Neurocomputing 649 (2025) 130715

Contents lists available at ScienceDirect

Neurocomputing

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

DICO: Distance-weighted Contrast and Instance Correlation for salient object ranking

Jinxia Zhang [a,b,*], Min Huang [a], Xinchao Zhu [a], Haikun Wei [a], Shixiong Fang [a], Kanjian Zhang [a]

[a] the Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China
[b] Advanced Ocean Institute of Southeast University, Nantong 226010, China

ARTICLE INFO

Communicated by R. Cong

Keywords:
Salient object ranking
Distance-weighted contrast
Instance correlation
Attention shift rank

ABSTRACT

Salient object ranking aims to infer the saliency levels of objects within an image and rank them accordingly. Existing methods mainly rely on attention mechanisms to model object-context relations. However, they tend to overlook two crucial aspects: the way spatial distance modulates contrast and the correlation between instances when assigning relative saliency. To address these issues, we propose a novel salient object ranking approach by modeling Distance-weighted COntrast and Instance COrrelation (DICO). Specifically, we propose the Distance-weighted Contrast (DCO) module, which utilizes Gaussian functions to simulate a dynamic attention distribution among regions. This distribution assigns higher weights to neighboring regions, capturing the inherent spatial relations in the scene. By integrating the dynamically generated attention distribution with feature-based contrast, the DCO module effectively enables more precise modeling of spatially-aware object-context and inter-object contrast. Furthermore, we propose an Instance Correlation (ICO) loss that takes into account both the inter-object correlation and individual object fitness. This dual consideration enables the model to more effectively learn the relative saliency of different objects. Specifically, the inter-object correlation helps to narrow the saliency gap between instance pairs that have close ranks, while individual object fitness aims to enhance the saliency scores of highly salient objects and reduce the saliency scores of less salient ones. Extensive experiments demonstrate that our method outperforms existing state-of-the-art approaches in terms of balancing computational complexity and performance.

1. Introduction

Salient object detection (SOD) is a crucial task in computer vision, aiming to locate objects or regions containing important information in visual scenes. Since its initial introduction by Liu et al. [1], various models have been proposed for this task. Many methods have described SOD as a binary prediction problem [2–7] without considering the varying degrees of saliency among different objects. However, when humans observe images, the visual system shifts attention from one object to another, resulting in an uneven distribution of visual attention among objects. In practical applications, saliency ranking of different objects can help prioritize critical visual elements in areas such as image compression [8], image captioning [9] and autonomous driving [10]. This prioritization enhances application performance, decision speed, and overall operational efficiency.

As a pioneering work, Islam et al. [11] introduced the concept of salient object ranking, but their work only focused on pixel-level relative saliency. Siris et al. [12] proposed an approach that leverages both bottom-up and top-down attention mechanisms to predict the saliency rank. Fang et al. [13] presented an end-to-end solution that combines positional information and attention mechanisms to improve the accuracy of ranking. Liu et al. [14] proposed an end-to-end model that combines graph reasoning with instance segmentation to rank the relative saliency of multiple objects within an image. Tian et al. [15] modeled the interactions between region-level and object-level features by unifying spatial attention and object-based attention, and devised a bi-directional object-context priority learning framework.

Nevertheless, these methods attempt to capture the fuzzy object-context relations rather than explicitly model information such as the distance between objects and the object-context contrast. According to human visual cognition principles, objects with higher contrast to their surroundings are more likely to be salient, which is a significant factor in ranking objects' saliency. Additionally, the interaction between two objects diminishes as their distance increases. As shown in Fig. 1, where A represents the right goose and B represents the left goose in the

* Correspondence to: School of Automation, Southeast University, Nanjing 210096, China.
E-mail addresses: jinxiazhang@seu.edu.cn (J. Zhang), 220242159@seu.edu.cn (M. Huang), 220211895@seu.edu.cn (X. Zhu), hkwei@seu.edu.cn (H. Wei), sxfang@seu.edu.cn (S. Fang), kjzhang@seu.edu.cn (K. Zhang).

https://doi.org/10.1016/j.neucom.2025.130715
Received 10 February 2025; Received in revised form 20 May 2025; Accepted 4 June 2025
Available online 18 June 2025

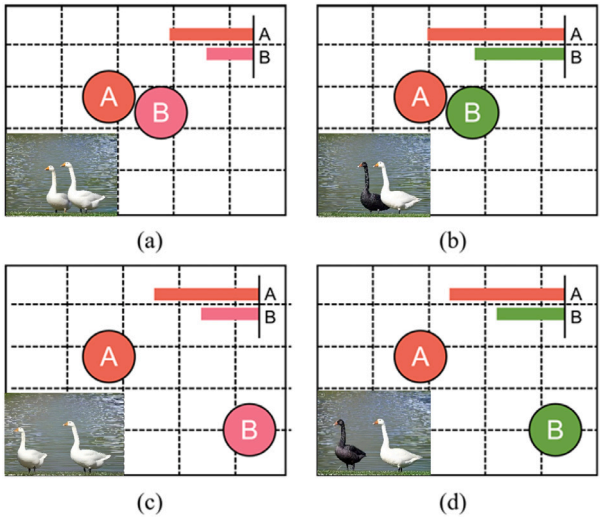

Fig. 1. Illustration of crucial factors that influence salient object ranking, where the histograms represent the saliency scores of individual objects. A represents the right goose and B represents the left goose in the lower left inset image. The contrast between objects in (a) is low, leading to mutual inhibition of their saliency scores. Conversely, the objects in (b) have high contrast, resulting in mutual facilitation of their saliency scores. A comparison of (c) with (a) reveals that increasing the distance between objects reduces the impact of mutual inhibition. Similarly, when comparing (d) with (b), increasing the distance between objects diminishes the impact of mutual facilitation.

lower left inset image, the low contrast between objects in (a) leads to mutual inhibition, which impacts their saliency scores by causing a decrease. Conversely, the objects in (b) have high contrast, resulting in mutual facilitation that impacts their saliency scores positively. Therefore, objects in (b) exhibit higher saliency scores compared to those in (a). Furthermore, comparing (c) with (a) reveals that increasing the distance between objects reduces the impact of mutual inhibition. Similarly, when comparing (d) with (b), increasing the distance between objects diminishes the impact of mutual facilitation. Beyond the impact of distance on contrast, another key factor affecting the results of salient object ranking is the correlation between object rankings, which necessitates explicit modeling.

Based on the above analysis, a method based on the Distance-weighted Contrast and Instance Correlation (DICO) is proposed for salient object ranking. Specifically, we propose the Distance-weighted Contrast (DCO) module, which effectively incorporates both contrast information and the impact of distance simultaneously. Contrast information between regions enables effective localization of salient objects, but not all inter-region contrasts yield an equal level of impact. When the distance between regions is greater, this impact becomes significantly diminished (as illustrated in Fig. 1). Therefore, a Gaussian function is employed to simulate the dynamic attention distribution between regions, aiming to model the distance-related impact.

Furthermore, we reconsider the issue of salient object ranking. Previous attempt by Islam et al. [11] aims to model it as a pixel-wise regression problem, but such a way leads to unsatisfactory results for both segmentation and ranking. Given the saliency properties of salient objects, it is more appropriate to treat them as instances. Consequently, most prior works have approached the problem from the perspective of instance segmentation and treat the saliency ranking prediction as a classification task. However, ranking tasks differ significantly from classification tasks. To better rank different salient objects, the Instance Correlation (ICO) loss is proposed in this paper, which incorporates both inter-object relations and the fitness of each individual object.

To sum up, the contributions of this work are as follows:

- Inspired by human visual perception, we propose a relative saliency ranking learning method by modeling Distance-weighted Contrast and Instance Correlation (DICO).
- We propose a novel Distance-weighted Contrast (DCO) module to model the contrast between salient objects and their contexts. By integrating dynamically generated attention distribution and feature-based contrast, the DCO module effectively quantifies the weighted influence of distance on contrast features.
- We reconsider the issue of salient object ranking and propose the Instance Correlation (ICO) loss that simultaneously takes into account the inter-object relations and individual object fitness, enabling the model to better learn the attention shift ranks of objects.
- Experimental results show that our method outperforms the state-of-the-art methods in terms of balancing computational complexity and performance.

2. Related work

2.1. Salient object detection

Since the introduction of the concept of saliency by Itti et al. [16], Liu et al. [1] made a pioneering contribution by defining SOD as a binary prediction problem. Since then, numerous algorithms have emerged in the field of SOD. Early works [17–19] mainly relied on low-level features in images to determine the salient regions. However, these manual features are unable to cope with various complex scenes. In recent years, heuristic features based on convolutional neural networks (CNNs) have gradually replaced manual features [20]. Some works combine basic processing units such as superpixels [21] and object proposals [22] with multilayer perceptrons to detect salient regions.

