Rethinking Object Saliency Ranking: A Novel Whole-Flow Processing Paradigm

Mengke Song, Linfeng Li, Dunquan Wu, Wenfeng Song, and Chenglizhao Chen

Abstract— **Existing salient object detection methods are capable of predicting binary maps that highlight visually salient regions. However, these methods are limited in their ability to differentiate the relative importance of multiple objects and the relationships among them, which can lead to errors and reduced accuracy in downstream tasks that depend on the relative importance of multiple objects. To conquer, this paper proposes a new paradigm for saliency ranking, which aims to completely focus on ranking salient objects by their "importance order". While previous works have shown promising performance, they still face ill-posed problems. First, the saliency ranking ground truth (GT) orders generation methods are unreasonable since determining the correct ranking order is not well-defined, resulting in false alarms. Second, training a ranking model remains challenging because most saliency ranking methods follow the multi-task paradigm, leading to conflicts and trade-offs among different tasks. Third, existing regression-based saliency ranking methods are complex for saliency ranking models due to their reliance on instance mask-based saliency ranking orders. These methods require a significant amount of data to perform accurately and can be challenging to implement effectively. To solve these problems, this paper conducts an in-depth analysis of the causes and proposes a whole-flow processing paradigm of saliency ranking task from the perspective of "GT data generation", "network structure design" and "training protocol". The proposed approach outperforms existing state-of-the-art methods on the widely-used SALICON set, as demonstrated by extensive experiments with fair and reasonable comparisons. The saliency ranking task is still in its infancy, and our proposed unified framework can serve as a fundamental strategy to guide future work. The code and data will be available at *https://github.com/MengkeSong/Saliency-Ranking-Paradigm*.**

Index Terms— **Object saliency ranking, adaptive circulative bagging, deep learning.**

Manuscript received 15 May 2023; revised 22 September 2023 and 5 November 2023; accepted 5 December 2023. Date of publication 15 December 2023; date of current version 20 December 2023. This work was supported in part the by the National Natural Science Foundation of China under Grant 62172246 and Grant 62102036, in part by the Youth Innovation and Technology Support Plan of Colleges and Universities in Shandong Province under Grant 2021KJ062, in part by the Beijing Natural Science Foundation under Grant 4222024, in part by the Research and Development Program of Beijing Municipal Education Commission under Grant KM202211232003, and in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems under Grant VRLAB2022A02. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ming-Ming Cheng. *(Corresponding author: Chenglizhao Chen.)*

Mengke Song is with the Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 257061, P. R. China.

Linfeng Li and Dunquan Wu are with the College of Computer Science and Technology, Qingdao University, Qingdao 266071, P. R. China.

Wenfeng Song is with the Computer School, Beijing Information Science and Technology University, Beijing 100029, P.R. China.

Chenglizhao Chen is with the Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 257061, P.R. China, and also with the Jiangsu Key Lab of Image and Video Understanding for Social Security and Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, P. R. China (e-mail: cclz123@163.com).

Digital Object Identifier 10.1109/TIP.2023.3341332

I. Introduction

SALIENCY detection is a fundamental task in computer vision. Previous research primarily focuses on **s**alient **o**bject **d**etection (SOD) task [1] and **e**ye **f**ixation **p**rediction (EFP) task [2]. SOD defines saliency in an absolute way using binary object-level pixel-wise saliency maps (Fig. 1-A(a)), while EFP aims to predict scattered human eye fixations (Fig. 1-A(b)). However, existing models for these tasks can only learn to detect all salient objects in an image equally without explicitly differentiating their different degrees of saliency (importance).

Therefore, **s**aliency **r**anking (SR) task is proposed to predict relative saliency of objects in a scene (Fig. 1-A(c)), which enables us to clearly distinguish which object is more salient and know the relative importance among objects, thus benefiting downstream tasks, such as image compression [3] (Fig. 1-B, C) and image retrieval [4] (Fig. 1-D). Unlike SOD, SR is more fine-grained since it assigns unique relative saliency (importance) ranking orders to all salient objects based on their visual saliency and differentiates them by their saliency (importance) degree. Therefore, it can provide more detailed visual information regarding the interrelationships of salient objects rather than simply detecting them.

Despite the promising performance of existing SR methods [5], [6], [7], they still encounter several ill-posed issues that may result in biased and inaccurate results. We can categorize these issues into three types:

*Issue 1: The **g**round-**t**ruths (GT) of the ranking orders are improperly generated* (Fig. 2-A). Existing saliency ranking GT orders generation methods are mainly fixation points-based (Mark ①), fixation maps-based (Mark ②) and attention shift-based (Mark ③). Among them, fixation points-based methods [8] assign the saliency ranking GT orders by calculating the number of fixation points within an object, while fixation maps-based methods [7], [9] count the average/maximum pixel values from the fixation map of an object, since the fixation maps, displaying the distribution and density of multiple fixation points across a visual stimulus, are smoother than the fixation points which represents discrete locations where the eyes fixate. Besides, attention shift-based methods [6], [10], [11] assign descending saliency scores to

1941-0042 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

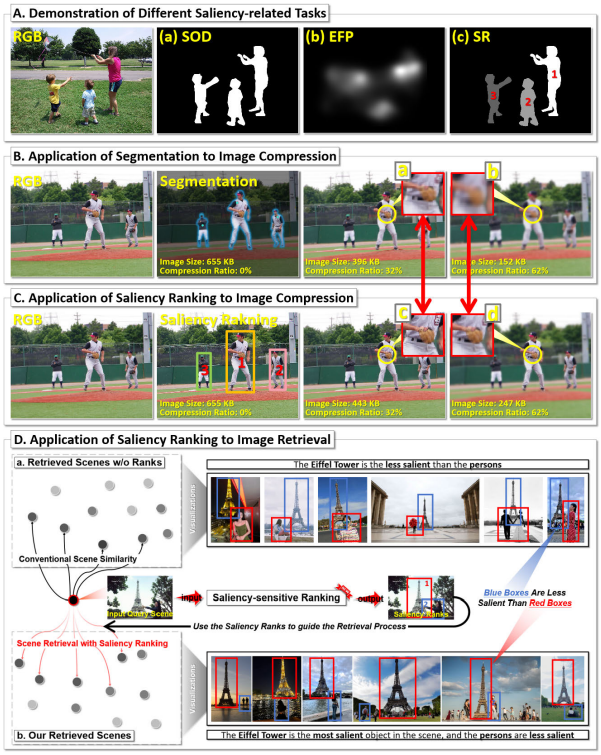

Fig. 1. Three saliency related tasks (A) regarding **s**alient **o**bject **d**etection (SOD), **e**ye **f**iaxtion **p**rediction (EFP), and **s**aliency **r**anking (SR). **B**, **C** and **D** provide a practical application of saliency ranking to image compression and image retrieval. Segmentation-based image compression (**B**) reveal that each object's importance or saliency degree is disregarded during the segmentation task. Consequently, when applying image compression, all objects undergo the same compression effect, leading to a certain loss of information in the more salient objects. Saliency ranking-based image compression (**C**) ensures that the compression effect on objects is inversely proportional to their saliency, resulting in higher preservation of information for more salient objects (comparing subfigure **B**-a and **C**-c, or **B**-b and **C**-d, in red boxes). **D**-a: existing semantically similar-based image retrieval results; **D**-b: our saliency-sensitive ranking-based image retrieval results. With the assistance of saliency ranking, our method can retrieve images based on saliency-sensitive rankings, making the retrieval results more fine-grained and accurate.

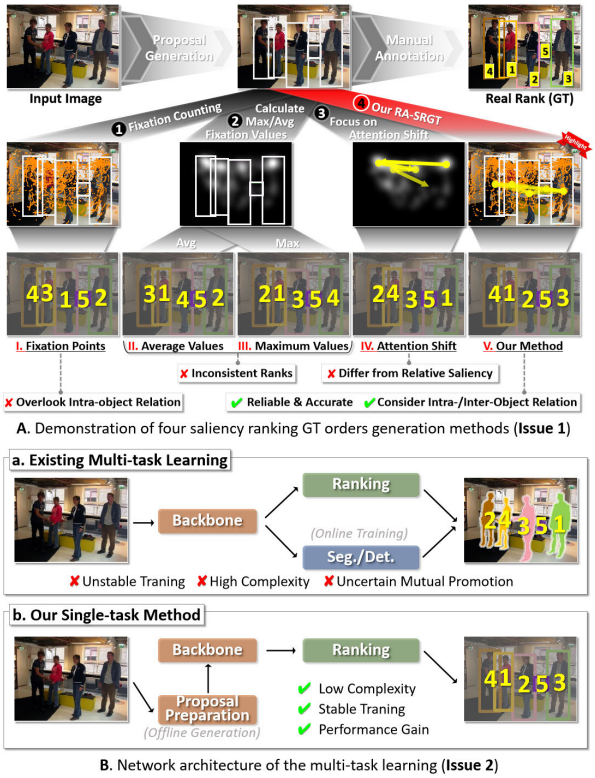

Fig. 2. In subfigure **A**, Marks ① → ④ denote fixation points-based, fixation maps-based, attention shift-based and our novel **r**elationship-**a**ware **s**aliency **r**anking **GT** orders generation method (RA-SRGT), respectively. Compared with other existing methods, our proposed RA-SRGT boosts the reliability and accuracy of saliency ranking GT orders generation. Subfigure **B** is the network architecture of the multi-task learning in saliency ranking task (a) and our proposed single-task method (b) (**Issue 2**).

objects based on the order of "attention shift[1]" of observers over different objects in an image. We will detail these three methods in the following from the perspective of "GT data generation".

1) Fixation points-based saliency ranking methods rely on the discrete points of eye fixations within the objects acquired from human observers to assign object saliency ranking orders, which is implemented by counting the total number of fixation points within each object and assigning a higher saliency ranking order to objects with more fixation points. However, this naive saliency ranking order assignment may lead to inaccurate results. The reason is that these approaches only consider intra-object relationships, meaning that they focus on isolated objects without considering inter-object relationships. This limitation prevents fixation points-based methods from accurately determining how objects relate to each other or the overall composition of the image. For example, Mark ① in Fig. 2-A assigning GT saliency ranking orders based on fixation points within an object leads to the wrong order.

[1]Attention shift order refers to the process of observers gazing from one object to another, which can reflect their interest and preference for the image content. The final saliency rank is an average across the saliency rank orders of multiple observers.

2) Fixation maps-based saliency ranking methods use fixation maps to assign object saliency ranking orders. These methods calculate the maximum or average pixel values within each object and assign a higher saliency ranking order to objects with larger pixel values. However, these methods suffer from inconsistent ranking results across different saliency ranking GT orders generation methods and are sensitive to noise occurring in fixation maps. For example, in Fig. 2-A (II, III), when calculating the average or maximum pixel values from the fixation map (Mark ②), indistinct objects may be deemed as more salient.

3) Regarding attention shift-based methods (Mark ③), it has been noted by [7] that the degree of an object's saliency mainly depends on the gaze duration (Fig. 2-A (IV)), *i.e.*, how humans sequentially select and shift attention from one object to another. It mainly relies on binary saliency prediction to rank objects rather than truly simulating the human attention shift process. Thus, attention shift differs mostly from relative saliency and is very close to the scan path prediction [12].

It is important to note that all three types of saliency ranking GT orders generation methods can be problematic and result in inferior outcomes, adversely affecting model training and evaluation. However, another issue regarding "network structure design" arises in the saliency ranking task.

Issue 2: The effect of multi-task learning in saliency ranking and segmentation or detection is still an open question. Most existing saliency ranking methods adopt multi-task learning ("segmentation + ranking" or "object detection + ranking").

