PoseSOR: Human Pose Can Guide Our Attention

Huankang Guan and Rynson W.H. Lau *

Department of Computer Science, City University of Hong Kong

Abstract. Salient Object Ranking (SOR) aims to study how human observers shift their attention among various objects within a scene. Previous works attempt to excavate explicit visual saliency cues, *e.g.*, spatial frequency and semantic context, to tackle this challenge. However, these visual saliency cues may fall short in handling real-world scenarios, which often involve various human activities and interactions. We observe that human observers' attention can be reflexively guided by the poses and gestures of the people in the scene, which indicate their activities. For example, observers tend to shift their attention to follow others' head orientation or running/walking direction to anticipate what will happen. Inspired by this observation, we propose to exploit human poses in understanding high-level interactions between human participants and their surroundings for robust salient object ranking. Specifically, we propose PoseSOR, a human pose-aware SOR model for the SOR task, with two novel modules: 1) a Pose-Aware Interaction (PAI) module to integrate human pose knowledge into salient object queries for learning high-level interactions, and 2) a Pose-Driven Ranking (PDR) module to apply pose knowledge as directional cues to help predict where human attention will shift to. To our knowledge, our approach is the first to explore human pose for salient object ranking. Extensive experiments demonstrate the effectiveness of our method, particularly in complex scenes, and our model sets the new state-of-the-art on the SOR benchmarks. Code and dataset are available at `https://github.com/guanhuankang/ECCV24PoseSOR`.

Keywords: Salient Object Ranking · Human Pose · Attention Shift

1 Introduction

Salient object detection (SOD) is a fundamental task in computer vision, aiming to recognize objects that naturally attract human attention. However, SOD methods [17, 30, 32, 45, 49, 50, 52, 55, 56] usually treat all salient objects equally, which does not align with human viewing behaviors [21, 38]. Psychological studies [19, 21, 22] show that humans tend to sequentially select and shift their attention from one salient object to another, instead of attending to all salient

* Rynson W.H. Lau is the corresponding author (Rynson.Lau@cityu.edu.hk).

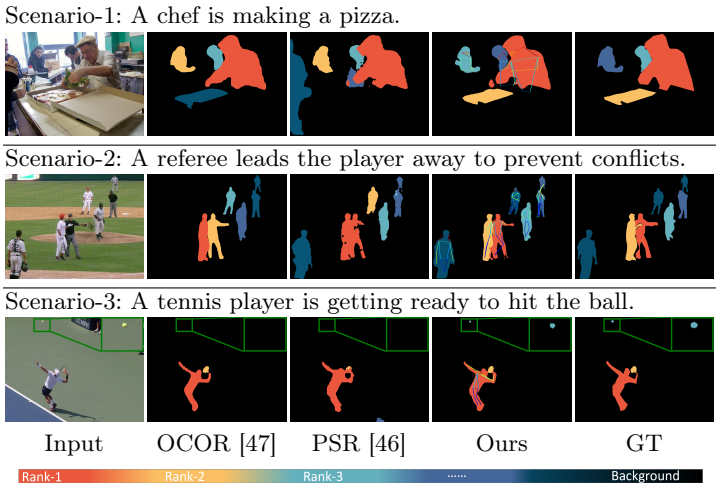

Fig. 1: Existing SOR methods are based on learning explicit visual saliency cues, and may fail to handle complex human activities and interactions. In Scenario 1, humans would tend to look at the chef in the middle first, and then the pizza that the chef is making. SOTA methods, OCOR [47] and PSR [46], are biased towards semantic-rich objects (*i.e.*, the persons), and do not include the pizza until the very end. In Scenario 2, SOTA methods tend to give higher ranks to high-contrast objects (*i.e.*, the two players in white shirts and red hats). In Scenario 3, SOTA methods tend to miss out small and less salient objects (*i.e.*, the tennis ball), even though they may be important. Our method integrates human pose knowledge in the prediction, and yields predictions that are more in line with human labels.

objects at the same time. This discrepancy has recently spurred the development of Salient Object Ranking (SOR) [44]. SOR aims to not only detect the salient objects, but also predict the visiting order of human visual attention on these objects. It can help better understand how humans interpret scenes and benefit many downstream tasks, such as image editing [1, 36], gaze communication [10], and scene understanding [9, 26].