Later works, such as UCF [23] and Picanet [24] usually adopt fully convolutional networks (FCNs) [25] to improve computational efficiency. While choosing advanced network structures as the basis, recent methods have also adopted strategies to enrich network features, such as multi-stream information fusion [26–28], multi-level feature fusion [4,29–31], and attention mechanisms [23,32,33]. Explicit modeling of part–whole relationships [34–36] has been shown to further enhance saliency detection by capturing hierarchical object structure. Liu et al. [34] introduce Part-Object Relational Visual Saliency, which builds a graph to learn interactions between local parts and the entire object, leading to more coherent saliency maps. BCNet [35] leverages self-supervised learning to discover part–whole correspondences without pixel-level annotations. Furthermore, TCGNet [36] exploits correlations between part features and object types, guiding the network to focus on semantically consistent regions across different scales. These strategies not only effectively improve the accuracy of SOD but also help the network obtain more refined saliency maps.

Recent studies [37,38] leverage the Transformer architecture to enhance salient object detection. VST++ [37] improves efficiency and accuracy with the Select-Integrate Attention module and a novel depth position encoding. VSCode [38] unifies salient and camouflaged object detection through 2D prompt learning, providing contextual guidance for diverse datasets and tasks.

The aforementioned models in the field of SOD have become mature, but they are still designed for pixel-level binary prediction tasks and mainly focus on the edge details of salient objects. They do not predict the saliency ranking values for different objects.

2.2. Salient instance segmentation

A few works have focused on segmenting salient objects as instances. Li et al. [6] first introduced the concept of salient instance segmentation (SIS) and constructed the first dataset using pixel-level SIS annotations. Fan et al. [7] proposed S4Net, which introduces a

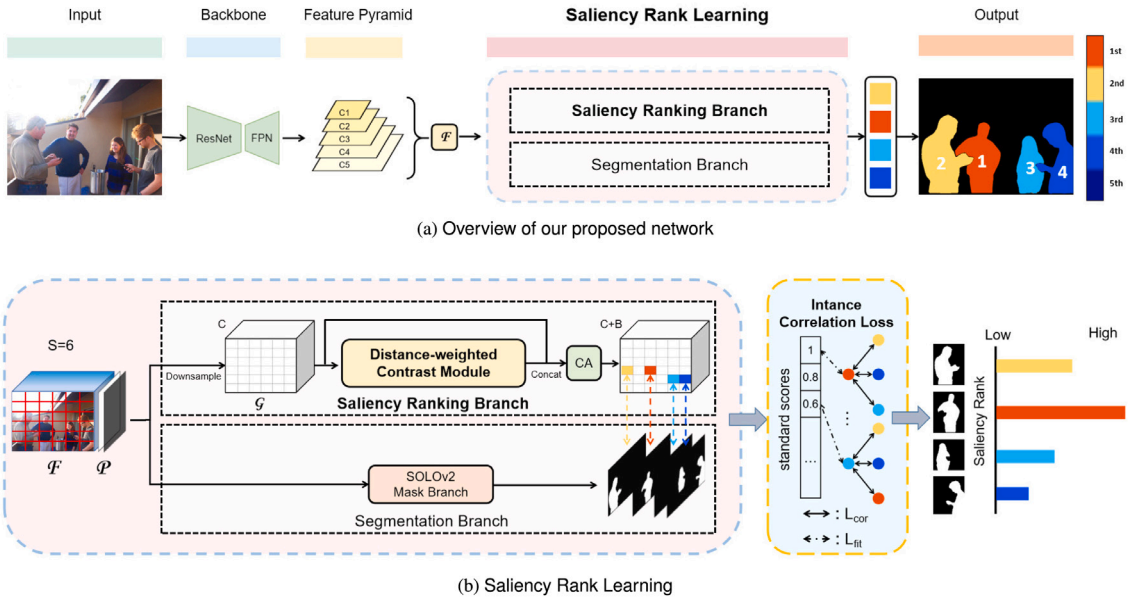

Fig. 2. Framework of DICO. Given an input image, we first apply the backbone to obtain the feature map $\mathcal{F}$, which serves as the input of the Saliency Rank Learning. Then, the SOR task is divided into two sub-tasks: saliency ranking prediction and mask segmentation, where the segmentation branch is derived from SOLOv2 [42]. The saliency ranking branch has: (1) a Distance-weighted Contrast module to extract weighted contrast features, (2) a Channel Attention (CA) module to adjust the weights between the grid feature $\mathcal{G}$ and the contrast feature $\mathcal{C}$ and (3) an Instance Correlation loss which considers both the inter-object relations and individual object fitness.

novel ROIMasking layer compared to the classical Mask-RCNN model. This innovative layer incorporates feature separation information between the objects and their surroundings, thus facilitating high-quality segmentation. Liu et al. [39] proposed a salient instance segmentation method based on Mask R-CNN by integrating saliency and contour information through a multi-scale global attention model. Tian et al. [40] proposed a weakly-supervised method that exploits class labels and subitizing labels for SIS. Chen et al. [41] proposed a keypoints-based SIS network, employing multiple keypoints as effective geometric guidance for dynamic convolutions to achieve precise segmentation of salient instances in images.

While SIS is unable to discern the relative saliency ranking among distinct salient objects, its instance-level information remains vital for salient object ranking. Therefore, we propose the Distance-weighted Contrast and Instance Correlation enhanced salient object ranking based on the advanced instance segmentation framework, SOLO [42,43]. By doing so, we achieve good segmentation performance while inferring the relative saliency ranking of each salient object.

2.3. Salient object ranking

Salient Object Ranking (SOR) is a new task. Islam et al. [11] first introduced the concept of SOR. They designed an end-to-end network based on FCN to solve multi-salient object detection problems. But this method only focused on pixel-level relative saliency. Siris et al. [12] constructed the first SOR dataset, ASSR, by combining gaze information with the existing MSCOCO dataset. Additionally, they proposed an approach that utilizes both bottom-up and top-down attention mechanisms to predict the saliency ranking of human attention shift. Although Siris et al.'s model has some object perception ability, it is not an end-to-end model. Fang et al. [13] pointed out the importance of the positional information and the interaction between objects, and proposed an end-to-end solution that combines object proposals with attention mechanisms to generate the final results. Liu et al. [14] summarized the defects of the ASSR dataset and proposed a new dataset, IRSR, with less noise and a larger salient instance number limitation. They designed a graph reasoning module based on GNNs to model the object-context relations. Tian et al. [15] also focused on the relations between objects and their contexts. They designed the selective object saliency module and the object-context-object relation module to unify spatial attention and object-based attention for SOR. Guan et al. [44] predicted the saliency ranking from attention shift through the dynamic interaction between foveal and peripheral vision.

The motivation behind these works is to model the object-context relations using tools such as attention mechanisms. This has inspired us to place a greater emphasis on the object-context relations. However, it is also worth noting that these methods rely on the network to capture fuzzy relations through learning, and do not model explicit factors between objects. Qiao et al. [45] captures semantic relationships between objects by constructing and utilizing scene graphs. However, Graph Neural Network-related modules result in a large computational load for the model. To address this limitation, we propose the Distance-weighted Contrast (DCO) module and the Instance Correlation (ICO) loss, which explicitly incorporate factors such as distance, contrast, and rank correlation as pivotal considerations in a more computationally friendly manner.

3. Methodology

3.1. Model overview

Observations from human visual perception suggest that the vision system is sensitive to contrast in visual signal [17]. On the one hand, when an object has a high contrast with its context, it is more likely to be salient. On the other hand, when two salient objects are close to each other and have high contrast, their saliency ranks will mutually enhance each other; when their contrast is low, their saliency ranks will mutually suppress each other. This interaction decays as the distance between the objects increases, as illustrated in Fig. 1. This motivates us to explicitly model this relation from the perspective of contrast.

The overall network architecture is shown in Fig. 2. Given an input image, we use the backbone to extract image feature $\mathcal{F}$. Then, the x and y coordinates are concatenated as position feature $\mathcal{P}$ and combined with the feature $\mathcal{F}$. We divide the image into $S \times S$ grids, and in Fig.

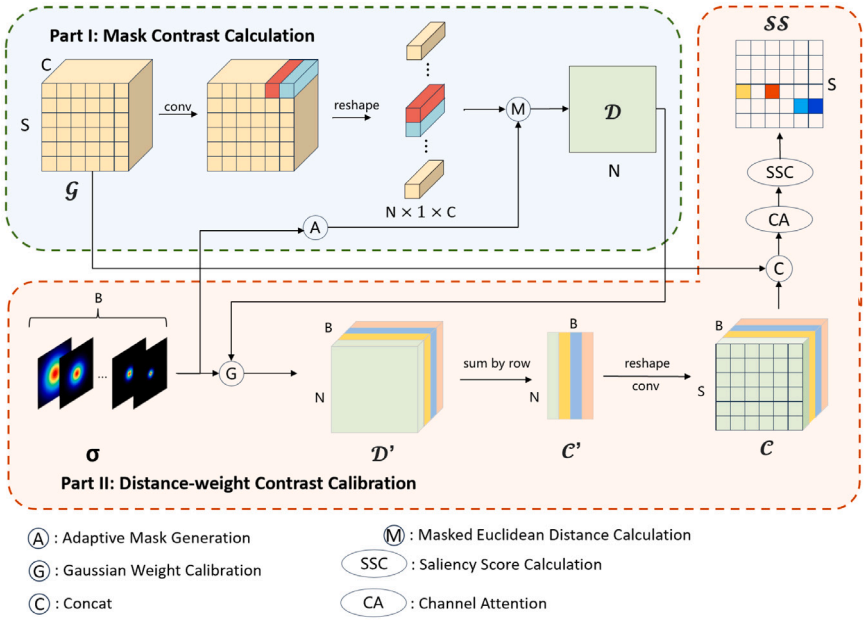

Fig. 3. Structure of our DCO module.