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

However, these combinations of tasks differ greatly in terms of datasets, models, and evaluation metrics. In particular, existing object detection methods already perform well; for the saliency ranking task, multi-task learning (Fig. 2-B (a)) could bring few benefits. Instead, it increases the complexity of the model (See Table IV for detailed proof) and leads to unstable training. Furthermore, the performance of multitask is not necessarily mutually enhance each other. Detailed analysis of the issue is presented in Sec. IV-G. Apart from the issue of multi-task learning, there exists another issue. We will detail it in the following.

Issue 3: Regression-based saliency ranking methods are unsuitable for ranking models. Existing saliency ranking methods rely heavily on regression-based instance segmentation, which is pixel-wise and requires a large quantity of data to achieve accurate results. These approaches can be quite complex and difficult to implement effectively.

Targeting the abovementioned issues in existing saliency ranking methods, this paper proposes a whole-flow processing paradigm of saliency ranking task, *i.e.*, "**GT data generation**" (**Issue 1**) → "**network structure design**" (**Issue 2**) → "**training protocol**" (**Issue 3**), which could serve as a new fundamental strategy for future saliency ranking work.

To address **Issue 1**, we propose a novel **r**elationship-**a**ware **s**aliency **r**anking **GT** orders generation method (RA-SRGT, Mark ④ in Fig. 2-A). This is designed to overcome the limitations of existing saliency ranking GT orders generation methods, which produce inconsistent and varied results that do not align with the principles of the human visual attention system (Fig.2-A (V)). Compared to existing methods, our new GT orders generated by RA-SRGT are more precise and rational, and can be derived directly from fixation points without requiring re-labeling.

Regarding **Issue 2**, we offer a simple yet efficient solution for saliency ranking tasks that avoids the need for complicated multi-task learning approaches. We avoid complex training while retaining high detection accuracy by leveraging already high-performance and pre-trained object detection models to obtain object proposals offline. Our adaptable approach makes it a valuable alternative for real-world applications.

To solve the **Issue 3**, we propose to utilize pure object-wise classification to predict saliency ranking orders. This approach has several advantages over pixel-wise instance segmentation-based regression, since regarding data amount, object proposals are sparser than pixel-wise instance masks, which are better suited for classification than regression. However, the flexible object input issue can arise with classification-based methods, which we address with our newly-devised **a**daptive **c**irculative **b**agging (ACB, Sec. III-C.3), allowing for varying number of input proposals generated by any object detection model and making it more applicable to real-world scenarios.

The conducted experiments have demonstrated the superiority of the proposed approach when compared to state-of-the-art methods. The obtained results highlight the effectiveness of the unified framework in accurately ranking salient objects.

To sum up, the main contributions of this work include:

- We have pioneeringly conducted a thorough analysis of the three fundamental issues that commonly occur in existing saliency ranking methods and proposed a novel whole-flow processing paradigm for saliency ranking;
- We propose a brand-new saliency ranking GT orders generation method which exhibits greater conformity with the real ranking orders (GT data generation);
- We propose a general solution to the challenges of multi-task learning in saliency ranking. The proposed method is a novel proposal-based single-task approach that uses offline object proposals and pure classification to accurately rank objects and reflect human visual perception, which offers unparalleled simplicity, efficiency, and reliability (network structure design);
- We present a novel adaptive circulative processing approach that can handle varying numbers of input proposals from any object detection model (training protocol);

II. Related Work

A. Salient Object Detection

Previous **s**alient **o**bject **d**etection (SOD) works have mainly used handcrafted features (*e.g.*, color contrast [13], background prior [14], and fixations [15]) to detect salient objects, which limits their generalization ability and performance in complex scenarios. However, the emergence of deep learning has led to the development of CNN-based SOD [16], [17], [18], [19], [20], [21], [22] methods that can better capture image representations. Although these methods have achieved impressive results, they still have computation and memory cost limitations. To address this, FCN-based methods [23], [24] have been proposed, which formulate SOD as a pixel-wise binary classification task. Recently, several works [25], [26] have integrated features from multiple layers of CNN to exploit context information at different semantic levels. In addition, some researchers [27], [28], [29], [30], [31] have used depth images as auxiliary information to improve RGB-D SOD performance, but this may be limited in adverse conditions. The development of thermal infrared sensors has facilitated the progress of RGB-T SOD [32], [33], which takes advantage of temperature cues to improve performance in complex scenes. These approaches have attracted much attention and are effective in challenging environments.

B. Eye Fixation Prediction

Eye **f**ixation **p**rediction (EFP), different from SOD, aims to predict where people get interested in natural scenes. Early EFP models [34], [35] used low-level features such as contrast, color, and brightness. However, recent advances in deep neural networks have allowed for learning high-level, top-down features, resulting in significant performance improvements. The SALICON dataset [36], which contains many real human eye observation points, has been used for EFP research [37], [38]. Generative adversarial networks [39] and new evaluation metrics have also been introduced into the field. Domain adaptation techniques [40] have been proposed to model image and video tasks in a unified way. However, predicting how objects relate to each other is more challenging than predicting where they are in scenes. This requires achieving two objectives: accurate localization with salient objects and mutual ranking of objects.

C. Saliency Ranking

The salient ranking is a newly proposed problem [5], [25], [41] to determine the relative order of different saliency

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

objects. Previous works have made steady progress in this field by adopting various approaches, such as human attention shifts [6], graph neural reasoning modules [7], and object-context interaction information [9]. Apart from this, Lv et al. [8] propose a new camouflage object detection model to rank camouflaged objects, and they adopt computing the fixation points on the instance to label the instance ranks. Fang et al. [11] propose an end-to-end SOR framework and introduce a Position-Preserved Attention module that preserves the coordinates of objects in an image.

However, the ground truth of the ranking orders generated by these methods is improper, and the regression-based multi-task learning approach may cause mutual interference between irrelevant tasks. Moreover, implementing regression-based instance segmentation is complex and challenging. To address the challenges faced in generating saliency ranking GT orders and conducting regression-based multi-task learning, we propose a comprehensive approach that considers three key areas: GT data generation, network structure design, and training protocol. By addressing these areas, we aim to create a whole-flow processing paradigm for the saliency ranking task.

III. The Proposed Method

A. Method Overview

The main objective of our research is to propose a whole-flow processing paradigm of saliency ranking task (refer to Fig. 3 for a better understanding) from the perspective of "**GT data generation**", "**network structure design**" and "**training protocol**" by addressing the three main issues that arise when generating saliency ranking GT orders (**Issue1**) and conducting regression-based multi-task learning (**Issue2** and **Issue3**). To address **Issue1**, we propose a **r**elationship-**a**ware **s**aliency **r**anking **GT** orders generation method (RA-SRGT, Sec. III-B) that utilizes varying levels of saliency thresholds to produce saliency ranking GT orders that closely align with the underlying principles of the **h**uman **v**isual **s**ystem (HVS). This is achieved by considering the interrelationships between image regions and incorporating HVS principles into the GT generation process. To address the interference (**Issue2**) of multi-task learning and complex implementation (**Issue3**) of regression techniques used in existing saliency ranking methods, we propose a brand-new approach that involves a newly-devised, simple yet efficient proposal-based single-task framework called **f**lexible **o**bject **s**aliency **r**anking network (FOSRNet, Sec. III-C). Based on classification to better deal with sparser proposals, our approach uses off-line proposals to enhance detection accuracy while reducing computational resources and training complexity compared to existing online proposals generation methods (Sec. III-C.1). Moreover, we introduce an **a**daptive **c**irculative **b**agging (ACB, Sec. III-C.3) method to solve the problem of flexible object inputs in the classification task. We will provide detailed explanations of each component in the following sections.

B. Relationship-Aware Saliency Ranking GT Order Generation

Salient object detection and eye fixation prediction can produce saliency maps of multiple objects. Still, these maps are limited in their ability to express these objects' relationships and degrees of saliency. Since these methods can only locate

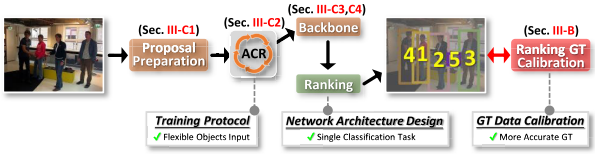

Fig. 3. Pipeline of our whole-flow processing paradigm of saliency ranking.

salient areas and do not distinguish which is more salient among them, consequently, some researchers use saliency ranking to address this challenge. However, the saliency ranking GT orders are assigned in various ways, and erroneous GT orders can negatively impact model training and fail to align with human attention and psychology studies (**Issue1**). To further illustrate the impact of different saliency ranking GT orders generation methods, we shall briefly detail two commonly-used methods: fixation points-based and fixation maps-based. Note that shift attention approaches are excluded because they differ from relative saliency.

1) Fixation Points-Based Saliency Anking GT Orders Generation Approaches (Fig. 4-line 2): These methods mainly rely on counting the total number of fixation points within each object and assigning higher saliency ranking orders to objects with more fixation points, *i.e.*, the saliency ranking orders of objects in an image are determined by the number of fixation points and the spatial size of the objects. This can be formulated as:

$$\mathrm{Rank}(\mathrm{Q}_i) = \underbrace{\frac{1}{\sqrt{\mathrm{size}(\mathrm{Q}_i)}}}_{\text{Penalty}} \overbrace{\sum_{(u,v)\in \mathrm{Q}_i} \mathrm{P}(u,v)}^{\text{Total Fixations in } \mathrm{Q}_i}, \tag{1}$$

where Q_i denotes the i-th object; $\mathrm{Rank}(\mathrm{Q}_i) \in \{1, 2, \ldots, n\}$ (n is the total number of objects in an image scene) returns the saliency ranking order of Q_i; $\mathrm{Size}(\mathrm{Q}_i)$ measures the spatial size of Q_i, *i.e.*, width×height; $\mathrm{P}(u, v) \in \{0, 1\}$ indicates whether there exists a fixation point at spatial coordinate (u, v). However, these fixation points-based approaches only consider intra-object relationships, which means they focus on isolated objects without considering inter-object relationships, *i.e.*, how those objects relate to one another and the overall composition of the image, leading to inaccurate saliency ranking orders (Mark ① in Fig. 2-A).

2) Fixation Maps-Based Saliency Ranking GT Orders Generation Approaches (Fig. 4-line 3): These methods assign saliency ranking GT orders by calculating the maximum/average pixel values of each object in fixation maps and assign higher saliency ranking orders to objects with larger pixel values. This process can be denoted as follows:

$$\mathrm{Rank}(\mathrm{Q}_i) = \underbrace{\mathrm{MAX}\big(\mathrm{V}(u,v)\big)}_{\text{Maximum pixel values}} \text{ or } \underbrace{\mathrm{AVG}\big(\mathrm{V}(u,v)\big)}_{\text{Average pixel values}}, \tag{2}$$

where $\mathrm{V}(u, v) \in \{0, 255\}$ denotes the pixel value at spatial coordinate (u, v) with the i-th object Q_i in a fixation map. MAX/AVG is to calculate the maximum/average pixel value of Q_i. Whereas, as shown in Fig. 2-A-Mark ②, the biggest problem of fixation maps-based methods is the inconsistent ranking results across different saliency ranking GT orders generation methods and sensitivity to noise occurring in fixation maps.