Recently, several works [12, 31, 44, 46, 47] are proposed to tackle the SOR task. For example, Siris *et al.* [44] propose to infer attention shifts by incorporating both object-scene context and spatial mask cues. Liu *et al.* [31] use graph convolution to model instance-level competition for saliency ranking. Fang *et al.* [12] propose to learn an explicit position embedding to enhance the ranking performance. Tian *et al.* [47] propose to predict saliency ranking by unifying spatial-attention and object-based attention. All these methods mainly focus on exploring explicit visual saliency cues, such as spatial frequency and semantic context, to model how objects compete for saliency. While visual saliency can help distinguish salient objects from their surroundings, it is not sufficient to handle complex scenarios, which often involve various human activities and

high-level interactions among human participants. For example, the first row in Figure 1 depicts a scenario where the chef is making a pizza while the customers are waiting. Our method first attends to the chef and then shifts the attention to the pizza, which he is interacting with, before reaching the customer. This is in line with the human label (GT). In contrast, existing methods [46, 47] shift the attention from the chef directly to the distant customer, and they either attend to the pizza at the end [47] or completely miss the pizza [46].

We observe that human observers' attention can be reflexively guided by the behaviors of people in the scene. For example, observers tend to shift their attention to follow people's head orientation [8, 24] to find out what they are looking at. They also tend to shift their attention to follow people's walking/running direction [3, 14, 43] to find out what they are doing. Based on this observation, we propose to incorporate human pose cues into the SOR model, to facilitate the understanding of human activities and interactions in SOR. To this end, we present *PoseSOR*, a novel human pose-aware SOR model. PoseSOR includes two novel modules: 1) a Pose-Aware Interaction (PAI) module to integrate human pose knowledge into salient object queries to facilitate the learning of high-level interactions between people and their surroundings, and 2) a Pose-Driven Ranking (PDR) module to leverage human pose knowledge as directional cues to help predict where human attention will shift to. Our PAI module contains a joint-level stage and an instance-level interaction stage. The former stage aims to model inter-joint relations for individual pose understanding, and the latter stage aims to model the interactions between people and their surroundings in a pose-aware manner. The proposed PDR module leverages the pose knowledge from the PAI module to help predict the next object that human attention will likely shift to. Finally, PoseSOR learns to predict the overall rank order with the knowledge learned from both modules. To our knowledge, we are the first to uncover human pose cues for salient object ranking. We have conducted extensive experiments to demonstrate the effectiveness of our approach, and our model achieves state-of-the-art performances on the existing SOR benchmarks.

In summary, our main contributions of this work are three-fold: i) We propose to explore human pose cues for salient object ranking. These human pose cues convey information related to human activities and interactions. ii) We propose PoseSOR, a human pose-aware SOR model, which includes two novel modules, *i.e.* Pose-Aware Interaction (PAI) module and Pose-Driven Ranking (PDR) module, to explore human pose knowledge for SOR. iii) We conduct extensive experiments to verify the effectiveness of our approach, and our model outperforms existing state-of-the-art methods both quantitatively and qualitatively on SOR benchmarks.

2 Related Work

2.1 Salient Object Ranking

Salient Object Ranking (SOR) is a rather new research field. It studies how humans sequentially prioritize their attention to daily scene objects. Islam *et*

al. [18] make an initial attempt to address this task by introducing a rank-aware approach, in which they utilize the frequency of an object labeled as salient to guide the learning process. However, their approach does not account for attention shifts, and is limited to pixel-level saliency competition. Later, Siris *et al*. [44] enhance the concept of SOR by incorporating psychological and neuroscientific evidences [7, 38], showcasing that SOR can be formulated as a process of predicting the order of an observer's attention on objects. They propose the first instance-level SOR model by exploring the object-scene context and spatial mask cues, and also constructing a large-scale SOR benchmark, which is commonly used by the subsequent works. Fang *et al*. [12] suggest adding position information to enhance the ranking performances, and propose a position-preserved attention module for their goal. Liu *et al*. [31] construct a new SOR dataset with fewer annotation errors and propose to use graph convolution for learning instance-level saliency competition. Tian *et al*. [47] propose to combine both spatial-attention and object-based attention to improve salient object ranking. Most recently, Sun *et al*. [46] propose ranking salient objects by partition other than ranking by sorting for alleviating the ranking ambiguities. Guan *et al*. [15] propose to mimic human visual behaviors and rank salient objects in sequence.