2 we set $S = 6$ for ease of visualization and explanation. Each grid is responsible for the SOR task of the object falling into it. The task is divided into two sub-tasks: saliency ranking prediction and mask segmentation. The segmentation branch is derived from SOLOv2 [42]. In the saliency ranking branch, the position feature $\mathcal{P}$ are removed and the feature $\mathcal{F}$ are downsampled to grid feature $\mathcal{G}$. To extract contrast features between objects and their contexts, we propose the Distance-weighted Contrast (DCO) module to convert the grid feature $\mathcal{G}$ into distance-weighted contrast feature $\mathcal{C}$. Then, we concatenate $\mathcal{C}$ and $\mathcal{G}$ along the channel dimension and input them into the Channel Attention module to dynamically adjust the weights between the channels based on the CA module proposed by Hu et al. [46]. Finally, by combining the results of saliency ranking branch and segmentation branch, we can obtain the saliency scores of the corresponding objects.

3.2. Distance-weighted contrast module

Based on human visual perception, the contrast features between regions play a critical role in detecting salient objects [17,18,47]. However, directly using region contrast features has some problems. For example, calculating the contrast between each pair of regions requires a lot of computing resources. Moreover, when the distance between two regions is far, the information provided by their contrast is almost zero. To solve these problems, we introduce distance-based weights to dynamically adjust the results of region feature contrast based on grids. The reasons for preferring grids over instances are as follows: First, grids are capable of modeling both inter-object interactions and object-context contrasts while region feature contrast based on instances can only model inter-object interactions; Second, computing region feature contrast based on grids can more effectively model Distance-weighted Contrast through matrix computations. Seychell et al. [48] pointed out that the probability of human visual fixation in the scene follows a Gaussian distribution. Therefore, we use the Gaussian function as the weight between regions to dynamically adjust the information provided by different regions according to their distance, thus obtaining more accurate results while avoiding excessive computation. Fig. 3 shows the details of the DCO module. The whole process of the DCO module primarily consists of two parts: Masked Contrast Calculation, and Distance-weighted Contrast Calibration. Through the DCO module, the saliency scores of different instances can be obtained, whose centers locate in one of the grids.

3.2.1. Masked contrast calculation

The goal of Masked Contrast Calculation is to compute the contrast between regions. We first generate the adaptive mask M by computing the Gaussian weight matrix W. The Gaussian function's output serves as the distance weight $W_{p,q}$ between each grid pair (S_p, S_q), where p and q are the grid indices. Given $S\times S$ grids, $p, q \in \{1, 2, \dots, N\}$ and $N = S\times S$. We use B different standard deviations $\sigma = [\sigma_1, \sigma_2, \dots, \sigma_B]$, resulting in B Gaussian weight matrices $W = [W^1, W^2, \dots, W^B] \in \mathbb{R}^{B\times N\times N}$:

$$W^b_{p,q} = exp(-\frac{(x_p - x_q)^2 + (y_p - y_q)^2}{2\sigma_b^2}), \tag{1}$$

where $b \in \{1, 2, \dots, B\}$ and (x_p, y_p) denotes the coordinates of grid p. Once W is obtained, we can determine the weights of the grid pair (S_p, S_q) as $W_{p,q} = [W^1_{p,q}, W^2_{p,q}, \dots, W^B_{p,q}]$. For the grid pairs where the maximum weight value still falls below the threshold T, the calculation of their contrast features can be skipped, thereby reducing the overall computational complexity. To skip the unnecessary calculation, we define the adaptive mask $M \in \mathbb{R}^{N\times N}$ as:

$$M_{p,q} = \begin{cases} 0, & max(W^1_{p,q}, W^2_{p,q}, \dots, W^B_{p,q}) < T. \\ 1, & others. \end{cases} \tag{2}$$

Then we combine feature distance calculation with the adaptive mask M to obtain the contrast between regions. Given the grid feature $\mathcal{G} \in \mathbb{R}^{C\times S\times S}$, we first split it along the channel dimension to generate channel-wise grid features $\mathcal{G} = [\mathcal{G}^1, \mathcal{G}^2, \dots, \mathcal{G}^N] \in \mathbb{R}^{C\times N}$, where $N = S \times S$. Subsequently, we compute the contrast matrix $\mathcal{D} \in \mathbb{R}^{N\times N}$ by pairwise calculating the Euclidean distance between the grid features based on the adaptive mask M:

$$\mathcal{D}_{p,q} = \begin{cases} \sum_{l=1}^{l=C} (\mathcal{G}^p_l - \mathcal{G}^q_l)^2, & M_{p,q} = 1. \\ 0, & M_{p,q} = 0. \end{cases} \tag{3}$$

3.2.2. Distance-weighted contrast calibration

Distance-weighted Contrast Calibration further utilizes the Gaussian weight matrix W and the contrast matrix $\mathcal{D}$ to generate Distance-weighted contrast feature C. Firstly, we perform element-wise multiplication between W and $\mathcal{D}$, resulting in $\mathcal{D}' \in \mathbb{R}^{B\times N\times N}$. Then, we compute the column-wise sum of $\mathcal{D}'$ to obtain the contrast each grid and its surrounding grids, denoted as $\mathcal{C}' \in \mathbb{R}^{B\times N}$:

$$\mathcal{C}'_{b,i} = \sum_{j=1}^{N} \mathcal{D}'_{b,i,j}, \quad i \in \{1, 2, \dots, N\}. \tag{4}$$

Finally, we reshape C' to its original form to get the distance-weighted contrast features $C \in \mathbb{R}^{B\times S\times S}$.

3.2.3. Saliency score calculation

To calculate saliency scores, the distance-weighted contrast features $C \in \mathbb{R}^{B\times S\times S}$ and the grid features $G \in \mathbb{R}^{C\times S\times S}$ are concatenated along the channel dimension and fed into the Channel Attention (CA) module. The combined features output by the CA module are then fed into a classification head and passed through a sigmoid activation function to generate the saliency scores $SS \in \mathbb{R}^{S\times S}$ for different instances, whose centers lie in one of the grid cells. After applying the sigmoid function, we compute the maximum value across channels to determine both the saliency rank $rank_n$ of an instance n and its associated confidence p_n. Subsequently, the saliency score SS_n of instance n is derived using the following transformation:

$$SS_n = \frac{1}{L(L+1)} \times R(rank_n) \times (p_n + R(rank_n)), \quad n \in \{1, 2, \dots, N\}. \tag{5}$$

As defined earlier, $N = S \times S$. In the above formula, L denotes the upper limit of saliency ranks (In the ASSR [12] dataset, L is 5, while in the IRSR [14] dataset, L is 8). And $R(rank_n)$ = L+1-$rank_n$, which is to ensure a one-to-one correspondence between saliency scores and instances' saliency ranks: The more front-ranked an instance is, the higher its corresponding saliency score. For example, when L is 5, the object ranked first (with $rank_n = 1$) has $R(rank_n) = 5$. Since the higher the confidence in the saliency ranking of an instance, the greater its saliency score, we also add the confidence p_n to $R(rank_n)$. For instance, in the ASSR dataset, if a object is inferred to have a rank of 1 with a confidence of 0.8, its saliency score can be calculated as follows:

$$SS_n = \frac{1}{5\times 6} \times (5 + 1 - 1) \times (0.8 + 5 + 1 = 1) = 0.97. \tag{6}$$

During the inference stage, the saliency scores are computed using the same process.