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

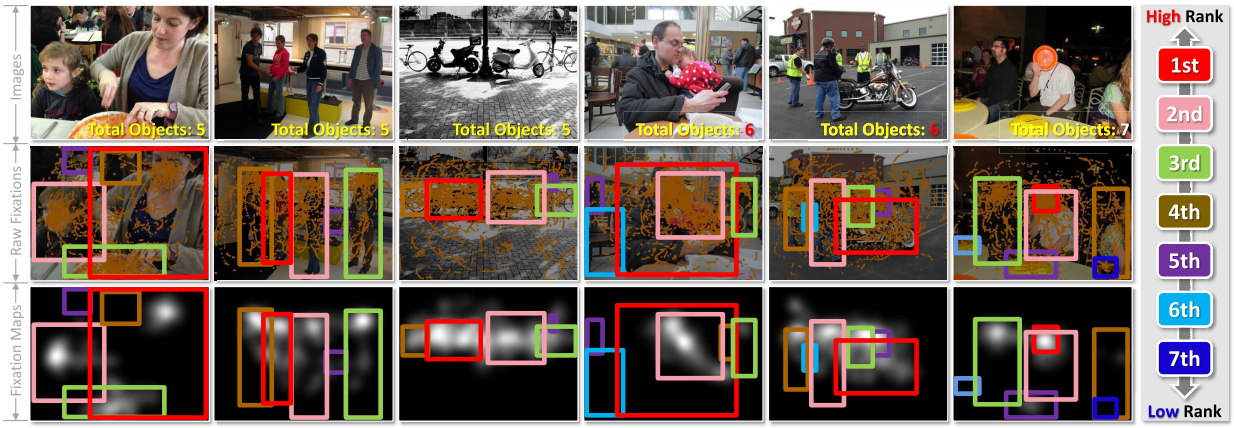

Fig. 4. Visual comparison between raw fixation points (2nd column) and fixation maps (3rd column), which are commonly used in saliency ranking GT orders generation methods. Saliency ranking orders are marked in rectangular boxes with different colors.

As mentioned in Fig. 2-A, the limitations of fixation points-based and fixation maps-based saliency ranking GT orders generation methods result in unreliable and inconsistent saliency ranking GT orders unaligned with the principles of the human visual attention system. These methods overlook the overall image composition and inter-object relationships, crucial in precisely reflecting human attention and psychology.

3) **Relationship-aware saliency ranking GT orders generation**. To address the limitation mentioned above, from the perspective of "**GT data generation**", we propose a novel method called **r**elationship-**a**ware **s**aliency **r**anking **GT** orders generation (RA-SRGT). This method is fixation points-based and considers intra-object and inter-object relationships among objects, enabling a comprehensive evaluation of saliency ranking. By considering the entire image composition, RA-SRGT generates reliable and consistent saliency ranking orders that accurately reflect human attention and psychology.

Intuitively, there are two ways to implement our RA-SRGT. The first way involves binarizing the fixation maps after graying and counting the white pixels of each object. The percentage of each object in the entire image is then calculated, and the final saliency ranking order is obtained by multiplying these two parts. The whole process can be defined as:

$$\mathrm{Rank}(\mathrm{Q}_i) = \mathrm{Sort}\left(\sum_{i=1}^{n} \frac{\mathrm{P}_{\mathrm{Q}_i}^{white}}{\mathrm{P}_{\mathrm{Q}_i}^{total}} \times \frac{\sqrt{\mathrm{size}(\mathrm{Q}_i)}}{\sqrt{\mathrm{size}(\mathrm{Q})}}\right), \tag{3}$$

where Q_i denotes the i-th object of the image scene Q; $\mathrm{Rank}(\mathrm{Q}_i) \in \{1, 2, \ldots, n\}$ (n is the total number of objects in an image scene) is the final saliency ranking GT order of Q_i; $\mathrm{Sort}(\cdot)$ represents the sorting operation, returning the saliency ranking order of Q_i; $\mathrm{Size}(\cdot)$ measures the spatial size; $\mathrm{P}_{\mathrm{Q}_i}^{white}$ and $\mathrm{P}_{\mathrm{Q}_i}^{total}$ separately indicate the number of white pixels and all pixels within the binary fixation map of Q.

Compared to existing fixation maps-based saliency ranking GT orders generation approaches, this binarized fixation maps-based scheme considers both intra-object and inter-object relationships among objects. But it still has a limitation, as it is sensitive to noise. The binary saliency maps with noise may not accurately reflect each object's true saliency, leading to ranking errors. Moreover, the binary threshold may vary depending on the saliency maps' quality and content, making it hard to generalize.

Therefore, we adopt the second option, which utilizes GT saliency annotations, *i.e.*, fixation points generated by mouse trajectories provided in the SALICON set [36]. These fixation points-based saliency annotations are used for eye fixation prediction tasks, and we resort to them to work for our saliency ranking. The main superiority of the fixation points-based way to implement RA-SRGT compared to the above-mentioned binarized fixation maps-based way is that counting fixation points are more reliable and controllable than binarizing fixation maps, which requires selecting suitable binary thresholds. Moreover, the key difference between our fixation points-based RA-SRGT and the existing fixation points-based saliency ranking GT orders generation methods mentioned above is that we do not rely solely on the number of fixation points. Instead, we consider the percentage of each object in the entire image.

Unlike the first scheme to count white pixels within binarized fixation maps, we count the fixation points of each single rectangle proposal, which object detection models generate. The motivation behind this is to provide a more accurate and reliable way to rank the saliency of objects in an image. By taking into account the relative importance of each object in the image, this approach is less sensitive to noise and threshold values, which can vary depending on the image's content and quality. Additionally, this approach reflects that the human visual system pays attention to objects that are contextually important rather than simply visually salient. In summary, our approach follows the process outlined below:

$$\mathrm{S}(\mathrm{Q}_i) = \begin{cases} \overbrace{\sum_{(u,v)\in \mathrm{Q}_i} \mathrm{P}(u,v)}^{\text{Total Fixations in } \mathrm{Q}_i} + \underbrace{\gamma \cdot \mathrm{e}^{\beta \cdot \sqrt{\mathrm{size}(\mathrm{Q}_i)}}}_{\text{Penalty}}, & \text{if } \mathrm{N}_i > 0, \\ 0, & \text{if } \mathrm{N}_i = 0, \end{cases} \tag{4}$$

where $\mathrm{S}(\mathrm{Q}_i)$ returns the combined score of the i-th object Q_i; $\mathrm{P}(u, v) \in \{0, 1\}$ indicates whether there exists a fixation point at spatial coordinate (u, v) of Q_i; $\sqrt{\mathrm{size}(\mathrm{Q}_i)}$ means the spatial size of Q_i. To mitigate the influence of various factors on the ranking scores, we propose a set of thresholds γ and β. The purpose of γ is to serve as the GT threshold (ablation study can be found in Fig. 11-B), while β is to align $\mathrm{P}(u, v)$ numerically, a value that can be considered as being hidden within γ (ablation study can be found in Fig. 11-C). Specifically, γ controls the extent to which the spatial size of the object affects the ranking, while β determines the impact of the fixation points. By varying the values of γ and β, we can

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

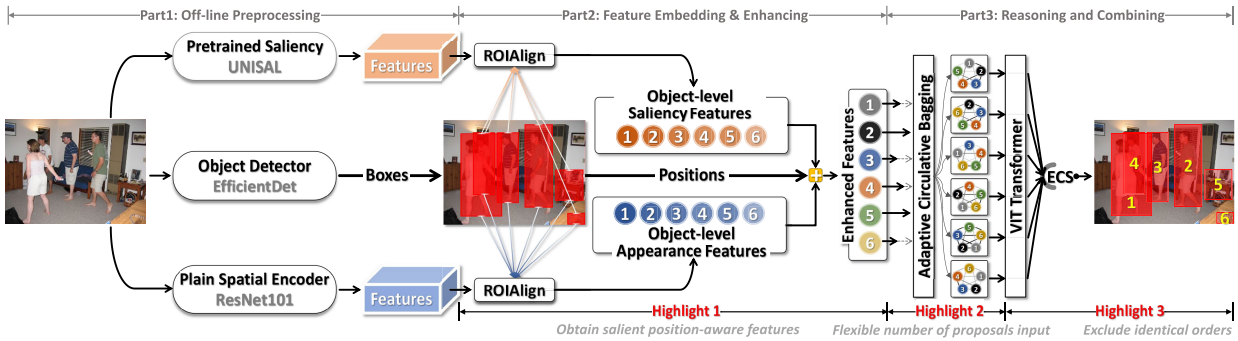

Fig. 5. The pipeline of our flexible object saliency ranking network. The ranking model mainly consists of three parts: Part1) off-line preprocessing to prepare off-line proposals and features for later processing; Part2) feature embedding & enhancing to embed and improve the object-level appearance and saliency features; Part3) reasoning and combining to settle down the flexible multiple objects input by the **a**daptive **c**irculative **b**agging (ACB) and implement classification with an **e**xclusive **c**lassification module (ECS) to avoid the same ranking order existing in traditional classification tasks. "⊕" is the feature concatenation operation.

analyze how these factors interact and identify the optimal combination for generating the saliency ranking GT orders.

Since we use object proposals instead of instance segmentation maps to represent objects (see Sec. III-C.1 for a reason), some proposals may have no fixation points. Therefore, if the fixation number of a proposal is zero, we label all such proposals as novel ranking orders, *i.e.*, 0. Then we can get the final saliency ranking GT orders $\mathrm{Rank}(\mathrm{Q}_i)$ of Q_i by applying:

$$\mathrm{Rank}(\mathrm{Q}_i) = \begin{cases} \underbrace{\mathrm{Sort}\big(\mathrm{S}(\mathrm{Q}_i)\big)}_{i\in(1,n)}, & \text{if } \mathrm{S}(\mathrm{Q}_i) > 0 \\ 0, & \text{if } \mathrm{S}(\mathrm{Q}_i) = 0 \end{cases} \tag{5}$$

where $\mathrm{Sort}(\cdot)$ represents the sorting operation, returning the saliency ranking order of Q_i and n is the total number of objects in an image scene.

Summary. Our proposed methods offer a more precise solution for obtaining accurate saliency ranking GT orders than fixation maps, which rely on arbitrary binary thresholds. By using GT saliency annotations in the form of fixation points, we can avoid such issues of sensitivity of noise, as each observer's annotation data is independent and exact. This approach enables us to obtain highly reliable identification of salient regions in an image across multiple observers. Therefore, our newly-proposed methods significantly improve over previous approaches and can provide valuable insights for various applications in computer vision and beyond.

C. Flexible Object Saliency Ranking Network

Multi-task learning refers to training a single model to perform multiple tasks simultaneously. In this approach, the model learns to share the learned representation across different tasks, which can improve the model's overall performance. However, a significant challenge in multi-task learning is that the different tasks may interfere, leading to unstable training and suboptimal performance (**Issue2**). For instance, when combining segmentation/object detection and saliency ranking, the segmentation/detection task may affect the saliency ranking task, making the model more complex and difficult to train. Moreover, current saliency ranking methods rely heavily on regression-based instance segmentation, a pixel-wise approach that demands a large amount of data to yield precise outcomes. This technique can be intricate and challenging to implement effectively, posing a significant challenge (**Issue3**).

To address these issues, we propose a single-task-based framework called **f**lexible **o**bject **s**aliency **r**anking network (FOSRNet) from the perspective of "**network structure design**". This framework simplifies the multi-task learning process by training the model to simultaneously perform a single saliency ranking task. It also solves the complex of regression-based methods and saliency ranking tasks by a classification-based architecture.

As visualized in Fig. 5, FOSRNet consists of three main parts: off-line preprocessing (**Part1**), feature embedding & enhancing (**Part2**), and reasoning and combining (**Part3**). In **Part1**, we generate object proposals offline. We avoid complicated training while retaining high detection accuracy by leveraging the already high-performance and pre-trained object detection models to obtain object proposals offline. In **Part2**, we utilize a pre-trained saliency model (*e.g.*, UNISAL) and plain spatial encoder (*e.g.*, ResNet101) to provide object-level saliency features and appearance features, respectively, and jointly embed them to object-level enhanced salient position-aware features for latter classification. Finally, in **Part3**, we address the issue of complex implementation of regression-based ranking models. To do so, we employ **e**xclusive **c**lassification (ECS) as described in Section III-C.3. ECS ensures that different objects are predicted with unique saliency ranking orders, which is necessary for a classification task. This approach helps to improve the accuracy and reliability of the classification results. Particularly, we use **a**daptive **c**irculative **b**agging (ACB, Sec. III-C.3) to process the flexible number of input proposals circulatively, which is more applicable to real-world scenarios where the input is dynamic and complex. We will elaborate on each part in the following.

