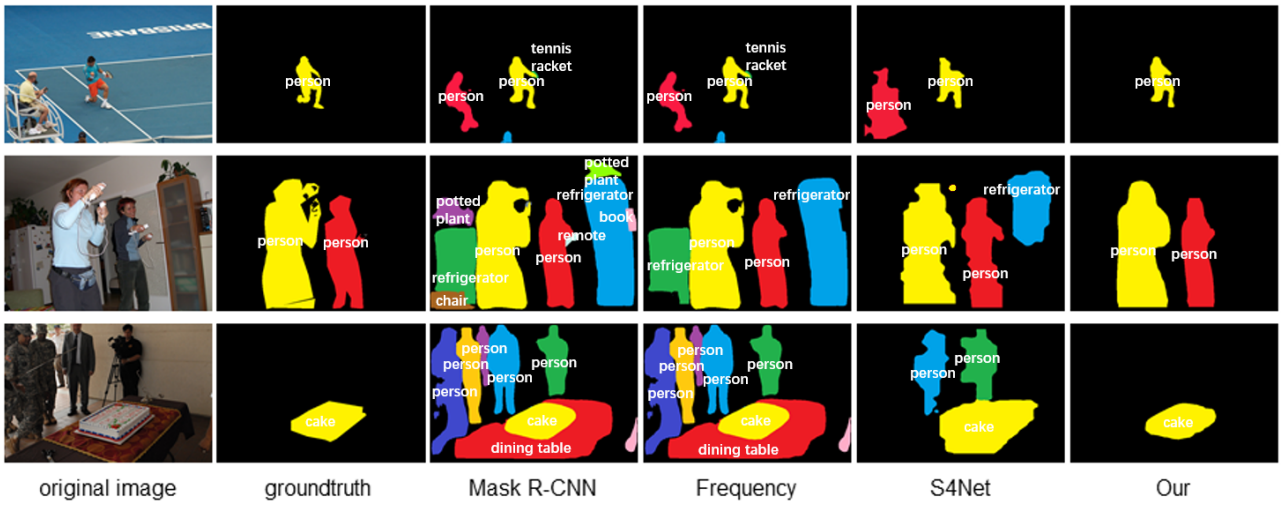

Figure 5: Qualitative examples of IOID using different methods.

5.3 Comparison with State-of-the-Arts

To further validate the overview performance of our method, we also compare it with several state-of-the-art methods of the related tasks, because there is no existing IOID method. The first baseline is Mask R-CNN [14], which is a typical and widely used method for instance segmentation. Here, we treat all the segmented instances as IOIs. The second baseline is based on the frequency of each category to be an IOI. We use Mask R-CNN to generate the instance segmentation results and only retain the instances with high frequency, top 50% in our experiments. The third baseline is S4Net [9], which is an advanced salient object segmentation method. Similarly, we treat all the segmented salient instances as IOIs.

Table 5 shows the comparison results. Here, Mask R-CNN and S4Net obtains 100% on recall* because all their segmented instances are retained. We have:

1) Both our method and Frequency use Mask R-CNN as the basis to extract instance, and obtain better performance than Mask R-CNN. It shows the effectiveness of the IOI selection strategies used in our method and Frequency. Furthermore, as compared to Frequency, our method obtains higher scores on precision and F, which means our method filters the false-positive IOIs more effectively.

2) Though the task of salient object segmentation has the same format of input and output as IOID, S4Net obtains worse performance on both precision and recall as compared to our method. It illustrates that IOID is a novel task, which cannot be effectively handled with the existing methods for other related tasks.

3) Mask R-CNN retains all the segmented instances, but its recall score is only 37.14%, which means over 60% IOIs in groundtruth cannot be extracted. Though our method improves the effect of instance extraction by combining instance segmentation and semantic segmentation, there are still many missed IOIs as shown in Figure 4. Hence, current instance extraction strategies limit the performance of IOID seriously, and more effort is required to provide better instance extraction effect for IOID.

Table 5: Comparison of our method and the state-of-the-art methods.

Method	precision	recall	F	recall*	F^*
Mask R-CNN [14]	41.48	**37.14**	40.39	**100.00**	47.95
Frequency	50.36	32.76	44.81	88.19	55.90
S4Net [9]	40.70	18.63	31.96	**100.00**	47.16
Our	**68.47**	30.15	**52.95**	49.80	**63.02**

Figure 5 shows some qualitative examples of the IOID results generated by our method and the compared ones. It shows that our method outperforms the state-of-the-art methods.