While these works [12, 15, 18, 31, 44, 46, 47] have achieved some progress in salient object ranking, their solutions are limited to exploiting explicit visual saliency cues, such as spatial frequency and semantic context, or reducing the ranking ambiguities [46]. In this work, we propose to excavate the human pose cues, which convey information related to human activities and interactions between people and their surroundings, for the SOR task.

2.2 Salient Object Detection

Salient Object Detection (SOD) is a topic closely related to SOR and has been widely studied. SOD aims to recognize objects that naturally capture human attention. Traditional SOD approaches mainly rely on local appearances [2,6] and low-level cues, such as center prior [6,54], boundary prior [51,57], and background prior [20, 27]. However, traditional methods usually suffer from insufficient semantic understanding. Subsequently, deep learning based methods [17,30,32,45,49,50,52,55,56] have become popular and significantly improved the SOD performances with various strategies, such as exploring multi-level or multi-scale feature fusion strategies [49, 55], utilizing boundary features [50, 52], and introducing advanced network architectures [30, 32].

Despite the success, SOD approaches can only produce binary saliency maps, without being able to differentiate object instances. Salient instance detection [25] (SID) is then proposed to close this gap, aiming to identify salient object instances. Most of the initial SID methods [11, 33, 53] adopt Mask R-CNN [16] to detect object instances, and learn to differentiate the salient ones from the background. Recently, Pei *et al*. [39] propose to detect salient instances using transformers, and to initialize the object queries with center prior [6] to accelerate convergence.

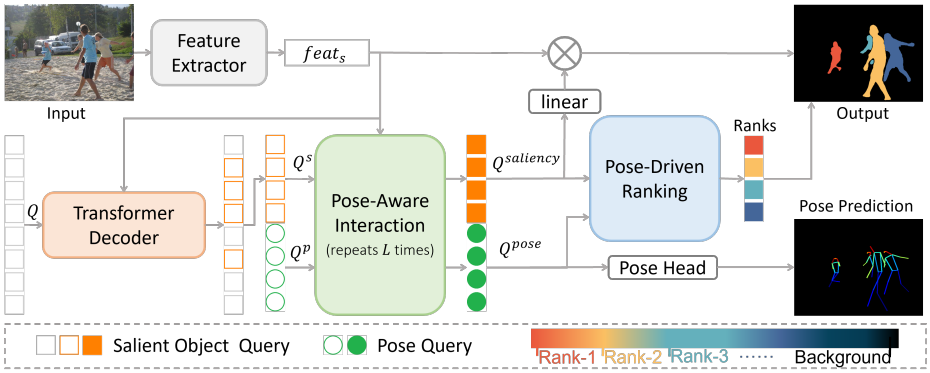

Fig. 2: The PoseSOR Framework. PoseSOR follows a query-based architecture that includes salient object queries to locate salient objects and pose queries for detecting human poses. The Pose-Aware Interaction (PAI) module learns to integrate human pose knowledge into salient object queries, and enables the modeling of high-level interactions between people and their surroundings in a pose-aware manner. The Pose-Driven Ranking (PDR) module excavates directional cues from pose queries to assist in the prediction of the overall rank order.

In contrast to SOD and SID, salient object ranking (SOR) is more challenging, as it further takes human attention shifts into account. SOR requires distinguishing the salient instances from the background, and assigning a rank to each instance to indicate the visiting order of human attention on these objects. In this work, we propose PoseSOR, a human pose-aware SOR model, to address the SOR task.

3 Method

Human attention can be reflexively guided by the poses and gestures of people in the scene, suggesting that human poses are strong cues for salient object ranking. We present PoseSOR to explore the human pose cues, which convey information about human activities and interactions, for the SOR task. We first describe the PoseSOR framework in Section 3.1. We then show how to integrate human pose knowledge into salient object queries and model interactions between people and their surroundings in a pose-aware manner via our novel Pose-Aware Interaction (PAI) module, described in Section 3.2. We propose our novel Pose-Driven Ranking (PDR) module, which leverages pose knowledge as directional cues to help predict where human attention will shift to, described in Section 3.3. Finally, we present our training strategy in Section 3.4.