1) Off-Line Preprocessing: Most existing saliency ranking methods are based on instance segmentation or object detection tasks, as shown in Fig. 6. Among them, instance segmentation tasks require producing pixel-wise masks that can be challenging. The ranking order is highly sensitive to the quality of the instance masks, and inaccurate ranking orders can result from poor-quality masks. And object detection tasks tend to perform online proposal generation, which increases the complexity of model training and leads to inferior saliency ranking results.

Therefore, instead of relying on pixel-wise instance segmentation and heavy online object detection, we focus on off-the-shelf object-wise proposals (Fig. 5-**Part1**), which is a simpler

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

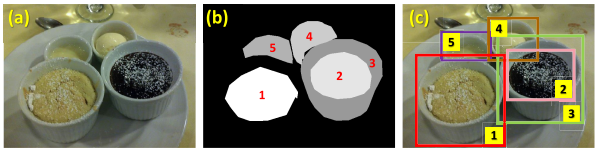

Fig. 6. Comparison of the saliency ranking orders between instance segmentation-based (b) and object detection-based saliency ranking (c).

task that can be effectively performed offline. To achieve off-line generation, we use an off-the-shelf state-of-the-art (SOTA) object detection method, such as EfficientDet[2] [42] to decompose the given image into at least five rectangular object proposals.[3] Then, filtering out overlapping object proposals with low **i**ntersection **o**ver **u**nion (IOU) rate between each pair of proposals. Additionally, we discard extremely large object proposals (those with an object proposal area proportion greater than 60% of the whole image area) and extremely small object proposals (those with a sum of pixel values less than 20). All these processes are conducted offline, avoiding complicated training while retaining high detection accuracy. We also resort to a pre-trained saliency model and plain spatial encoder to obtain enhanced salient position-aware features for the latter classification, which will be detailed later.

2) Feature Embedding & Enhancing: After obtaining proposals from Sec. III-C.1, several feature backbones are available, *e.g.*, CNN, RNN, Transformer, GNN, and MLP. Here we select Transformer (*e.g.*, VIT [43]) as our feature extractor due to its powerful modeling of long-range dependency, which is significant to explore the inter-object relationships in a scene. Unlike traditional plain patch embedding methods such as image-level partition or CNN-based feature-level partition in Transformer, we propose to infuse salient position-aware knowledge inherited from fixations and deep appearance features from deep feature extractors to the embedded features (Fig. 5-**Part2**). This knowledge is position-aware and saliency-aware, meaning it can establish the relationships between the individual proposal and the whole image, which can help accurately rank the objects.

The feature embedding and enhancing process shown in Fig. 7 comprises two parts: a **p**retrained **s**aliency **m**odel (PSM) and a **p**lain **s**patial **e**ncoder (PSE). The PSM extracts object-level saliency features that inherit salient position-aware knowledge from fixations. This is achieved using the off-line position-aware eye fixation prediction model UNISAL [40]. On the other hand, the PSE obtains object-level appearance features that contain deep spatial and semantic information. This is done using the off-the-shelf pre-trained ResNet101 to obtain high-dimensional semantic deep features. The object-level saliency features and appearance features work together to enhance the feature embedding process for improved performance. We will explain them below.

The PSM takes the whole original image I_n as input, outputting f_u, which then is fed into ROIAlign to produce two proposals of different scales: a local feature-level proposal

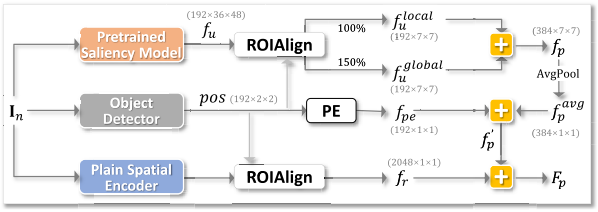

Fig. 7. Feature embed & enhance. PE: Position Embed; ⊕: Concatenate.

f_u^{local} (same size as f_u), and a global feature-level proposal f_u^{global} (50% larger than f_u). The global feature-level proposal can capture more global information from the original image and enhance the proposal features. The ROIAlign operation requires exact proposal location information, which is lost in f_u. We have added extra location information (denoted as $pos \in \mathbb{R}^{192\times2\times2}$, containing coordinates of the top-left and bottom-right corners of each proposal) in it. Then, fuse the two feature-level proposals to obtain salient position prior-oriented feature f_p.[4] This process can be formulated as follows:

$$\begin{aligned} &\underbrace{\text{ROIAlign}\big((f_u),\, 1.5r,\, pos\big)} \\ &\qquad\Downarrow \\ &f_p = \text{Concat}(f_u^{local},\, f_u^{global}),\;\; f_u = \text{PSM}(I_n), \\ &\qquad\Uparrow \\ &\overbrace{\text{ROIAlign}\big((f_u),\;\; r,\;\; pos\big)} \end{aligned} \tag{6}$$

where, r is the region of interest (ROI) from image-level object proposals; Concat$(\cdot)$ denotes feature concatenation; ROIAlign$(\cdot)$ is to extract ROIs and map them to a fixed size.

Note that salient position prior-oriented feature f_p is inherited from f_u generated by PSM, *i.e.*, UNISAL, it naturally contains a small amount of location information. However, the location information is not accurate. To enhance the position information among proposals, we add each proposal's central point coordinate information to the feature, which begins by scaling down the feature-level proposal $f_p \in \mathbb{R}^{384\times36\times48}$ to the size of $f_p^{avg} \in \mathbb{R}^{384\times1\times1}$ through global average pooling. Then, integrate it with position embedding f_{pe}, which represents the center coordinate (x, y) of f_p^{avg}. The object-level saliency feature $f_p' \in \mathbb{R}^{576\times1\times1}$ can be formulated as:

$$f_p' = \text{Concat}(f_p^{avg},\;\; f_{pe}),\;\; f_p^{avg} = \text{AvgPool}(f_p) \tag{7}$$

where AvgPool$(\cdot)$ is the average pooling operation; $f_{pe} = \text{PE}(\cdot)$ denotes position embedding consisting of a 1×1 convolution to make position information pos learnable.

However, though object-level saliency feature f_p' is equipped with enhanced position information, it only contains shallow features without deep semantic information since the PSM is a lightweight architecture that can merely generate shallow position-related features. To overcome, we resort to a **p**lain **s**patial **e**ncoder (PSE) to obtain object-level appearance feature $f_r \in \mathbb{R}^{2048\times1\times1}$ that contain deep spatial and semantic information to compensate for the lack of high-level semantic cues. Finally, a more powerful object-level enhanced salient position-aware feature $F_p \in \mathbb{R}^{2642\times1\times1}$, which contains

[2]Other SOTA object detection models can replace it, and the ablation study is shown in Table. IV-I.1.

[3]We assume that an image has at least five object proposals since most images (nearly 70%) in SALICON set contain about five objects. When the proposed model processes an image containing fewer than five objects, it automatically fills to five objects using empty objects, a.k.a. dummy nodes.

[4]f_p, generated by the output of PSM (f_u), naturally possesses saliency information, since PSM is derived from EFP method, which is also to predict salient regions.

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

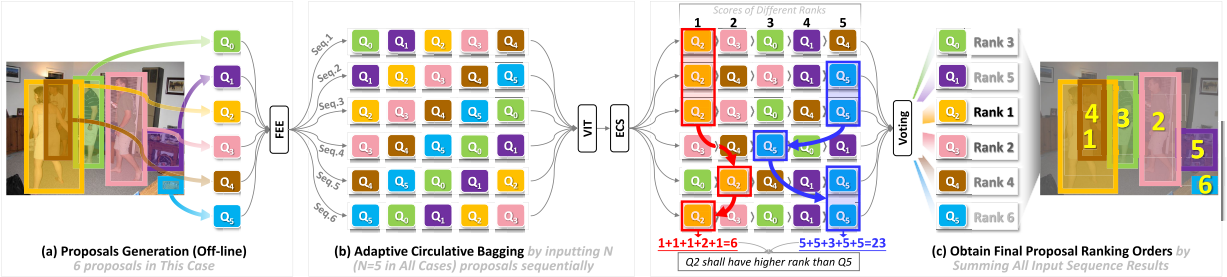

Fig. 8. Pipeline of adaptive circulative bagging. The process includes two phases: adaptive circulative bagging (b) and final proposal ranking orders obtaining (c). Specifically, the times of circulative bagging are determined by the number of proposals, *e.g.*, six times in the case shown in the figure. And each time, we input five proposals in all cases. The final ranks are achieved by summing the ranking score results of each proposal's six times circulative input and voting each proposal's final saliency ranking order according to the ranking scores (smaller scores, smaller orders, higher saliency ranking/importance).

object-level saliency features and deep object-level appearance features simultaneously is generated by fusing f_p' and f_r. It can be formulated as follows:

$$F_p = \text{Concat}(f_p', \ f_r), \tag{8}$$

Then, the object-level enhanced salient position-aware feature F_p will be used to predict saliency ranking orders. Next, we will detail how to use this enhanced feature to implement a saliency ranking task.

3) Reasoning and Combining: As detailed in Sec. III-C.2, we select Transformer (*e.g.*, VIT [43]) as our feature extractor due to its powerful modeling of long-range dependency, which is significant to explore the inter-object relationships in a scene. However, there exists a flexible input proposals issue, which conventional Transformer networks cannot solve. Thus, adapting it to a classification-based saliency ranking method with a fixed number of proposal classes, *i.e.*, a saliency ranking network with a fixed input, remains challenging. This motivates us to conduct further exploration, as seen below.

Adaptive Circulative Bagging. Regression-based saliency ranking methods are a more practical choice in real-world scenarios because they can handle varying numbers of objects in an image. Unlike classification-based methods, which require a fixed number of proposal classes, regression models predict the saliency scores of each object directly based on its visual features. This means they can analyze any number of objects in the image without needing a predefined number. In contrast, classification-based methods may be limited in recognizing a fixed number (*e.g.*, 5) of objects and may not predict saliency ranking orders for scenes with more than or less than the fixed number of objects. To address the limitation of classification-based methods, concerning "**training protocol**", we propose a brand-new method called **a**daptive **c**irculative **b**agging (ACB, Fig. 5-**Part3**) to handle varying numbers of input proposals from any object detection model in a circulative manner.

The ACB each time processes five object-level enhanced salient position-aware features (each corresponds to a proposal within an image) obtained by Feature Embedding & Enhancing (Fig. 5-**Part2**). Suppose there are more than five object-level features. In that case, we start by selecting the object-level features Q_0[5] as the center object-level features and sort the other object-level features based on their distance to Q_0 in ascending order. Then, we process each object-level feature in sequence, separately starting from Q_1, Q_2, and so on, until all object-level features have been processed. To better understand the process, please refer to Fig. 8-(b). The whole processing phase is as follows:

$$\text{Seq}_i = \bigcup_{k=0}^{4} Q_{\{(i+k) \bmod n\}}, \quad 0 \le i \le n-1 \tag{9}$$

where Seq_i refers to the i-th input object-level features list sequence of an image; $\bigcup$ denotes the union operation of object-level feature sequences by order, and $\{(j+k) \bmod i\}$ means circularly taking the remainder. n is the total number of proposals in an image. Before feeding the object-level features list sequence Seq_i into the classification reasoning part, there exists a fatal issue — same classification results (*i.e.*, same saliency ranking orders for objects with different saliency ranking orders in traditional classifications to be solved.

Exclusive Classification. Current saliency ranking methods rely heavily on complex regression-based instance segmentation, a pixel-wise approach requiring significant data to produce accurate results. Consequently, this technique can be challenging, leading to a substantial obstacle (**Issue3**). Thus, we choose a classification to develop effective ranking models.