6 CONCLUSIONS

We proposed a novel task named IOID, which aims to detect all instances of interest in a given image. To handle the challenge in IOID, we proposed a IOID method consisting of instance extraction, interest estimation and IOI selection. Moreover, we constructed the first IOID dataset containing 45,000 images with manually labeled IOIs. The experimental results on the dataset demonstrate that the existing methods for the related tasks cannot effectively handle IOID task and our method outperforms the state-of-the-art baselines.

ACKNOWLEDGEMENTS

This work is supported by National Science Foundation of China (61202320, 61732007), Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20180307151516166), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned Salient Region Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision*.

[3] Sergiu Nedevschi Arthur Daniel Costea, Andra Petrovai. 2018. Fusion Scheme for Semantic and Instance-level Segmentation. In *International Conference on Intelligent Transportation Systems*.

[4] Liangchieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P Murphy, and Alan L Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets and Atrous Convolution and and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2018), 843–848.

[5] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. 2018. Reverse Attention for Salient Object Detection. In *European Conference on Computer Vision*.

[6] Xinlei Chen and C Lawrence Zitnick. 2015. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Conference on Empirical Methods in Natural Language Processing*.

[8] Andrew T Duchowski. 2007. Eye Tracking Methodology: Theory and Practice. In *Springer Science and Business Media*.

[9] Ruochen Fan, Qibin Hou, Mingming Cheng, Taijiang Mu, and Shimin Hu. 2017. S4Net: Single Stage Salient-Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[10] Hao Fang, Saurabh Gupta, Forrest N Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, C Lawrence Zitnick, and Geoffrey Zweig. 2015. From Captions to Visual Concepts and Back. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[11] Liang Lin Guanbin Li, Yuan Xie and Yizhou Yu. 2017. Instance-Level Salient Object Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Dan Witzner Hansen and Qiang Ji. 2010. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3 (2010), 478–500.

[13] Jonathan Harel, Christof Koch, and Pietro Perona. 2006. Graph-Based Visual Saliency. In *Annual Conference on Neural Information Processing Systems*.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross B Girshick. 2017. Mask R-CNN. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[15] Cosmin Toca Anca Ioana Petre Carmen Patrascu Mihai Ciuc Ionut Ficiu, Radu Stilpeanu. 2018. Automatic Annotation of Object Instances by Region-Based Recurrent Neural Networks. In *International Conference on Intelligent Computer Communication and Processing*.

[16] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.

[17] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. SALICON: Saliency in Context. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[18] Trevor Darrell Jonathan Long, Evan Shelhamer. 2015. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[19] Kevin Mcguinness Noel E Oconnor Jordi Torres Elisa Sayrol Xavier Giro I Nieto Junting Pan, Cristian Cantonferrer. 2017. SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[20] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[21] Vahid Kazemi and Ali Elqursh. 2017. Show and Ask and Attend and Answer: A Strong Baseline for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[22] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollr. 2018. Panoptic Feature Pyramid Networks. In *arXiv:1901.02446*.

[23] Alexander Kirillov, Kaiming He, Ross B Girshick, Carsten Rother, and Piotr Dollar. 2018. Panoptic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[24] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra M Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[25] Philipp Krahenbuhl and Vladlen Koltun. 2011. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Annual Conference on Neural Information Processing Systems*.

[26] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. 2017. Instance-Level Salient Object Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[27] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. 2019. Attention-guided Unified Network for Panoptic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[28] Tsungyi Lin, Piotr Dollar, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. 2017. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[29] Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2015. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.

[30] Nian Liu, Junwei Han, and Minghsuan Yang. 2018. PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[31] Zhiming Luo, Akshaya Kumar Mishra, Andrew Achkar, Justin A Eichel, Shaozi Li, and Pierremarc Jodoin. 2017. Non-local Deep Features for Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[32] Xiaowei Hu Ali Borji Zhuowen Tu Philip H S Torr Qibin Hou, Mingming Cheng. 2017. Deeply Supervised Salient Object Detection with Short Connections. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[33] Eleonora Vig, Michael Dorr, and David D Cox. 2014. Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[35] Alex Graves Koray Kavukcuoglu Volodymyr Mnih, Nicolas Heess. 2014. Recurrent Models of Visual Attention. In *Annual Conference on Neural Information Processing Systems*.

[36] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R Dick, and Anton Van Den Hengel. 2016. What Value Do Explicit High Level Concepts Have in Vision to Language Problems. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[37] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*.

[38] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. 2013. Hierarchical Saliency Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[40] Shuai Zheng, Sadeep Jayasumana, Bernardino Romeraparedes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H S Torr. 2015. Conditional Random Fields as Recurrent Neural Networks. In *International Conference on Computer Vision*.