3.3. Instance correlation loss

Islam et al. [11] modeled SOR as a pixel-wise regression problem and used the pixel-wise Euclidean loss between the predicted saliency map and the ground truth (GT) as the loss function. Nevertheless, this approach did not consider salient objects as individual instances, making it challenging to accurately infer the saliency ranks of each object. Siris et al. [12] approached saliency rank prediction from the perspective of instance segmentation, modeling it as a rank order classification problem. In contrast, there are notable differences between ranking and classification tasks: for ranking tasks, there is a strong relation between different ranks, whereas for classification tasks, the relation between different classes is less significant. Liu et al. [14] proposed a ranking loss for ranking tasks, optimizing saliency scores from the perspective of instance pairs, but they did not focus on optimizing for individual instances. Furthermore, they achieved explicit optimization of instances with very high or very low ranks by assigning greater weights for pairs with large rank differences. However, these instance pairs inherently possess distinguishable characteristics in feature distributions, making them relatively easier to discriminate. The true challenge lies in modeling the correlation between instances with close ranks. Therefore, we propose the Instance Correlation (ICO) loss, which combines the considerations of inter-object relations and the fitting of the individual instance. The ICO loss is defined as:

$$L_{ico} = \alpha L_{cor} + L_{fit}, \tag{7}$$

where L_{cor} is the correlation loss, designed to model the correlation between instance pairs, particularly those with close ranks. L_{fit} is the fitting loss, aiming to promote higher saliency scores for highly salient objects and suppress saliency scores for less salient ones by fitting the saliency scores of each object to the standard scores. α represents a hyperparameter that balances the contributions of the two losses and is set to 2 in all our experiments.

3.3.1. Correlation loss

Concretely, considering a total of L saliency levels, for a training image with K instances, the GT ranks are denoted as $\{r_1, r_2, \dots, r_K\}$, where $r_k \in \{1, 2, \dots, L\}$ represents the saliency rank of instance k, and smaller values indicate higher ranks. Based on permutation, we select a total of C_K^2 instance pairs for training. For an instance pair m consisting of m_1 and m_2, with ground truth (GT) ranks (r_{m_1}, r_{m_2}), and inferred saliency scores (s_{m_1}, s_{m_2}), the correlation loss can be defined as follows:

$$L_{cor} = \sum_{m=1}^{C_K^2} \beta (s_{m_1} - s_{m_2})^2, \tag{8}$$

where β represents the weight assigned to instance pairs (m_1, m_2). When m_1 and m_2 are close to each other, it is appropriate to assign a larger value to β, thereby giving greater weight to instances that are more difficult to distinguish. Inspired by Shepard's Universal Law of Generalization [49], which shows that generalization probability decays with distance following a Gaussian or exponential curve, we employ a Gaussian function to model the weight between instance pairs. This weight naturally assign high weight to very close instances and rapidly reduces weight for those slightly farther apart. The definition of β is as follows:

$$\beta = \frac{exp(-(r_{m_1} - r_{m_2})^2/(2\mu^2))}{\sum_{o=1}^{C_K^2} exp(-(r_{o_1} - r_{o_2})^2/(2\mu^2))}, \tag{9}$$

where μ^2 is set to 1 in all our experiments.

3.3.2. Fitting loss

L_{fit} is defined as the mean-square error between the predicted saliency score and the standard saliency score, and it can be expressed as follows:

$$L_{fit} = \frac{1}{K} \sum_{k=1}^{K} (s_k - \bar{s}_k)^2, \tag{10}$$

where $\bar{s}_k$ is the standard saliency score of instance k. We consider the highest rank to have a standard saliency score of 1, while the background has a standard saliency score of 0. Assuming there are L saliency levels equally divided, for instance i with the GT rank of r_k, its standard saliency score $\bar{s}_k$ equals to $1 - (r_k - 1)/L$.

3.4. Loss function

Our training loss function is defined as follows:

$$L = L_{sor} + \lambda L_{mask}, \tag{11}$$

where L_{sor} is the loss function for training saliency ranking branch. For a salient object, $L_{sor} = L_{cls} + \gamma L_{ico}$. L_{cls} corresponds to the Focal Loss [50]. L_{ico} is the ICO loss. When the ICO loss is used, we set γ to 0.3. L_{mask} represents the loss function for segmentation Branch, which employs the Dice Loss [51]. In all our experiments, we set λ to 3.

4. Experiments

4.1. Datasets

Our experiments are conducted on the publicly available ASSR [12] and IRSR [14] datasets. ASSR is the first large-scale dataset for saliency object ranking, which is a combination of the MSCOCO [52] dataset and SALICON [53] dataset. Each image in the dataset is annotated with eye gaze information and contains up to 5 salient objects labeled with saliency ranks. It provides 7464, 1436, and 2418 images for training, validation, and testing, respectively. IRSR considers both eye gaze information and fixation durations to annotate the saliency ranks. Additionally, IRSR optimizes the distribution of salient object instances, filters out inappropriate images, and limits the number of salient instances to 8 per image. It consists of 8988 images, with 6059 images used for training and 2929 images used for testing.

Table 1
Quantitative comparison with state-of-the-art methods on the ASSR and IRSR datasets. The corresponding backbone and parameter for each method is provided. "–" indicates that the corresponding results are not provided in the source. "↑(↓)" indicates that higher (lower) values indicate better performance. The bold number is the top score and the underlined number is the second.

Method	Year	Backbone	Params	FLOPs	ASSR test set [12]				IRSR test set [14]			
					SA-SOR↑	SOR↑	Imgs Used↑	MAE↓	SA-SOR↑	SOR↑	Imgs Used↑	MAE↓
RSDNet [54]	2018	ResNet-101	–	–	–	0.728	**2418**	0.139	0.460	0.735	–	0.129
BASNet [55]	2019	ResNet-34	332M	–	–	0.707	2402	0.115	–	–	–	–
S4Net [7]	2019	ResNet-50	–	–	–	0.891	1507	0.150	–	–	–	–
CPD-R [56]	2019	ResNet-50	183M	–	–	0.766	2417	0.100	–	–	–	–
SCRN [57]	2019	ResNet-50	97M	–	–	0.756	**2418**	0.116	–	–	–	–
ASRNet [12]	2020	ResNet-101	–	–	0.667	0.792	2365	0.101	0.388	0.714	–	0.125
IRSRNet [14]	2021	ResNet-50	489M	128.79G	0.709	0.811	–	0.105	0.565	0.806	–	0.085
SOR-PPA [13]	2021	VoVNet-39	454M	311.29G	–	0.841	2371	0.081	–	–	–	–
OCOR [15]	2022	Swin-L	1.5G	592.34G	**0.738**	**0.904**	–	0.078	0.578	0.834	–	0.079
RLSOR [58]	2024	ResNet-50	–	–	0.713	0.883	–	0.074	0.570	0.822	–	0.093
PoseSOR [59]	2024	Swin Transformer	–	–	0.673	0.871	–	0.072	0.568	0.817	–	**0.063**
CBDI-SOR [60]	2024	VoVNet-39	–	–	0.725	0.850	2375	0.082	–	–	–	–
Ours(light)	–	ResNet-18	182M	68.15G	0.652	0.841	2056	0.093	0.470	0.859	2556	0.097
Ours	–	ResNet-101	601M	220.66G	0.729	0.861	2337	**0.065**	**0.587**	**0.863**	**2745**	**0.063**

4.2. Evaluation metrics

To ensure fairness, we adopt the same evaluation settings as [13, 14], using three evaluation metrics: SOR [12], SA-SOR [14], and MAE.

SOR represents the Spearman's Rank-Order correlation between the prediction and GT of the saliency ranks. A higher SOR indicates a stronger correlation between the two ranks. For ease of interpretation, we normalize the SOR score to the range of $[0, 1]$. However, in cases where there are no common salient objects between the prediction and GT, the SOR metric cannot be computed. Such cases are excluded and the number of images used for SOR calculation is reported. A higher number of images used for SOR calculation indicates a more reliable SOR metric. Moreover, another limitation of SOR is that it can only measure the accuracy of ranking but cannot characterize the precision of segmentation.

To enhance the SOR metric, the SA-SOR metric is introduced [14] to assess the Pearson correlation between the predicted saliency ranks and the GT ranks, effectively penalizes both instance omissions and redundant segmentation.

MAE metric calculates the average pixel-wise difference between the predicted saliency maps and the GT saliency maps, providing a measure of the quality of salient object ranking.

4.3. Implementation details

4.3.1. Model settings

Our experiments are based on SOLOv2 [42]. Following its setup, the number of grids corresponding to the feature pyramid levels $C1$ to $C5$ are $[40, 36, 24, 16, 12]$. We use a ResNet pretrained on ImageNet [61] as our backbone. For SOLOv2, we make some modifications in certain details. Specifically, we assign grid labels based on masks instead of bounding boxes. During the inference process, for each individual grid, we only consider the class with the highest probability.

In the DCO module, we set the threshold T as 0.01 to mask out negligible weights for computational efficiency and then assign corresponding standard deviations in Eq. (1) based on the number of grids which can be regarded as the context to compute the contrast. Specifically, for each of the five feature pyramid scales, proper standard deviations are set to make sure the context regions (i.e., the regions whose weights are higher than T) can cover the surrounding $[40, 24, 12, 8, 4]$ (modified from the grid numbers to cover a wider range of scales) grids. B is set to 5 and through calculations, the standard deviations are obtained for the five levels as $\sigma = [6.59, 3.95, 1.98, 1.32, 0.66]$.

4.3.2. Training schedule settings

We train our model for 36 epochs on each dataset. We adopt SGD [62] as our optimizer with an initial learning rate of 0.0025, which is then divided by 10 at 27th and again at 33th epoch. Weight decay of 0.0001 and momentum of 0.9 are used. The input image resolution is 640×480, and the data augmentation operations follow the settings in [42]. The training is conducted on two Tesla V100 GPUs, with a batch size of 4.