Since saliency ranking is similar to multiple classification tasks, intuitively, we can use two **f**ully-**c**onnected layers (FC) to generate the final saliency ranking orders directly. However, traditional classification can suffer from a fatal issue, *i.e.*, due to the limitation of the "Max" function used in FC layers to predict classes, it might assign the same classification results (*i.e.*, same saliency ranking orders) for objects with different saliency ranking orders. For example, in a scene with six objects, which should be assigned six distinct saliency ranking orders, inversely, traditional classification tasks might assign two objects with the same order, which does not fit the essence of the saliency ranking task.

Therefore, in order to achieve a more accurate and reliable saliency ranking, we employ an **e**xclusive **cla**ssification (ECS) method to predict the saliency ranking orders of proposals in each iteration. The ECS includes two FC layers, a SoftMax function, and a Hungarian algorithm to handle the same ranking order issue. Since the Hungarian algorithm is used to solve the bipartite graph matching problem, which can ensure that each category contains unique data, thus, solving the problem of identical saliency ranking orders in conventional classification problems.

Each proposal's final saliency ranking order can be collected by: 1) summing the saliency ranking score outputs of the total number of n input object-level features list sequences and 2) voting each proposal's final saliency ranking order (smaller

[5]Here, we start the subscript from 0 instead of 1 to fit better the following formula (eq. 9).

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

scores, smaller orders, higher saliency ranking/importance). As shown in Fig. 8-(c), the whole processing can be denoted as:

$$\begin{aligned}\mathrm{Rank}(\mathrm{Q}_j) = \mathrm{Sort}\Big(\textstyle\sum_{j=0}^{n-1}\sum_{i=0}^{n-1}\Big(\mathrm{Net}\big(\mathrm{Seq}_{i,j}\big)\Big)\Big),\\ \overbrace{\mathrm{ECS}\Big(\mathrm{VIT}\Big(\mathrm{ACB}\big(\mathrm{FEE}(\mathrm{Seq}_{i,j})\big)\Big)\Big)}^{\Uparrow}\\ \overbrace{\mathrm{Hungary}\Big(\mathrm{SoftMax}\big(\mathrm{FC}(\mathrm{F}_g)\big)\Big)}^{\Uparrow}\end{aligned} \tag{10}$$

where $\mathrm{Rank}(\mathrm{Q}_j)$ is the final saliency ranking order of the j-th object-level feature (*i.e.*, proposal) Q_j; $\mathrm{Seq}_{i,j}$ means the i-th input proposal list sequence of j-th proposal of an image, where $i! = (j-1+n)\ mod\ n$; F_g means the output of $\mathrm{VIT}(\cdot)$; $\mathrm{FEE}(\cdot)$ is the feature embedding & enhancing (Sec. III-C.2); $\mathrm{ACB}(\cdot)$ is the adaptive circulative bagging (Sec. III-C.3); $\mathrm{Sort}(\cdot)$ represents the sorting operation, returning the saliency ranking order of Q_j.

4) Loss Function: The essence of our saliency ranking network is a typical classification task; thus, we use CrossEntropy Loss (L_{cls}) and MarginRanking Loss (L_{rank}) to optimize the model. The total loss (L_{all}) is defined as follows:

$$\begin{aligned}\underbrace{-\frac{1}{N}\textstyle\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} t_{i,j}\log(\hat{y}_{i,j})}_{\Downarrow}\\ \mathrm{L}_{all} = \mathrm{L}_{cls} + \alpha\cdot\mathrm{L}_{rank}(\hat{y}_j, \hat{y}_i, z),\\ \overbrace{\textstyle\sum_{i,j}^{N}\max(0, -z\times(\hat{y}_j-\hat{y}_i)+m)}^{\Uparrow}\end{aligned} \tag{11}$$

where $\hat{y}_j$ and $\hat{y}_i$ are the predicted outputs, while z is the true label belonging to either $\{1, -1\}$. "1" indicates that $\hat{y}_j$ has a higher rank than $\hat{y}_i$, and "-1" means that $\hat{y}_j$ has a lower rank than $\hat{y}_i$. The minimum value of the correct ranking difference, denoted by m, is set to "0" in this case.

IV. Experiments

A. Datasets and Evaluation Metrics

Following the approach in [6], we have relabeled the SALICON dataset using our proposed RA-SRGT (Sec. III-B). Specifically, we divided the SALICON validation set into a validation set and a test set since the SALICON test set did not contain fixation points that our RA-SRGT method could utilize. The resulting training, validation, and test sets contained 10, 000, 1, 200, and 3, 800 samples. We adopt the **s**pearman **r**ank-order **c**orrelation **c**oefficient (SRCC) and the F1 score to evaluate the performance of our method. The higher the SRCC and F1 score, the better the ranking performance.

B. Implementation Details

We implement our approach in Python with the Pytorch toolbox on an NVIDIA GTX2080Ti GPU (with 11G RAM). We optimize the network via SGD with a momentum of 0.9 and weight decay of 10^{-4}. The learning rate is set to 0.001 and exponentially decayed by 0.1 after each ten epoch. The batch size is set to 5 since each image has at least five objects. The weights of UNISAL and ResNet101 are frozen during the whole training phase. The number of Transformer encoder blocks is set as 4 in our network.

C. Comparison With State-of-the-Art Methods

1) Comparison With Eye Fixation Prediction Methods: While diverse objects and instances are generated by object detection or segmentation methods among saliency ranking methods, it can be challenging to make a direct comparison. Therefore, we conducted a quantitative comparison mainly with SOTA eye fixation prediction models. To demonstrate the effectiveness of the proposed method, we have compared it with the eight most recent SOTA eye fixation prediction models, *e.g.*, $\mathrm{SalFBNet}_{22}$ [44], $\mathrm{TranSalNet}_{21}$ [45], UNISAL_{20} [40], $\mathrm{EML\text{-}NET}_{20}$ [46], SAM_{18} [47], SalGAN_{18} [39], $\mathrm{ML\text{-}Net}_{16}$ [48], and $\mathrm{SALICON}_{16}$ [36]. To ensure a fair comparison, all quantitative evaluations were conducted using the saliency maps provided by the authors or obtained from the available codes, with parameters unchanged and trained on the SALICON dataset.

Eye fixation prediction methods only generate saliency maps. Therefore, we followed previous saliency ranking methods to obtain saliency ranking orders of other methods by counting the number of white pixels within each object after graying and binarizing the fixation maps. The more white pixels within an object, the more salient the object is. However, the binary threshold obtained by averaging all the values of the grayed whole image in other methods could lead to a false alarm issue, where an object is not salient, but the number of white pixels might still be relatively high.

We propose obtaining a more accurate binary threshold to generate binary maps to ensure a fair comparison. First, we gray (denoted as Θ) the saliency map. Next, we separately calculated the average values of each object in an image by summing the grayed values of an object and dividing the total by the object's area. These average values determine our binary threshold T. The whole process can be formulated as:

$$T = \frac{1}{n\times\lambda}\sum_{i=1}^{n}\frac{\mathrm{Sum}\big(\Theta(\mathrm{Q}_i)\big)}{\sqrt{\mathrm{size}(\mathrm{Q}_i)}}, \tag{12}$$

where Q_i is the i-th object proposal; n is the total number of object proposals, and $\sqrt{\mathrm{size}(\cdot)}$ is the spatial size of Q_i; Sum denotes the summation of the values of a gray map; $\sum$ represents the summation operation of the scores of all the proposals; Θ mean the graying operation. Specifically, λ serves as a weight ranging from 0-1 to ensure the best performance of other methods.

We present the quantitative comparison results for the SALICON dataset regarding the SRCC and F1 metrics in Fig.9. We assign the saliency ranking orders of the compared methods using our newly proposed binary threshold T. Specifically, we set the GT threshold γ (eq. 4) as 0.1, 0.2, 0.5, and 0.8 (we will detail the rationale of this choice in Fig. 11-B later). Our classification-based method, denoted as Ours (Cls.), outperforms all other methods regarding the SRCC metric on any GT/binary threshold, even when the best results of the compared methods are selected by the dynamic binary thresholds searching. Importantly, the SRCC metric of our

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

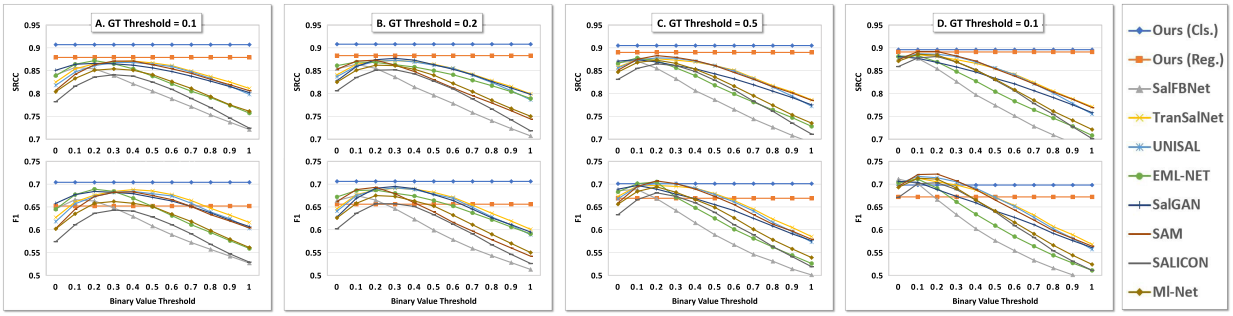

Fig. 9. Performance illustration of the proposed method with other SOTA models in terms of SRCC (upper line) and F1 (lower line) curves with the binary value threshold of saliency maps at four GT thresholds (γ = 0.1, 0.2, 0.5 and 0.8 from left to right).

TABLE I

QUANTITATIVE COMPARISONS. FOR A FAIR COMPARISON, ALL COMPARED THREE REPRESENTATIVE SOTA METHODS HAVE BEEN RETAINED RESPECTIVELY USING THE DEFAULT TRAINING SPLIT OF THE THREE TESTED SETS

Sets	Siris' Dataset		Liu's Dataset		Our Dataset	
Methods	SRCC	F1	SRCC	F1	SRCC	F1
ASSR	.792	.601	.714	.523	.885	.683
ILRSR	.811	.618	.806	.609	.891	.698
Ours	**.897**	**.695**	**.882**	**.701**	**.908**	**.706**

method (0.907, 0.908, 0.905, and 0.896) remains consistent across all different binary value thresholds, indicating the robustness of our proposed method. Furthermore, our method performs competitively with other methods in terms of the F1. More qualitative visual comparisons with the competing methods are presented in Fig. 10.

In addition to our classification-based method, we also evaluated a regression-based version of our approach (referred to as "Ours (Reg.)"). However, the results demonstrate that the regression-based method is inferior to the classification-based method in all evaluation metrics. Our proposed classification-based method achieves average performance gains of 1.8% and 4% across four GT thresholds, as measured by the SRCC and F1 metrics. These results confirm the effectiveness of our classification-based approach.

2) Comparison With Saliency Ranking Methods: We compare our proposed model with two other representative SOTA saliency ranking methods, *i.e.*, ASSR [6] and ILRSR [7]. For more comprehensive evaluations, we compare three models on three datasets, *i.e.*, Siris' dataset [6], Liu' dataset [7], and our proposed dataset. Following [7], we only select at most five instances in our prediction results for fair comparisons due to the limited salient instances in the GT saliency map of the ASSR' dataset. Other training settings are the same as [7]. Specifically, our model (we choose the best model with $\gamma = 0.2$) and the compared two models are sorted differently, *i.e.*, our model ranks the salient object orders in the proposal level (use rectangles to frame objects), while the two in pixel-wise level (segment the instances). Thus, we only compare the numerical ranking orders with SRCC and F1.