3.1 PoseSOR

Figure 2 shows the PoseSOR framework. We first feed the input image to the feature extractor to extract multi-scale image features, $feat_s \in R^{C \times H_s \times W_s}$, where $s \in \{2, 3, 4, 5\}$ is the stage number and C corresponds to the feature dimension. The extracted image features $feat_s$ are shareable for the subsequent Transformer decoder, the PAI module, and the PDR module.

We start with N salient object queries (*i.e.*, □ in Figure 2), denoted as $Q \in R^{N \times C}$, where N is set to be significantly larger than the typical number of salient objects in the image. Q represents all potential salient objects in the input image and is initialized with zeros at the beginning. We next employ a 6-layer Transformer decoder [48] to enrich Q with salient object features, and select the queries with high confidence of being salient for the subsequent PAI module. This learning and selection process is formulated as:

$$Q = \text{Transformers}(Q, feat_{s|s\in\{3,4,5\}}), \tag{1}$$

$$\hat{y} = \sigma(\text{proj}(Q)), \tag{2}$$

$$Q^s = \{Q_i \,|\, \hat{y}_i > \tau \,,\, i = 1, 2, ..., N\}, \tag{3}$$

where $feat_s$ serves as the *key* and *value*, while Q serves as the input sequence to the Transformer decoder. proj is a linear layer that projects C-dimensional features to 1-dimensional features. σ is a sigmoid function. $\hat{y} \in R^N$ indicates the likelihood of each query being salient. $\tau \in R$ is a threshold for determining salient object queries. $Q^s \in R^{M \times C}$ (*i.e.*, □ in Figure 2) represents the selected salient object queries that are likely to be salient, where M is the number of detected salient objects in the input image.

Our PAI module uncovers human pose knowledge in a top-down manner, by first detecting the persons and then locating the skeletal joints for each person. We introduce pose queries (*i.e.*, ○ in Figure 2), denoted as $Q^p \in R^{M \times K \times C}$, which contains M instances with K joints each. We obtain Q^p by summing up every pair of salient object query and skeletal joint representation:

$$Q^p_{i,j} = Q^s_i + J_j, \tag{4}$$

where $J \in R^{K \times C}$ is the latent joint representation, with $i \in \{1, 2, ..., M\}$ and $j \in \{1, 2, ..., K\}$. We use the pose queries Q^p to capture human pose knowledge.

The PAI module learns to integrate pose knowledge into salient object queries, and model high-level interactions between people and their surroundings, generating pose-aware salient object queries (*i.e.*, ■ in Figure 2), denoted as $Q^{saliency}$, and refined pose queries (*i.e.*, ● in Figure 2), denoted as Q^{pose}. We apply a dot-product between $Q^{saliency}$ and $feat_2$ to obtain salient instance masks as:

$$masks = \sigma(\text{linear}(Q^{saliency}) \cdot feat_2), \tag{5}$$

where linear is a linear layer, and $masks \in R^{M \times H_2 \times W_2}$ are salient instance masks. The rankings of salient instances are determined by the PDR module as:

$$ranks = \text{PDR}(Q^{saliency}, Q^{pose}). \tag{6}$$

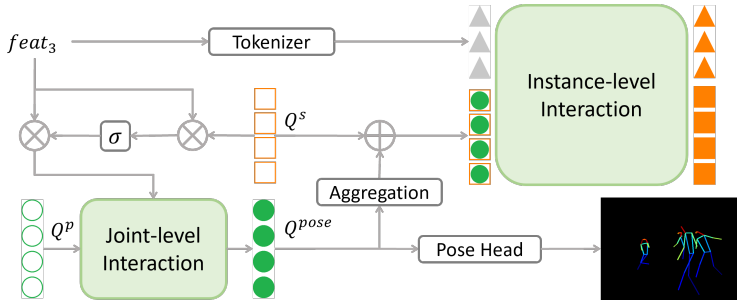

Fig. 3: Pose-Aware Interaction (PAI) Module. The PAI module consists of a joint-level interaction stage to model inter-joint relations for acquiring individual pose knowledge, and an instance-level interaction stage to capture the interactions between people and other salient items or scene elements, for understanding human activities and high-level interactions in a pose-aware manner.

Finally, we combine the salient instance masks and rankings to get final output.