By employing this strategy, the saliency score for the background is maintained at 0. Additionally, in cases where two objects have the same inferred saliency rank, the probability term allows for differentiation, determining which object has a higher saliency score.

4.4. Main results

4.4.1. Quantitative comparison

We compare our method with various state-of-the-art approaches, including BASNet [55], S4Net [7], CPD-R [56], SCRN [57], RSDNet [54], ASRNet [12], IRSRNet [14], SOR-PPA [13], OCOR [15], RLSOR [58], PoseSOR [59] and CBDI-SOR [60]. Table 1 shows the quantitative results. The lightweight version of our model uses ResNet-18 as the backbone, with specific parameters described in [42]. On the ASSR test set, OCOR achieves the highest SA-SOR and SOR scores. However, they do not provide the number of images used for SOR calculation, and their model has more than twice the parameters and FLOPs compared to ours. S4Net ranks second in SOR but uses a small number of images and has the worst MAE. Our method achieves similar SA-SOR to OCOR and has the best MAE. On the IRSR test set, our method outperforms all other models on all metrics. Additionally, the lightweight version of our model achieves satisfactory performance with relatively few parameters. Overall, our method outperforms all methods on the test sets in terms of balancing computational complexity and performance. Particularly, it exhibits outstanding performance on the IRSR dataset, indicating its unique advantages when dealing with a larger number of salient objects and more complex scenes.

4.4.2. Qualitative comparison

We show visualization results in Fig. 4 for qualitative comparison. In columns (a)–(c), due to the explicit modeling of contrast between regions in our approach, we better preserve the edge details of individual objects, such as the feet of the girl in column (a), the skis in column (b) and the contours of the commercial truck in column (c). In columns (d)–(h), when it comes to complex scenes and multiple salient objects, our method provides better inference of the ranking among objects. For example, in column (f), our method accurately identifies the giraffe as the most salient object and get the correct result. However, the

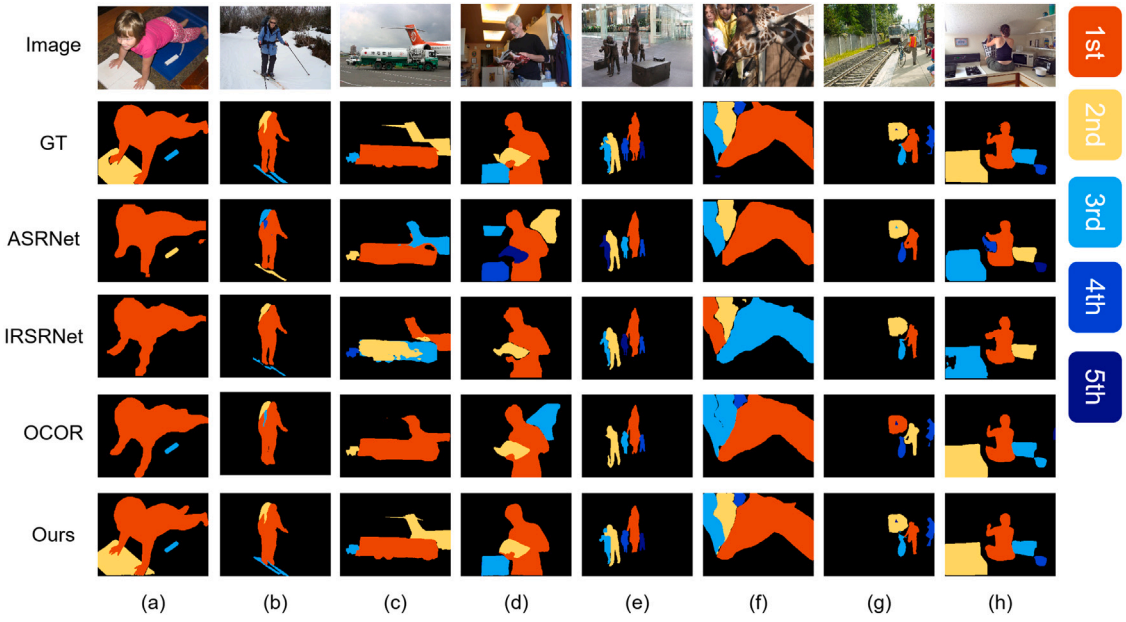

Fig. 4. Qualitative comparison of our method with other state-of-the-art approaches.

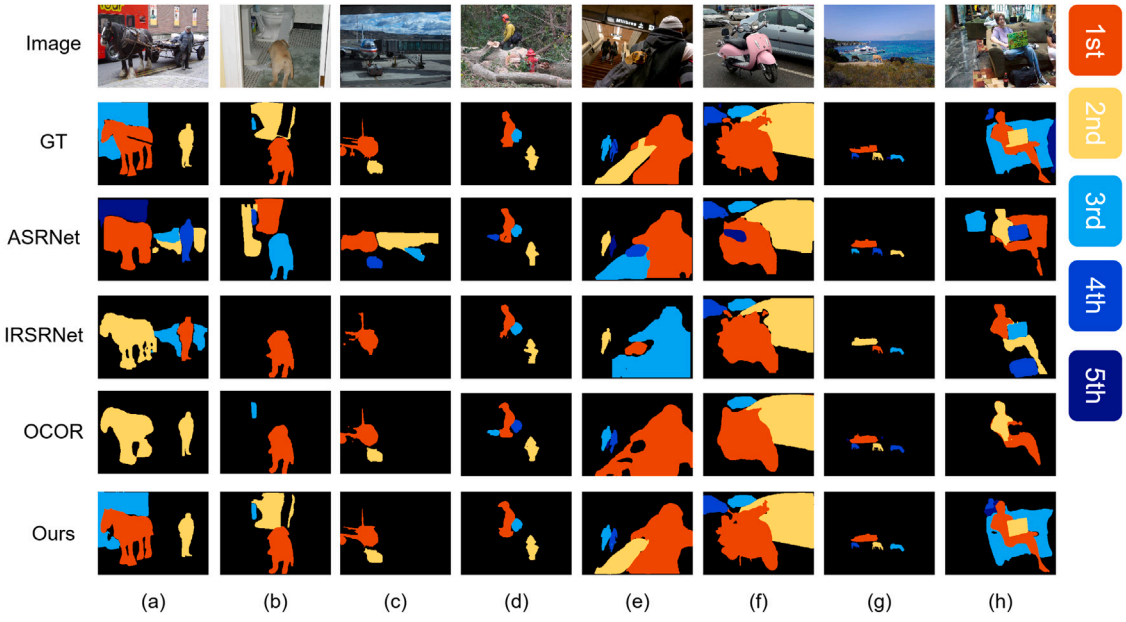

Fig. 5. Further qualitative comparison of our method with other state-of-the-art approaches in complex scenes.

IRSRNet, which favor objects with strong semantics (e.g., people), mistakenly assigns the highest saliency rank to the leftmost audience member. Conversely, while the ASRNet correctly identifies most of the salient objects, it overlooks an audience member in the corner. On the other hand, the OCOR excessively focuses on highly salient objects while disregarding those that are less salient, leading to ranking errors among several audience members.

To demonstrate the advantages of our approach in complex scenes, we provide additional visualization results as depicted in Fig. 5. These images encompass multiple salient objects and exhibit certain background interferences. Notably, our method provides better inference of the ranking among objects, further enhancing the quality of the results. For example, in column (g), the larger vessel stands out as the most salient object. However, determining the relative saliency among the three objects on the shore is challenging. Our method, by explicitly modeling distance-weighted contrast, allows us to identify the middle cow as less salient. Additionally, through modeling the instance correlation, we accurately capture the relation between the remaining two objects.

These visualization results demonstrate the enhanced capability of our method in accurately inferring relative saliency rankings.

4.5. Ablation study

4.5.1. Modifications to SOLOv2

We devise several pre-designed improvement directions for SOLOv2, which are outlined as follows:

- *masklabel*: assign grid labels based on masks instead of bounding boxes
- *maxprob*: consider only the saliency rank with the highest probability for a grid during the inference process

Table 2
Comparison of different modifications to SOLOv2.

Method				Evaluation Metrics			
masklabel	*maxprob*	*centralsample*	*centerbranch*	SA-SOR↑	SOR↑	Imgs Used↑	MAE↓
				0.655	0.864	2309	0.073
✓				0.692	0.864	2309	0.073
✓	✓			**0.713**	**0.865**	2310	0.073
✓	✓	✓		0.706	0.864	**2322**	**0.071**
✓	✓		✓	0.706	0.859	2307	0.073

Table 3
Ablation study of our method.