As shown in Table I, we can observe that our model generally outperforms the other two methods by a large margin on all three datasets, *e.g.*, compared with ILRSR, our model achieves an average performance gain of 8.6%, 7.6%, and 1.7% in the SRCC metric across the three datasets. These results show the effectiveness of our proposed saliency ranking model and its superiority and efficiency for practical usage.

TABLE II

COMPONENTS EVALUATION REGARDING THE MAJOR PARTS OF FEATURE EMBEDDING & ENHANCING (FEE, SEC. III-C.2)): A-C, PE; AND HA (EQ. 10)) OF PROPOSED METHODS ON OUR RELABELLED SALICON SET

	Major Components: FEE			Transformer		Cls.	Dataset: SALICON	
	PSE	LF	GF	PE	TEB	HA	SRCC	F1
⓪ 1	✓	✗	✗	✗	✗	✗	.778	.531
2	✗	✓	✗	✗	✗	✗	.746	.519
① 3	✗	✓	✓	✗	✗	✗	.784	.564
4	✓	✓	✓	✗	✗	✗	.865	.624
5	✓	✓	✓	✓	✗	✗	.871	.629
② ③ 6	✓	✓	✓	✓	✓	✗	**.912**	.691
④ 7	✓	✓	✓	✓	✓	✓	.908	**.706**

⓪ Baseline ① *Verify* FEE ② *Verify* Transformer ③ ④ *Verify* HA

LF/GF: Local/Global Feature-level Proposals **PSE:** Plain Spatial Encoder
PE: Position Embedding **Cls.:** Classification
TEB: Transformer Encoder Block **HA:** Hungarian Algorithm

D. Component Evaluation

To validate the effectiveness of our method, we conducted an extensive component evaluation on our relabelled SALICON set. The results are shown in Table II. To enable successful code running, we replaced the key components that needed to be verified with simpler operations. For example, we replaced the proposed SPAFE with simple ResNet101 and the Transformer with two fully-connected layers. We treated this replaced model as a baseline, and the qualitative result is shown in the 1st column denoted by mark ⓪.

Comparing line 1 and line 2, we can see that the features extracted by ResNet101 are more effective than the local feature-level features of the proposals obtained by ROIAlign, as evidenced by the higher SRCC and F1 metrics (0.778 *v.s.* 0.746 and 0.531 *v.s.* 0.519, respectively). This is likely because ResNet101's deep architecture makes its features more informative than the limited feature extraction capability of UNISAL, a lightweight network.

Lines 2-4 show the effectiveness of SPAFE in combining local and global feature-level proposals with ResNet101. Removing ResNet101 (line 3) or both ResNet101 and global feature-level proposals (line 2) decreased the SRCC metric in the SALICON set (mark ①). Global feature-level proposals enhance perception ability by capturing more surrounding features. ResNet101 provides more feature representations for Transformer blocks. Position embedding and Transformer Encoder Blocks are effective in capturing long-range dependency features (mark ②). We also conducted an ablation study on other backbones, such as FC and GCN (Table III-C-a).

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

Fig. 10. Visulizations of our proposed method and several competing methods. Zoom in for better observation.

The Hungarian algorithm was found to be necessary in addressing the issue of identical classification results, as shown in marks ③ and ④. Replacing the Hungarian algorithm with the max function resulted in decreased ranking accuracy (SRCC) despite improved F1 metric performance. This is because false-alarm ranking orders caused by the identical classification result issue may reduce ranking accuracy, which does not affect the SRCC metric.

E. Our RA-SRGT v.s. SOTA Competitors

1) Effectiveness and Rationality of RA-SRGT: Saliency ranking is a subjective task. We aimed to eliminate induction bias by comparing four different saliency ranking GT orders generation methods to explore which one is closer to the real relative saliency of the human visual system. These methods included "our proposed method", "average value", "maximum value", and "fixation points". Thus, we conducted a user study to validate our RA-SRGT (Sec. III-B). We selected 300 images, each with four distinct saliency ranking GT orders generated by these four methods, and asked subjects to annotate the saliency ranking orders. We then calculated the number of images with the same ranking orders between annotated saliency ranking orders and saliency ranking orders generated by each of the four saliency ranking GT orders generation methods. As shown in Fig. 11-A, the number of images with the same saliency ranking orders between our method (the lightest bar) and the annotated real relative saliency of the human visual system steadily outnumbered the other methods as the number of subjects increased. This result demonstrates that our method is more feasible and closer to the human visual system and effectively generates accurate saliency ranking orders.

2) Physiological Exploration of RA-SRGT: To conduct an in-depth analysis of saliency ranking GT orders from a physiological perspective, we performed a thorough investigation to determine the saliency ranking GT orders that best match the human visual system and are preferred by different individuals, using various thresholds. Specifically, we randomly selected 3,000 images from the SALICON dataset and generated saliency ranking GT orders by applying GT thresholds (eq. 4) ranging from 0.1 to 1.0. We then computed the total number of saliency ranking discrepancy offsets $\mathrm{T}^{t}_{\mathrm{offset}}$ between adjacent thresholds $t-0.1$ and t. This involved summing the saliency ranking order changes of objects in all images under adjacent thresholds. It can be mathematically expressed as:

$$\mathrm{T}^{t}_{\mathrm{offset}} = \sum_{n=1}^{N}\sum_{i=1}^{I}\left|\mathrm{Rank}^{t}_{n,i} - \mathrm{Rank}^{t-0.1}_{n,i}\right|, \tag{13}$$

where $0.1 \leq t \leq 1$ with a step size of 0.1; the index i ($1 \leq i \leq I$) and I represents the i-th and total proposal(s) within the n-th image ($1 \leq n \leq N$, $N = 3,000$), respectively; $\sum$ denotes summation operation; Rank ($1 \leq \mathrm{Rank} \leq I$) is the saliency ranking order.

As shown in Fig. 11-B, we observed that the total number of saliency ranking discrepancy offsets among thresholds from 0.3 to 0.6 and 0.7 to 1.0 are essentially the same. Therefore, we chose the transitional end-point values as our final GT thresholds γ (eq. 4), namely, 0.1, 0.2, 0.5, and 0.8 (0.5 and 0.8 are the average thresholds of 0.3 to 0.6 and 0.7 to 1.0, respectively). These four GT thresholds could generate four different kinds of saliency ranking GT orders to the maximum extent. However, we did not know which one was preferred by the HVS and different individuals with diverse life backgrounds, *e.g.*, gender. Therefore, we conducted another user study with over 100 subjects (aged 19 to 27) with diverse backgrounds to explore which of the four different kinds of saliency ranking GT orders of randomly selected 100 images under the four GT thresholds they preferred the most. We have also presented a quantitative comparison in Table III-C-b.

Results in Fig. 11-D showed that both males and females preferred the saliency ranking GT orders under GT threshold 0.2, which conforms to the quantitative comparison in Fig. 9, where the SRCC and F1 metrics also obtained the best results under GT threshold 0.2. This finding verifies the effectiveness of our approach in generating saliency ranking GT orders, as our saliency ranking GT orders are closer to the real HVS. At the same time, we further explored the influence of gender in saliency ranking, as shown in Fig. 11-E, where males were more inclined to saliency ranking GT orders under GT threshold 0.1, while females tended to choose GT threshold 0.5 and 0.8. Nevertheless, GT threshold 0.2, which both males and females preferred, was equal, demonstrating our approach's robustness.

3) Impact of the Hyperparameter β: To verify the impact of the hyperparameter β in eq. 4 on the generation of saliency ranking GT orders, we conduct an ablation analysis using

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

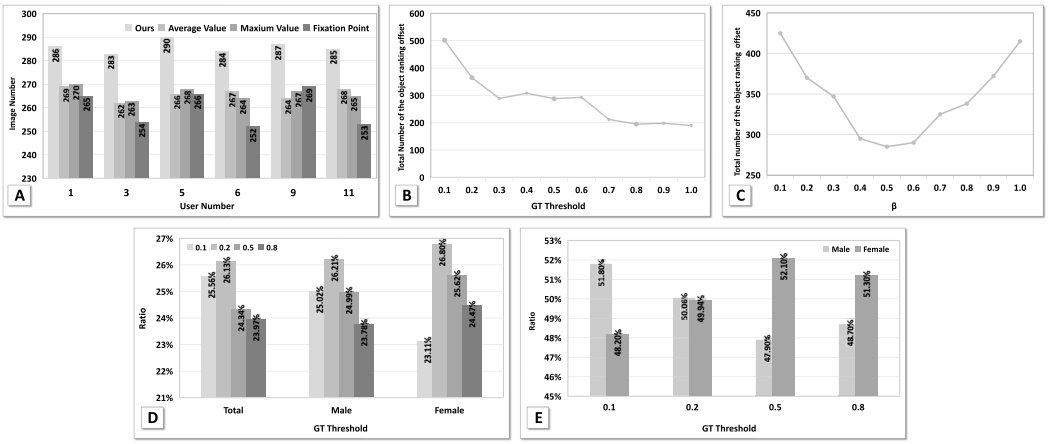

Fig. 11. A: a user study on the consistency between real relative saliency ranking of the HSV and our saliency ranking GT orders, and other saliency ranking GT orders, respectively. B: a user study on discrepancy rate. That is to compute the total number of saliency ranking discrepancies T^t_{offset} between adjacent thresholds $t-0.1$ and t. This involved summing the ranking changes of objects in all images under adjacent thresholds. C: ablation study on different choices of β. D: a user study on discrepancy rate. That is to compute the total number of saliency ranking discrepancies T^t_{offset} between adjacent thresholds $t-0.1$ and t. This involved summing the ranking changes of objects in all images under adjacent thresholds. E: ablation study on different choices of β.

TABLE III

COMPARISONS BETWEEN MODELS TRAINED BY "REAL GT" AND "SHUFFLED GT" (A), BETWEEN MULTI-TASK LEARNING AND SINGLE-TASK LEARNING (B), AND BETWEEN DIFFERENT CHOICES OF BACKBONES AND GT THRESHOLD γ (C) ON THE SALICON SET

A

Metrics	a. Real GT SRCC	a. Real GT F1	b. Shuffled GT SRCC	b. Shuffled GT F1
ASSR	.885	.683	.882	.681
ILRSR-GR	**.891**	**.698**	**.888**	**.695**

B

Metrics	a. PPA SRCC	a. PPA F1	b. Ours SRCC	b. Ours F1
Detection + Ranking	.841	.652	.882	.691
Detection→Ranking	**.867**	**.695**	**.908**	**.706**

C

Metrics	a. Different Backbones FC	GCN	VIT	b. Different Threshold γ 0.1	0.2	0.5	0.8
SRCC	.891	.904	**.908**	.907	**.908**	.905	.896
F1	.689	.701	**.706**	.704	**.706**	.701	.698

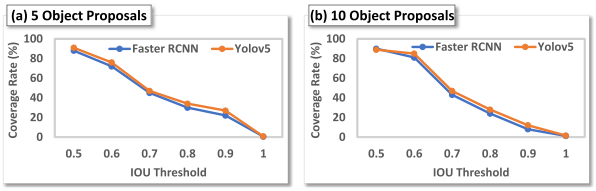

Fig. 12. Charts of the changing trend of the proposal coverage rates when replacing our object detector with four other object detectors, *i.e.*, Faster RCNN [49], Yolov5 [50], DETR [51] and Sparse RCNN [52].

the saliency ranking orders obtained from the user study, which includes 300 manually annotated images. To control variables, we fix the threshold γ at 0.2. Next, we calculate the number of image offsets between the saliency ranking orders generated by our RA-SRGT (Sec. III-B) and the saliency ranking orders manually annotated by humans under different β values. As shown in Fig. 11-C, the image offsets vary greatly with an increase in β, but remain almost constant in the range of 0.4∼0.6. Results showed that when β was too small, saliency ranking focused only on local fixations. When β was in the range of 0.4∼0.6, the ranking was more stable, and a large β could ignore local objects. We set β to 0.5 to balance the contribution of fixations and object spatial size.