Method	SA-SOR↑	SOR↑	Imgs Used↑	MAE↓
SOLOv2	0.655	0.864	2309	0.073
Baseline(Modified SOLOv2)	0.713	**0.865**	2310	0.073
Baseline + ICO Loss	0.720	0.861	2323	0.066
Baseline + DCO	0.718	0.864	2332	0.069
Baseline + DCO + CA	0.722	0.861	2327	0.066
Baseline + DCO + CA + ICO Loss	**0.729**	0.861	**2337**	**0.065**

- *centralsample*: modify the labeling scheme from assigning labels based on the surrounding nine-grid area of a grid to assigning labels based on a 0.2 times the bounding box's width/height
- *centerbranch*: add a branch for the network to regress the true coordinates

The impact of the above modifications to SOLOv2 can be seen in Table 2. From the table, it can be observed that the combination of modifications *masklabel* and *maxprob* yields the best performance for SOLOv2. Therefore, we adopt this combination as our baseline.

4.5.2. Effectiveness of each component of our method

To demonstrate the effectiveness of the proposed DCO module and ICO loss, we investigate the effectiveness of them respectively. In addition, we explore the impact of modifying the SOLOv2 model and introducing the CA module on network performance. Table 3 shows the incremental effects of adding each module to the Baseline, indicating the importance of both the DCO module and the ICO loss in salient object ranking performance.

4.5.3. Choice of backbone

To further investigate the impact of advanced backbones on our method, we employed the Swin series, as detailed in Table 4. Specifically, Swin-B denotes Base-size variant of the Swin Transformer, while Swin-L denotes Large-size variant of the Swin Transformer. When utilizing more advanced backbones, our method consistently demonstrates improved performance.

As shown in Table 4, on the ASSR dataset, the SA-SOR metric of our method with Swin-L as the backbone achieves the best performance, indicating that with the support of advanced backbone, our method can more effectively segment and rank salient objects. Although the SOR metric of the OCOR method is higher than that of our method, the number of images used to evaluate this metric for OCOR has not been reported. Moreover, the SOR metric only measures ranking accuracy without penalizing either instance omissions or redundant segmentations, while the SA-SOR metric does account for these penalties.

On the IRSR dataset, our model using ResNet-101 as the backbone outperforms other methods. Moreover, its performance can be further improved by adopting more advanced backbones.

4.5.4. Alternatives of the DCO module

In the DCO module, the Gaussian weight is utilized for Distance-weighted Contrast Calibration. To further validate our DCO module, we compared the Gaussian weight in the DCO module with linear decay and uniform matrix. The results are presented in Table 5. The terms "DCO→Uni-DCO" and "DCO→Lin-DCO" signify the substitution of the Gaussian function in the DCO module with a uniform matrix and a linear decay function respectively, while keeping all other operations unchanged. The uniform matrix is employed to model long-term relationships from a global perspective. From this global perspective, even if two objects are far apart, if their contrast is high, it can still significantly influence each other's saliency scores.

Based on the results obtained from the first and third rows of Table 5, upon replacing the Gaussian function with a uniform matrix, the SA-SOR metric shows a big decrease. This can be attributed to the fact that this global contrast modeling approach does not consider the influence of distance, leading to redundant segmentation, which is penalized in SA-SOR. Since the redundant segmentation is not penalized in SOR, the SOR metric shows a spurious increase.

When the results from the second and third rows are combined and a linear decay function is substituted for the Gaussian function, both the SA-SOR metric and the SOR metric exhibit a noticeable decline. This finding underscores the effectiveness of employing the Gaussian function to model relative distance in contrast modeling, a practice that proves beneficial for the task of salient object ranking.

4.5.5. Exploration of the performance impact of the ICO loss

By modifying the loss function of our model, we compare the performance before and after incorporating the ICO loss and evaluate its effectiveness. To expedite verification, we employ a lightweight version of our model with a ResNet-18 backbone as the baseline, incorporating both the DCO and CA modules. The experimental results are presented in the upper part of Table 6. It can be seen that both components of the ICO Loss, namely the Correlation loss (L_{cor}) and the Fitting loss (L_{fit}), contribute to the enhancement of the model's performance. And the combination of these two components further improves the model's effectiveness. L_{fit} facilitates the alignment of object saliency scores with the standard scores. However, for objects that are challenging to rank solely based on their features, the effectiveness of L_{fit} is limited. In such cases, L_{cor} is necessary to model the correlation between instance pairs and correct their saliency scores accordingly. For example, when an object with a GT of Rank2 is uncertain about being predicted as Rank1, 2, or 3, it can be guided by its correlation with Rank1 as well its correlation with Rank3, to constrain its saliency score between Rank1 and Rank3.

In addition, the selection of appropriate hyperparameters also has a certain degree of impact on the performance of the model. The weight of L_{cor} is α while the strength of its influence is impacted by μ^2. Optimal effectiveness of L_{cor} can be achieved with appropriate α and μ^2. When α is very small, L_{cor} has no effect while a large α pushes saliency scores of all objects together. A small μ^2 makes L_{cor} mainly consider instance pairs with closest ranks. Higher μ^2 results in L_{cor} focusing on instance pairs with a broader range of influence and huge μ^2 makes

Table 4
Ablation study of backbone.

Method	Backbone	ASSR test set [12]				IRSR test set [14]			
		SA-SOR↑	SOR↑	Imgs Used↑	MAE↓	SA-SOR↑	SOR↑	Imgs Used↑	MAE↓
ASRNet [12]	ResNet-101	0.667	0.792	2365	0.101	0.388	0.714	–	0.125
OCOR [15]	Swin-L	0.738	**0.904**	–	0.078	0.578	0.834	–	0.079
PoseSOR [59]	Swin-L	0.673	0.871	–	0.072	0.568	0.817	–	**0.063**
Ours	Swin-B	0.745	0.843	2050	0.069	0.602	**0.867**	2870	0.074
Ours	Swin-L	**0.756**	0.865	**2389**	0.067	**0.612**	0.859	**2882**	0.070
Ours	ResNet-101	0.729	0.861	2337	**0.065**	0.587	0.863	2745	**0.063**

Table 5
Ablation study of the DCO module.

Method	SA-SOR↑	SOR↑	Imgs Used↑	MAE↓
Ours(DCO → Uni-DCO)	0.721	**0.865**	**2348**	0.069
Ours(DCO → Lin-DCO)	0.715	0.857	2319	0.071
Ours	**0.729**	0.861	2337	**0.065**

Table 6
Ablation study of hyperparameters in ICO loss. To expedite verification, we employ a lightweight version of our model with a ResNet-18 backbone as the baseline.

Method	SA-SOR↑	SOR↑	Imgs Used↑	MAE↓
Baseline	0.640	0.835	2070	0.094
Baseline + L_{fit}	0.644	0.835	2047	**0.093**
Baseline + $L_{cor}(\alpha = 2, \mu^2 = 1)$	0.643	**0.848**	2021	**0.093**
Baseline + $L_{cor}(\alpha = 2, \mu^2 = 1)$+ L_{fit}	0.652	0.841	2056	**0.093**
Baseline + $L_{cor}(\alpha = 2, \mu^2 = 3)$+ L_{fit}	0.644	0.832	2052	**0.093**
Baseline + $L_{cor}(\alpha = 1, \mu^2 = 1)$+ L_{fit}	0.645	0.834	2049	0.095
Baseline + $L_{cor}(\alpha = 3, \mu^2 = 1)$+ L_{fit}	**0.659**	0.841	2080	**0.093**
Baseline + $L_{cor}(\alpha = 4, \mu^2 = 1)$+ L_{fit}	0.651	0.834	**2091**	**0.093**
Baseline + $L_{cor}(\alpha = 6, \mu^2 = 1)$+ L_{fit}	0.638	0.835	2051	**0.093**

Table 7
Ablation study of the ICO loss. To expedite verification, we employ a lightweight version of our model with a ResNet-18 backbone.

Method	SA-SOR↑	SOR↑	Imgs Used↑	MAE↓
Ours(Square Loss → Ranking Loss)	0.649	0.840	**2330**	**0.089**
Ours(Square Loss → Triplet Loss)	0.64	0.834	2090	**0.089**
Ours(Gaussian Weight → Lin-Weight)	0.651	0.839	2145	**0.089**
Ours(Gaussian Weight → Non-Weight)	0.649	0.838	2180	0.093
Ours	**0.652**	**0.841**	2056	0.093

all the instance pairs share the equal weight. As shown in Table 6, satisfactory performance can be achieved when α is within the range of [2,4], and μ^2 is set to 1. In all other experiments, we adopt the same hyperparameters and set $\alpha = 2$ and $\mu^2 = 1$.

We further experimentally validate the applicability of square Loss in Formula (8) by replacing the square-loss term with two Margin-based losses: Ranking loss and Triplet loss. As shown in Table 7, the experimental results demonstrate that the square loss achieves superior performance. The performance degradation when using Triplet Loss originates from its requirement for at least three instances in an image: an anchor, a positive sample, and a negative sample. Consequently, Triplet Loss is ill-suited for images containing only two salient instances.