F. Our Single-Task Classification-Based Method v.s. SOTA

1) Irrationality of Instance Segmentation-Based Saliency Ranking Methods: To verify whether combining saliency ranking and instance segmentation tasks for multi-task learning [7] may not necessarily improve model performance, we shuffled the ground truth segmentation mask ranking order *i.e.*, shuffling the segmentation of each object in the segmentation mask by their brightness orders, and used it to retrain the existing saliency ranking model to test whether the model relies on the semantic information in the masks or not. We expected this would eliminate the side effect of segmentation and only leave the multi-task effect. However, as shown in Table III-A, we found that the performance of original models trained by "Shuffled GT" stays the same as the original models ("Real GT"). This suggests that the segmentation masks with consistent brightness orders and saliency ranking orders of each object do not significantly influence the saliency ranking task, which means that the segmentation task only provides segmented instance results for the saliency ranking task and has not much in the way of performance improvement. Thus, it might not need instance segmentation as an auxiliary task for saliency ranking.

2) Effectiveness of Single-Task Method With Offline Proposal Generation: To demonstrate the drawbacks of the existing multi-task model [11] that uses online object proposal generation and training of the detector for saliency ranking task, we design a single-task model that uses offline object proposal extraction and no training of the detector. The hypothesis is that the online extraction and training of the detector might introduce noise or inconsistency in the object features and masks, which may affect the ranking performance. The single-task model avoids this problem by using a fixed and reliable object detector pre-trained on a large-scale dataset. We compare the two models *i.e.*, 1) original model [11] with multi-task learning of two parallel object detection branch and saliency ranking branch, and 2) modified model [11] with single-task learning by offline generated object proposals and an online saliency ranking on SALICON dataset. Experimental results in Table III-B show that the single-task model will

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

COMPARISONS OF COMPUTATIONAL EFFICIENCY AND RESOURCE REQUIREMENTS BETWEEN OUR PROPOSED AND OTHER COMPETING METHODS. FPS IS CONSIDERED TO EVALUATE THE MODEL EFFICIENCY. **BOLD** INDICATES THE BEST PERFORMANCE

A. Running time comparisons among different saliency ranking methods.

a: Segmentation-based Methods | b: Object Detection-based Method

Metrics	ASSR	ILRSR	PPA	RSDNet	Bi-OCPL	Ours
SRCC	.714	.806	.818	.705	.834	**.882**
F1	.523	.609	.642	.518	.684	**.701**
FPS	2.2	15.4	17.1	4.6	13.8	**20.5**
Platform	1080Ti	-	TITAN RTX	-	Tesla GPU	2080Ti

B. Running time comparisons among different object detection methods.

Models	Sparse RCNN	Faster RCNN	DETR	Efficient Det	YOLOV5	RT-DETR	YOLOV8
FPS	23	26	28	41	140	114	311

TABLE V

COMPARISONS AMONG MULTI-TASK-LEARNING SEGMENTATION/ DETECTION-BASED SALIENCY RANKING MODELS (A AND B) AND OUR SINGLE-TASK-LEARNING MODEL (C). THE METRIC "MIOU" IS USED TO EVALUATE THE PERFORMANCE OF SEGMENTATION MODELS, WHILE THE METRICS "AP/AP_{50}" ARE EMPLOYED FOR ASSESSING THE PERFORMANCE OF DETECTION MODELS

	a. RSDNet	b. PPA		c. Ours	
Metrics	mIoU	AP	AP_{50}	AP	AP_{50}
Segmentation+ Ranking	37.9	-	-	-	-
w/o Ranking	**38.4**	-	-	-	-
Detection + Ranking	-	37.4	41.1	**51.6**	**70.3**
w/o Ranking	-	**39.3**	**43.6**	49.8	68.2

achieve comparable or better performance than the multi-task model, while being more efficient and stable.

G. In-Depth Analysis of Multi-Task Learning-Based Saliency Ranking Methods

1) High Complexity: In Fig. 2-B (a), we discussed the high complexity of existing multi-task learning methods that use segmentation as a basis. These methods are more intricate than plain object detection-based approaches. As saliency ranking is designed to determine the visual significance of objects/regions in an image, without needing pixel-level segmentation or salient detection of the entire image, it simplifies the saliency ranking tasks and improves efficiency. To evaluate the computational efficiency of our proposed method, we conducted a comparison of the running time between our method and several segmentation-based state-of-the-art (SOTA) methods, including ASSR [6], ILRSR [7], PPA [11], RSDNet [10], and Bi-OCPL [9]. As shown in Table IV-A, our method achieved real-time speed, running at 20.5 frames per second (FPS) during the **total inference phase**, including the offline generation of proposals.

Furthermore, the computational cost of offline object proposal generation depends on the selected object detection models. We present several representative SOTA object detection models in Table IV-B. Although the FPS of EfficientDet - the model chosen in our method - lies in the middle, it still outperforms other models (Table IV-A).

2) Uncertain Mutual Promotion: In multi-task learning, "mutual promotion" is a strategy or approach aimed at improving the performance of multiple related tasks by promoting mutual interaction and sharing of information between the tasks. This implies that shared features and knowledge between tasks can mutually promote and benefit all tasks. However, in practice, "mutual promotion" may encounter some issues because the relationships between different tasks are not always clear or certain. This gives rise to the concept of "uncertain mutual promotion", which emphasizes the uncertainty in the relationships between tasks, where certain tasks' features may have positive impacts on other tasks. In contrast, others may have no effect or even negative effects.

To demonstrate the phenomenon of "uncertain mutual promotion" in existing saliency ranking models, we conducted two validations using a multi-task learning strategy. Specifically, 1) we tested the performance of only one of the tasks in multi-task learning, *e.g.*, the performance of segmentation in segmentation-based saliency ranking models or the performance of object detection in object detection-based saliency ranking models;[6] 2) we trained and tested only one task by excluding the saliency ranking task from multi-task learning, *e.g.*, the segmentation task or object detection task. selected RSDNet [10] as the representative segmentation-based model, and PPA [11] as the object detection-based model due to its easy code availability.

Suppose segmentation-based multi-task-learning model **a** (RSDNet) can achieve mutual promotion in multi-task learning. In that case, the expected outcome should be that segmentation performance during simultaneous segmentation and ranking is higher than when only segmentation is performed. However, according to the results in Table V, the data decreased from 38.4 to 37.9 regarding the mIOU metric, which provides sufficient evidence that segmentation-based multi-task-learning model **a** (RSDNet) is an example of uncertain mutual promotion. As is the same case in object detection-based multi-task-learning model **b** (PPA), the AP and AP_{50} separately decreased, which also proves that object detection-based multi-task-learning model **b** (PPA) is also an example of uncertain mutual promotion. However, in our approach (model **c** (Ours)), a single-task learning-based saliency ranking model, the ranking task offers support to its pre-object detection processing, boosting the AP and AP_{50} separately from 49.8 to 51.6 and 68.2 to 70.3. It proves that our single-task learning method can achieve mutual promotion.

The "uncertain mutual promotion" in segmentation-based multi-task learning models can be attributed to two factors. Firstly, there is task misalignment between the segmentation task and the ranking task. These tasks have different objectives and optimization criteria, which hinders their mutual promotion. Secondly, training dynamics play a role, as the optimization process may prioritize one task over the other, leading to decreased performance in the segmentation task. In contrast, our approach avoids "uncertain mutual promotion" for several reasons. We use a single-task learning approach, focusing solely on saliency ranking without conflicting tasks. Additionally, we employ pre-object detection processing, ensuring accurate object delineation and providing valuable information for ranking. This targeted optimization

[6]Note that nearly all the existing saliency ranking models are segmentation-based, only a few models resort to object detection before segmenting the objects, we consider these kinds of methods containing object detection methods as object detection-based saliency ranking models.

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

TABLE VI

THE FINAL RANKING RESULTS COMPARISONS AMONG FIVE OBJECT DETECTION METHODS. THE FINAL RANKING RESULTS EXHIBIT ONLY A SMALL RANGE OF CHANGE

Metrics	Faster RCNN	Yolov5	DETR	Sparse RCNN	EfficientDet
SRCC	.906	.903	.902	.907	**.908**
F1	.703	.705	**.707**	.703	.706

and alignment of objectives enhance the ranking performance without compromising segmentation accuracy.

Based on this observation, we conclude that existing multi-task learning-based saliency ranking models can exhibit "uncertain mutual promotion" and our single-task learning-based saliency ranking method is "mutual promotion".

3) Unstable Training: We conducted experiments by replacing our original object detection methods with EfficientDet, Faster RCNN [49], Yolov5 [50], DETR [51], and Sparse RCNN [52]. The quantitative results presented in Table VI demonstrate that the final ranking results only exhibit a small range of change. This validates that our approach possesses a strong capability for "stable training".

We argue that multi-task learning models exhibit "unstable training" due to two primary reasons: 1) In multi-task learning, the saliency ranking tasks require retraining the entire model whenever the segmentation model changes. In contrast, our model only needs to be trained once using high-accuracy, efficient object detection methods. The similarity of generated proposals from different detection methods allows us to directly feed offline-generated object proposals into the trained saliency ranking model without training a new model for each detection method. 2) Joint training of segmentation-based multi-task learning models can prioritize one task over the other, leading to an out-of-sync optimization process and "unstable training". Our single-task learning method focuses solely on the saliency ranking loss, simplifying implementation and avoiding resource allocation conflicts.

H. Method Generality Analysis

We evaluated our detection-based saliency ranking method's compatibility with segmentation tasks, specifically segmentation-based saliency ranking. We used large-scale segmentation models such as SAM [47] and SegGPT [53] to generate rectangular segmentation proposals, which were fed into our proposed saliency ranking model. These proposals differ from object proposals in that they only contain foreground objects (with the background filled in black) while also being rectangular in shape. We initially obtained object proposals using object detection models, which we then fed into SAM or SegGPT for segmentation (refer to Fig. 13).

The results in Table VII show that our object proposal-based method (referred to as Ours (Detection)) performs slightly lower compared to the segmentation proposal-based methods, namely Ours (SAM) and Ours (SegGPT). Furthermore, these segmentation proposal-based methods outperform other segmentation-based saliency ranking models listed in Table IV. These segmentation proposal-based methods demonstrate higher performance because they only focus on ranking the objects, while our object proposals include unnecessary background information, which may impact the model's learning process. However, it is worth noting that the segmentation proposal-based methods have longer running times due to the complexity of pixel calculations involved in segmentation tasks, whereas the object proposal-based method deals with sparse tasks. This discrepancy arises from the nature of these two approaches.

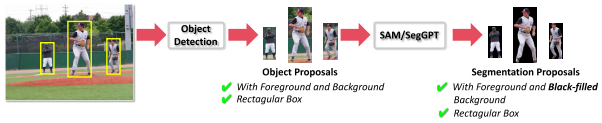

Fig. 13. Comparisons between object proposals and segmentation proposals.

TABLE VII

GENERALITY ANALYSIS OF OUR DETECTION-BASED SALIENCY RANKING METHOD TO SEGMENTATION-BASED TASKS

Metrics	SAM	Ours (SegGPT)	Ours (Detection)
SRCC	**.886**	.885	.882
F1	.704	**.706**	.701
FPS	19.2	18.7	**20.5**

TABLE VIII

COMPARISONS BETWEEN TRANSFORMING SEGMENTATION MASKS TO OBJECT PROPOSALS AND OUR DIRECTLY GENERATED OBJECT PROPOSALS

Metrics	AP	AP_{50}	SRCC	F1	FPS
SAM (mask → proposal)	59.2	76.8	.883	.703	16.8
SegGPT (mask → proposal)	**59.8**	**77.1**	**.885**	**.704**	17.3
EfficientDet (proposal)	51.6	70.3	.882	.701	**20.5**

To further explore the effectiveness of segmentation models in saliency ranking, we used large-scale models like SAM and SegGPT to generate masks. These masks were then used to fill proposals based on object sizes. Results in Table VIII show that using SAM and SegGPT for mask-to-proposal transformation improves saliency ranking performance. However, their running speed (16.8 and 17.3 FPS) is slower compared to our object detection-based model (20.5 FPS). This is because SAM and SegGPT perform pixel-wise segmentation across the entire image, resulting in more accurate object localization. Despite the slight improvement in ranking performance, these methods reduce processing speed. Considering the goal of saliency ranking is to assign a saliency order to each object, object detection-based methods strike a balance between effectiveness and efficiency. These findings validate the strong stability and generality of our proposed method.

I. Ablation Study

1) Robustness of Our Object-Wise Proposal Detection: We evaluated the robustness of our proposal detection method by comparing the coverage rates of proposals detected by different object detectors. We used EfficientDet as the baseline and compared it with Faster RCNN [49], Yolov5 [50], DETR [51] and Sparse RCNN [52]. Varying the IOU threshold from 0.5 to 1, we observed consistent coverage rates for proposals detected by our used detector (*e.g.*, EfficientDet) and other detectors (*e.g.*, Faster RCNN) (see Fig. 12). This indicates the robustness of our approach across scenarios with 5 (a) and 10 (b) proposals. Our training approach ensured stable performance across various detectors, demonstrating strong generalizability and applicability of our proposed method.

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

2) Transformer v.s. Other Backbones: We conducted an ablation study on the Vision Transformer (VIT), Graph Convolutional Network (GCN), and fully-connected layer (FC) for saliency ranking. The object proposals determined GCN's number of graph nodes, while FC directly received the salient position-aware feature F_p (Sec. III-C.2) as input for predicting the final results. The results in Table III-C-a indicate that VIT outperformed GCN and FC in modeling object relationships, scoring 0.908 compared to 0.904 for GCN and 0.891 for FC. VIT's self-attention mechanisms enabled it to capture long-range dependencies and understand spatial relationships between objects comprehensively. GCN faced challenges in achieving this efficiently through graph-based techniques. FC directly input the salient feature without considering spatial relationships, resulting in information loss and inferior performance compared to VIT and GCN.

3) Impact of GT Threshold γ: To further verify the advantages of our proposed relationship-aware saliency ranking GT orders generation method, we conducted a quantitative test to explore the impact of GT threshold γ (eq. 4), where we chose $\gamma = 0.1$, 0.2, 0.5, and 0.8. Results in Table III-C-b show that as the γ value increases, the SRCC and F1 metrics first rise and then fall simultaneously. For instance, when the γ value changes from 0.1 to 1.0, the SRCC and F1 metrics decline by 1.1% and 0.6%. However, when the γ value is 0.2, the SRCC and F1 metrics reach their maximum. This is because when γ is small, the spatial size of the object has less impact on the final saliency ranking GTs, whereas when γ is large, the object size has too much influence, resulting in false alarms. These results confirm the robustness of our proposed method compared to other saliency ranking GT order generation methods.

J. Limitations

1) In Fig. 14, we demonstrate the limitations of our model through visual examples. These cases highlight challenges in distinguishing visually similar objects with comparable saliency ranking orders (line 1). To address this, we can incorporate color saliency and leverage existing fixation prediction models to locally rank visual stimuli based on color and shape. Another challenge arises when our model fails to detect all object proposals, making it difficult to assign saliency ranking orders (line 2). To mitigate this, we can utilize a more accurate object detection model or employ multiple detection models in coordination. Additionally, treating the problem of missing detections as a camouflaged object detection task [54], [55] could be a straightforward approach. Camouflaged object detection methods are specifically designed to handle scenarios where objects blend with the background. While this problem differs from saliency detection [56], we will not explore it further in this discussion.

2) Our RA-SRGT method (Sec. III-B) has several limitations that require attention. Firstly, the approach relies heavily on object proposals generated by detection models, which may introduce errors in object localization, impacting saliency ranking accuracy. It is crucial to find ways to mitigate these errors. Secondly, the method's thresholds, γ and β, are manually set, introducing subjectivity. This subjectivity poses challenges in generalizability across datasets or scenarios where optimal thresholds may vary.

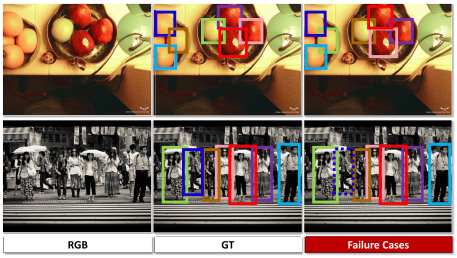

Fig. 14. Failure cases with similar appearances and shapes (1st line) and object detector failing to detect all object proposals (2nd line).

V. Conclusion

In summary, this paper introduces a novel paradigm for saliency ranking that outperforms existing methods on the SALICON dataset. The approach addresses challenges in generating ground truth orders, network design, and training protocols. A key contribution is recognizing the limitations of existing saliency detection methods in determining the relative importance of multiple objects and their relationships. The proposed whole-flow processing paradigm demonstrates superior performance, accurately ranking salient objects and enhancing downstream tasks. These findings pave the way for improved saliency ranking methods and offer valuable insights into the importance ordering of visually salient objects. Future work can focus on improving data efficiency through semi-supervised learning or transfer learning, and enhancing generalization capabilities for diverse datasets and real-world deployment.

References

[1] M. Song, W. Song, G. Yang, and C. Chen, "Improving RGB-D salient object detection via modality-aware decoder," *IEEE Trans. Image Process.*, vol. 31, pp. 6124–6138, 2022.

[2] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-stage learning to predict human eye fixations via SDAEs," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 487–498, Feb. 2016.

[3] G. Lu, T. Zhong, J. Geng, Q. Hu, and D. Xu, "Learning based multi-modality image and video compression," in *Proc. CVPR*, Jun. 2022, pp. 6073–6082.

[4] H. Wu, M. Wang, W. Zhou, H. Li, and Q. Tian, "Contextual similarity distillation for asymmetric image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9479–9488.

[5] M. A. Islam, M. Kalash, and N. D. B. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7142–7150.

[6] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Inferring attention shift ranks of objects for image saliency," in *Proc. CVPR*, Jul. 2020, pp. 12133–12143.

[7] N. Liu, L. Li, W. Zhao, J. Han, and L. Shao, "Instance-level relative saliency ranking with graph reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8321–8337, Nov. 2022.

[8] Y. Lv et al., "Simultaneously localize, segment and rank the camouflaged objects," in *Proc. CVPR*, 2021, pp. 11591–11601.

[9] X. Tian, K. Xu, X. Yang, L. Du, B. Yin, and R. W. H. Lau, "Bi-directional object-context prioritization learning for saliency ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5872–5881.

[10] M. Kalash, M. A. Islam, and N. D. B. Bruce, "Relative saliency and ranking: Models, metrics, data and benchmarks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 204–219, Jan. 2021.

[11] H. Fang et al., "Salient object ranking with position-preserved attention," in *Proc. ICCV*, 2021, pp. 16311–16321.

[12] W. Sun, Z. Chen, and F. Wu, "Visual scanpath prediction using IOR-ROI recurrent mixture density network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2101–2118, Jun. 2021.

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.

[13] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[14] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*, 2012, pp. 29–42.

[15] G. Li et al., "Personal fixations-based object segmentation with object localization and boundary preservation," *IEEE Trans. Image Process.*, vol. 30, pp. 1461–1475, 2021.

[16] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Scene context-aware salient object detection," in *Proc. ICCV*, 2021, pp. 4136–4146.

[17] N. Liu, K. Nan, W. Zhao, X. Yao, and J. Han, "Learning complementary spatial–temporal transformer for video salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–17, Feb. 2023.

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

[19] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. CVPR*, 2019, pp. 8546–8556.

[20] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, Jun. 2022.

[21] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proc. CVPR*, 2019, pp. 5961–5970.

[22] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. CVPR*, 2019, pp. 1448–1457.

[23] Y. Pang, C. Wu, H. Wu, and X. Yu, "Unsupervised multi-subclass saliency classification for salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 2189–2202, 2023, doi: 10.1109/TMM.2022.3144070.

[24] J. Li, Z. Wang, Z. Pan, Q. Liu, and D. Guo, "Looking at boundary: Siamese densely cooperative fusion for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3580–3593, Jul. 2023, doi: 10.1109/TNNLS.2021.3113657.

[25] J. Lin, H. Guan, and R. W. H. Lau, "Rethinking video salient object ranking," 2022, *arXiv:2203.17257*.

[26] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3738–3752, Mar. 2023.

[27] N. Zhang, J. Han, and N. Liu, "Learning implicit class knowledge for RGB-D co-salient object detection with transformers," *IEEE Trans. Image Process.*, vol. 31, pp. 4556–4570, 2022.

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

[29] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "RGBD salient object detection via disentangled cross-modal fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 8407–8416, 2020.

[30] J. Zhang et al., "Uncertainty inspired RGB-D saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5761–5779, Sep. 2022.

[31] J. Zhang et al., "RGB-D saliency detection via cascaded mutual information minimization," in *Proc. ICCV*, 2021, pp. 4318–4327.

[32] C. Xu, Q. Li, M. Zhou, Q. Zhou, Y. Zhou, and Y. Ma, "RGB-T salient object detection via CNN feature and result saliency map fusion," *Int. J. Speech Technol.*, vol. 52, no. 10, pp. 11343–11362, Aug. 2022.

[33] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection," in *Proc. IEEE MIPR*, May 2019, pp. 141–146.

[34] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Feb. 1998.

[35] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proc. NIPS*, 2007, pp. 241–248.

[36] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proc. CVPR*, 2015, pp. 1072–1080.

[37] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[38] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1913–1927, Aug. 2020.

[39] J. Pan et al., "SalGAN: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*.

[40] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 419–435.

[41] Y. Liu, D. Zhang, N. Liu, S. Xu, and J. Han, "Disentangled capsule routing for fast part-object relational saliency," *IEEE Trans. Image Process.*, vol. 31, pp. 6719–6732, 2022.

[42] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[43] D. Alexey, B. Lucas, K. Alexander, W. Dirk, and Z. Xiaohua, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[44] G. Ding, A. Caglayan, M. Murakawa, and R. Nakamura, "SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks," *Image Vis. Comput.*, vol. 120, Apr. 2022, Art. no. 104395.

[45] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "TranSalNet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, pp. 455–467, Jul. 2022.

[46] S. Jia and N. D. B. Bruce, "EML-NET: An expandable multi-layer network for saliency prediction," *Image Vis. Comput.*, vol. 95, Mar. 2020, Art. no. 103887.

[47] A. Kirillov et al., "Segment anything," in *Proc. ICCV*, 2023, pp. 4015–4026.

[48] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3488–3493.

[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[50] (2017). *Ultralytics: YOLOV5*. [Online]. Available: https://github.com/ultralytics/yolov5

[51] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020, pp. 213–229.

[52] P. Sun et al., "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14449–14458.

[53] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "SegGPT: Towards segmenting everything in context," in *Proc. ICCV*, 2023, pp. 1130–1140.

[54] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, Oct. 2022.

[55] D.-P. Fan, G.-P. Ji, P. Xu, M.-M. Cheng, C. Sakaridis, and L. Van Gool, "Advances in deep concealed scene understanding," *Vis. Intell.*, vol. 1, no. 1, pp. 1–10, Aug. 2023.

[56] D.-P. Fan et al., "Re-thinking co-salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4339–4354, Aug. 2022.

Authorized licensed use limited to: China University of Petroleum. Downloaded on October 30,2024 at 06:31:44 UTC from IEEE Xplore. Restrictions apply.