To further validate the effectiveness of the Gaussian weight in Formula (8), we conduct two ablation experiments, i.e. deleting the weight and replacing the Gaussian weight with the linear weight. As shown in Table 7, the experimental results demonstrate the effectiveness of the Gaussian weight.

4.6. Failure case analysis

When there are many salient objects in the image, it becomes challenging to differentiate between lower-ranked objects. As shown in Fig. 6, in row (a), our method confuses the third and fourth instances. In row (b), we confuses the rank of the leftmost and rightmost people. In row (c), we mix up the order of the third-ranked truck and the fourth-ranked car, both of which are located on the far left.

Actually, the above problem is also very challenging for the existing salient object ranking methods. For example, in row (b), ASRNet, IRSR-Net and OCOR did not correctly rank the second salient object. In the future, more efforts are needed to address the ranking of lower-ranked objects.

5. Conclusion

In this paper, a novel approach DICO is proposed for salient object ranking, which effectively model Distance-weighted Contrast and Instance Correlation. We propose the Distance-weighted Contrast (DCO) module, which utilizes Gaussian functions with different standard deviations to simulate the dynamic attention distribution between regions, explicitly modeling the object-context relations from a contrast perspective. Furthermore, we propose the Instance Correlation (ICO) loss which takes into account both inter-object relations and individual object fitness. Extensive experiments have verified that our approach outperforms existing state-of-the-art methods while balancing computational complexity and performance. But when there are many salient objects in the image, it becomes challenging to differentiate between lower-ranked objects. In the future, more cognitive characteristics of human visual system can be explicitly exploited to better predict the salient object ranking.

CRediT authorship contribution statement

Jinxia Zhang: Writing – review & editing, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Min Huang:** Writing – review & editing, Software, Validation, Data Curation, Visualization. **Xinchao Zhu:** Writing – original draft, Software, Methodology, Investigation, Data curation. **Haikun Wei:** Supervision. **Shixiong Fang:** Validation, Supervision. **Kanjian Zhang:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Research Fund for Advanced Ocean Institute of Southeast University, Nantong, China (GP202411), Guangdong Basic and Applied Basic Research Foundation, China (2022A1515011435), ZhiShan Scholar Program of Southeast University, China and the Fundamental Research Funds for the Central Universities, China. We also thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

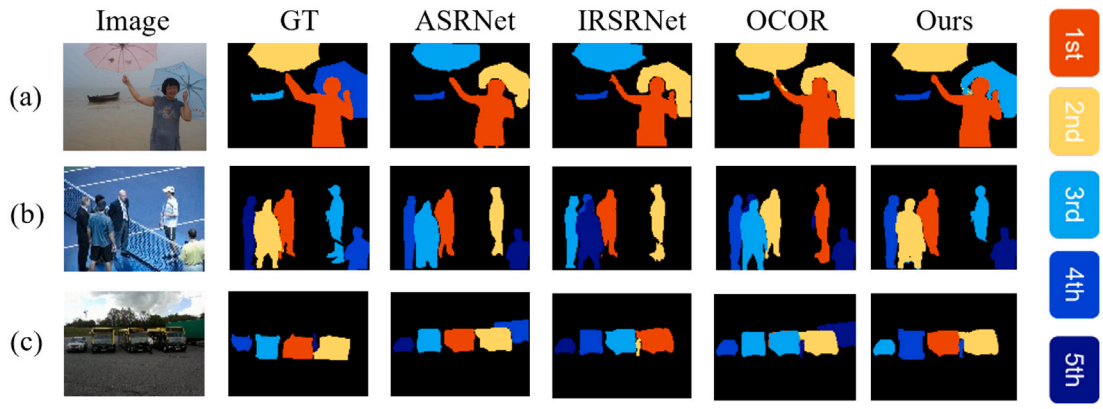

Fig. 6. Visual examples of failure cases.

Data availability

Data will be made available on request.

References

[1] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2010) 353–367, http://dx.doi.org/10.1109/CVPR.2007.383047.

[2] S. He, R.W. Lau, W. Liu, Z. Huang, Q. Yang, SuperCNN: A superpixelwise convolutional neural network for salient object detection, Int. J. Comput. Vis. 115 (2015) 330–344, http://dx.doi.org/10.1007/s11263-015-0822-0.

[3] N. Liu, J. Han, DHSNet: Deep hierarchical saliency network for salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 678–686, http://dx.doi.org/10.1109/CVPR.2016.80.

[4] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P.H. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 3203–3212, http://dx.doi.org/10.1109/TPAMI.2018.2815688.

[5] L. Zhang, J. Dai, H. Lu, Y. He, G. Wang, A bi-directional message passing model for salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1741–1750, http://dx.doi.org/10.1109/CVPR.2018.00187.

[6] G. Li, Y. Xie, L. Lin, Y. Yu, Instance-level salient object segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 2386–2395, http://dx.doi.org/10.1016/j.cviu.2021.103207.

[7] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, S.-M. Hu, S4Net: Single stage salient-instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6103–6112, http://dx.doi.org/10.1109/CVPR.2019.00626.

[8] M. Xu, L. Jiang, X. Sun, Z. Ye, Z. Wang, Learning to detect video saliency with HEVC features, IEEE Trans. Image Process. 26 (1) (2017) 369–385, http://dx.doi.org/10.1109/TIP.2016.2628583.

[9] X. Li, S. Jiang, Know more say less: Image captioning based on scene graphs, IEEE Trans. Multimed. 21 (8) (2019) 2117–2130, http://dx.doi.org/10.1109/TMM.2019.2896516.

[10] A. Palazzi, D. Abati, s. Calderara, F. Solera, R. Cucchiara, Predicting the driver's focus of attention: The dr(eye)ve project, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2019) 1720–1733, http://dx.doi.org/10.1109/TPAMI.2018.2845370.

[11] M.A. Islam, M. Kalash, N.D. Bruce, Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7142–7150, http://dx.doi.org/10.48550/arXiv.1803.05082.

[12] A. Siris, J. Jiao, G.K. Tam, X. Xie, R.W. Lau, Inferring attention shift ranks of objects for image saliency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12133–12143, http://dx.doi.org/10.1109/CVPR42600.2020.01215.

[13] H. Fang, D. Zhang, Y. Zhang, M. Chen, J. Li, Y. Hu, D. Cai, X. He, Salient object ranking with position-preserved attention, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16331–16341, http://dx.doi.org/10.48550/arXiv.2106.05047.

[14] N. Liu, L. Li, W. Zhao, J. Han, L. Shao, Instance-level relative saliency ranking with graph reasoning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 8321–8337, http://dx.doi.org/10.1109/TPAMI.2021.3107872.

[15] X. Tian, K. Xu, X. Yang, L. Du, B. Yin, R.W. Lau, Bi-directional object-context prioritization learning for saliency ranking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5882–5891, http://dx.doi.org/10.1109/CVPR52688.2022.00579.

[16] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259, http://dx.doi.org/10.1109/34.730558.

[17] M.-M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2014) 569–582, http://dx.doi.org/10.1109/TPAMI.2014.2345401.

[18] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 733–740, http://dx.doi.org/10.1109/CVPR.2012.6247743.

[19] Y. Ji, H. Zhang, K.-K. Tseng, T.W. Chow, Q.M.J. Wu, Graph model-based salient object detection using objectness and multiple saliency cues, Neurocomputing 323 (2019) 188–202, http://dx.doi.org/10.1016/j.neucom.2018.09.081.

[20] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, J. Li, Salient object detection: A survey, Comput. Vis. Media 5 (2019) 117–150, http://dx.doi.org/10.1007/s41095-019-0149-9.

[21] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 1265–1274, http://dx.doi.org/10.1109/CVPR.2015.7298731.

[22] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, R. Mech, Unconstrained salient object detection via proposal subset optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 5733–5742, http://dx.doi.org/10.1109/CVPR.2016.618.

[23] P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 212–221, http://dx.doi.org/10.48550/arXiv.1708.02031.

[24] N. Liu, J. Han, M.-H. Yang, PiCANet: Learning pixel-wise contextual attention for saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089–3098, http://dx.doi.org/10.1109/CVPR.2018.00326.

[25] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440, http://dx.doi.org/10.1109/TPAMI.2016.2572683.

[26] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 4019–4028, http://dx.doi.org/10.1109/ICCV.2017.433.

[27] X. Chen, A. Zheng, J. Li, F. Lu, Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic cnns, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 1050–1058, http://dx.doi.org/10.1109/ICCV.2017.119.

[28] X. Zhou, K. Shen, L. Weng, R. Cong, B. Zheng, J. Zhang, C. Yan, Edge-guided recurrent positioning network for salient object detection in optical remote sensing images, IEEE Trans. Cybern. 53 (1) (2023) 539–552, http://dx.doi.org/10.1109/TCYB.2022.3163152.

[29] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 202–211, http://dx.doi.org/10.1109/ICCV.2017.31.

[30] Y. Liu, M.-M. Cheng, X.-Y. Zhang, G.-Y. Nie, M. Wang, DNA: Deeply supervised nonlinear aggregation for salient object detection, IEEE Trans. Cybern. 52 (7) (2022) 6131–6142, http://dx.doi.org/10.1109/TCYB.2021.3051350.

[31] H. Wang, Y. Wang, H. Wang, J. Zhao, Hierarchical-model salient object detection based on manifold ranking, Neurocomputing 398 (2020) 460–468, http://dx.doi.org/10.1016/j.neucom.2019.04.096.

[32] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1448–1457, http://dx.doi.org/10.1109/CVPR.2019.00154.

[33] J. Li, Z. Pan, Q. Liu, Y. Cui, Y. Sun, Complementarity-aware attention network for salient object detection, IEEE Trans. Cybern. 52 (2) (2022) 873–886, http://dx.doi.org/10.1109/TCYB.2020.2988093.

[34] Y. Liu, D. Zhang, Q. Zhang, J. Han, Part-object relational visual saliency, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2022) 3688–3704, http://dx.doi.org/10.1109/TPAMI.2021.3053577.

[35] Y. Liu, X. Dong, D. Zhang, S. Xu, Deep unsupervised part-whole relational visual saliency, Neurocomputing 563 (2024) 126916, http://dx.doi.org/10.1016/j.neucom.2023.126916.

[36] Y. Liu, L. Zhou, G. Wu, S. Xu, J. Han, TCGNet: Type-correlation guidance for salient object detection, IEEE Trans. Intell. Transp. Syst. 25 (7) (2024) 6633–6644, http://dx.doi.org/10.1109/TITS.2023.3342811.

[37] N. Liu, Z. Luo, N. Zhang, J. Han, VST++: Efficient and stronger visual saliency transformer, IEEE Trans. Pattern Anal. Mach. Intell. 46 (11) (2024) 7300–7316, http://dx.doi.org/10.1109/tpami.2024.3388153.

[38] Z. Luo, N. Liu, W. Zhao, X. Yang, D. Zhang, D.-P. Fan, F. Khan, J. Han, VSCode: General visual salient and camouflaged object detection with 2d prompt learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17169–17180, URL https://arxiv.org/abs/2311.15011.

[39] N. Liu, W. Zhao, L. Shao, J. Han, SCG: Saliency and contour guided salient instance segmentation, IEEE Trans. Image Process. 30 (2021) 5862–5874, http://dx.doi.org/10.1109/TIP.2021.3088282.

[40] X. Tian, K. Xu, X. Yang, B. Yin, R.W. Lau, Learning to detect instance-level salient objects using complementary image labels, Int. J. Comput. Vis. 130 (3) (2022) 729–746, http://dx.doi.org/10.1007/s11263-021-01553-w.

[41] J. Chen, R. Cong, H.H.S. Ip, S. Kwong, KepSalinst: Using peripheral points to delineate salient instances, IEEE Trans. Cybern. (2023) 1–14, http://dx.doi.org/10.1109/TCYB.2023.3326165.

[42] X. Wang, R. Zhang, T. Kong, L. Li, C. Shen, SOLOv2: Dynamic and fast instance segmentation, Adv. Neural Inf. Process. Syst. 33 (2020) 17721–17732, http://dx.doi.org/10.48550/arXiv.2003.10152.

[43] X. Wang, T. Kong, C. Shen, Y. Jiang, L. Li, SOLO segmenting objects by locations, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 649–665, http://dx.doi.org/10.1007/978-3-030-58523-5_38.

[44] H. Guan, R.W. Lau, SeqRank: Sequential ranking of salient objects, Proc. AAAI Conf. Artif. Intell. 38 (3) (2024) 1941–1949, http://dx.doi.org/10.1609/aaai.v38i3.27964.

[45] M. Qiao, M. Xu, L. Jiang, P. Lei, S. Wen, Y. Chen, L. Sigal, HyperSOR: context-aware graph hypernetwork for salient object ranking, IEEE Trans. Pattern Anal. Mach. Intell. 46 (9) (2024) 5873–5889.

[46] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, http://dx.doi.org/10.1109/CVPR.2018.00745.

[47] Y. Hu, D. Rajan, L.-T. Chia, Robust subspace analysis for detecting visual attention regions in images, in: Proceedings of the ACM International Conference on Multimedia, 2005, pp. 716–724, http://dx.doi.org/10.1145/1101149.1101306.

[48] D. Seychell, C.J. Debono, Ranking regions of visual saliency in rgb-d content, in: International Conference on 3D Immersion, IEEE, 2018, pp. 1–8, http://dx.doi.org/10.1109/IC3D.2018.8657902.

[49] R.N. Shepard, Toward a universal law of generalization for psychological science, Science 237 (4820) (1987) 1317–1323, http://dx.doi.org/10.1126/science.3629243.

[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 2980–2988, http://dx.doi.org/10.1109/ICCV.2017.324.

[51] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: International Conference on 3D Vision, IEEE, 2016, pp. 565–571, http://dx.doi.org/10.1109/3DV.2016.79.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 740–755, http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[53] M. Jiang, S. Huang, J. Duan, Q. Zhao, SALICON: Saliency in context, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 1072–1080, http://dx.doi.org/10.1109/CVPR.2015.7298710.

[54] A.P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, N. Padoy, RSDNet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations, IEEE Trans. Med. Imaging 38 (4) (2018) 1069–1078, http://dx.doi.org/10.1109/TMI.2018.2878055.

[55] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, BASNet: Boundary-aware salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7479–7489, http://dx.doi.org/10.1109/CVPR.2019.00766.

[56] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3907–3916, http://dx.doi.org/10.1109/CVPR.2019.00403.

[57] Z. Wu, L. Su, Q. Huang, Stacked cross refinement network for edge-aware salient object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7264–7273, http://dx.doi.org/10.1109/ICCV.2019.00736.

[58] Y. Guo, S. Chen, G. Yan, S. Di, X. Lv, Salient Object Ranking: Saliency model on relativity learning and evaluation metric on triple accuracy, Displays 85 (2024) 102855, http://dx.doi.org/10.1016/j.displa.2024.102855.

[59] H. Guan, R.W.H. Lau, PoseSOR: Human pose can guide our attention, in: ECCV, Springer Nature Switzerland, 2024, pp. 350–366, http://dx.doi.org/10.1007/978-3-031-72649-1_20.

[60] H. Zhai, Z. Chen, C. Liu, H. Bai, Q.J. Wu, Category-based depth incorporation for salient object ranking, J. Vis. Commun. Image Represent. 101 (2024) 104165, http://dx.doi.org/10.1016/j.jvcir.2024.104165.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255, http://dx.doi.org/10.1109/CVPR.2009.5206848.

[62] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. (1951) 400–407, http://dx.doi.org/10.1109/TSMC.1971.4308316.

Jinxia Zhang received the B.S. degree in computer science and technology and Ph.D. degree in control science and engineering from the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China, in 2009 and 2015, respectively.

From 2012 to 2014, she was a Visiting Scholar with Visual Attention Lab at Brigham and Womens Hospital and Harvard Medical School. She is currently an Associate Professor with the School of Automation, Southeast University. Her research interests include saliency detection, knowledge transfer, computer vision, and machine learning.

Xinchao Zhu received the B.S. degree in the School of Automation, Southeast University, Nanjing, China in 2024. He has received the master's degree with the School of Automation, Southeast University, Nanjing, China. His research interests include deep learning, machine learning, and saliency object ranking.

Min Huang received the B.S. degree in the School of Intelligent Science and Technology, University of Shanghai for Science and Technology, Shanghai, China in 2024. She is currently pursuing the master's degree with the School of Automation, Southeast University, Nanjing, China. Her research interests include multi-modal learning, machine learning, and industrial anomaly detection.

Haikun Wei received the B.S. degree in industrial automation from the Department of Automation, North China University of Technology, Beijing, China, in 1994, and the M.S. and Ph.D. degrees in control theory and control engineering from the Research Institute of Automation, Southeast University, Nanjing, China, in 1997 and 2000, respectively.

From 2005 to 2007, he was a Visiting Scholar with RIKEN Brain Science Institute, Japan. He is currently a Professor with the School of Automation, Southeast University. His research interest is real and artificial in neural networks and industry automation.

Shixiong Fang received the B.S. degree in electrical engineering from the School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, China, in 2009, and the M.S. and Ph.D. degrees in control theory and control engineering from Southeast University, Nanjing, China, in 2002 and 2009, respectively.

From 2004 to 2006, he was a Visiting Scholar with the European Organization for Nuclear Research. He is currently a Lecturer with the School of Automation, Southeast University. His research interests include computer vision and cyber physical systems.

Kanjian Zhang received the B.S. degree in mathematics from Nankai University, Tianjin, China, in 1994, and the M.S. and Ph. D. degrees in control theory and control engineering from Southeast University, Nanjing, China, in 1997 and 2000, respectively.

He is currently a Professor with the School of Automation, Southeast University. His research interests include nonlinear control theory and its applications, with particular interest in robust output feedback design and optimization control.