3.2 The Pose-Aware Interaction (PAI) Module

Our PAI module is proposed to uncover human pose knowledge, and learn to integrate pose knowledge into salient object queries for a deeper understanding of human activities and interactions. Figure 3 shows the design of our PAI module. The PAI module includes a joint-level interaction stage to model inter-joint relations for individual pose understanding, and an instance-level interaction stage to model interactions between people and other items that they may be interacting with. We repeat the PAI module for L times to ensure comprehensive pose knowledge learning.

The joint-level interaction is specific to each instance, as each instance has its own pose information. Thus, we first generate instance-specific features by multiplying the image features with the intermediate instance maps as:

$$maps_i = \sigma(Q_i^s \cdot feat_3), \tag{7}$$

$$feat_i^{ins} = maps_i * feat_3, \tag{8}$$

where $maps \in R^{M \times H_3 \times W_3}$ are intermediate instance maps, $*$ is the element-wise multiplication, and $feat^{ins} \in R^{M \times C \times H_3 \times W_3}$ are the generated instance-specific features. We then use cross-attention and self-attention to model the joint-level interactions as they are effective in capturing complex long-range relations:

$$Q^{pose} = \text{CrossAttn}(Q^p, \rho(feat^{ins})), \tag{9}$$

$$Q^{pose} = \text{FFN}(\text{SelfAttn}(Q^{pose})), \tag{10}$$

where ρ is the flatten and permute operation, $\rho(feat^{ins}) \in R^{M \times H_3 W_3 \times C}$ acts as the *key* and *value* for the cross-attention layer. This enables the pose queries to

attend to all pixels and aggregate the features of human skeletal joints. SelfAttn is a self-attention layer for modeling joint-level interactions, and FFN is a feed-forward neural network. $Q^{pose} \in R^{M \times K \times C}$ are directly used to predict the coordinates of human skeletal joints.

We transform the pose queries Q^{pose}, which have M human poses with K joints each, with a linear layer to aggregate the K joints to one global human pose. We then incorporate this pose knowledge, *i.e.*, aggregate(Q^{pose}), into the salient object queries Q^s, to allow the pose knowledge to be exploited in the latter instance-level interaction stage.

Since the surroundings where people stay also influences human activities and interactions, we introduce a scene tokenizer to add some scene tokens:

$$T^s = \text{LN}(\rho(\text{MaxPool}(feat_3))), \tag{11}$$

where LN is the layer normalization operation, and MaxPool denotes an adaptive max pooling layer. $T^s \in R^{S \times C}$ are scene tokens. We concatenate scene tokens T^s and salient object queries Q^s to form the input sequence to the instance-level interaction stage, which is composed of a self-attention layer and a feed-forward neural network:

$$[T^{scene}, Q^{saliency}] = \text{FFN}(\text{SelfAttn}([T^s, Q^s])), \tag{12}$$

where $[\cdot, \cdot]$ indicates the concatenation operator. $T^{scene} \in R^{S \times C}$ are the refined scene tokens. $Q^{saliency} \in R^{M \times C}$ are the refined salient object queries, which can be used to generate the final salient instance masks, as stated in Eq. 5. $Q^{saliency}$ and Q^{pose} are then sent to the PDR module for rankings inference.

3.3 The Pose-Driven Ranking (PDR) Module

The PDR module is proposed to apply the pose knowledge as directional cues to help predict where human attention will shift towards, and combine both directional cues and interaction knowledge to predict the overall ranking. Figure 4 shows the design of our PDR module, which contains two parts: directional cues learning and rankings inference.

Directional cues learning models the human attention shift process by introducing directional queries $Q^{direction}$, which have M instances with D shifting directions each, and coordinate tokens $T^{coordinate}$, which have M instances with K coordinates each. We compute the likelihood of shifting by measuring the similarity between each shifting direction and coordinate. We then learn the salient features of the targets based on the shifting likelihood. This attention shift process can be formulated as:

$$\mathcal{L}^{shift} = \text{softmax}(\rho(Q^{direction}) \cdot \rho(T^{coordinate})), \tag{13}$$

$$Q^{next} = \varphi_3(\mathcal{L}^{shift} \cdot \rho(Q^{saliency})), \tag{14}$$

where ρ is the flatten and permute operation. The shifting likelihood is $\mathcal{L}^{shift} \in R^{MD \times MK}$. $Q^{next} \in R^{M \times C}$ are the queries pointing to the next region/object that human attention will shift towards.

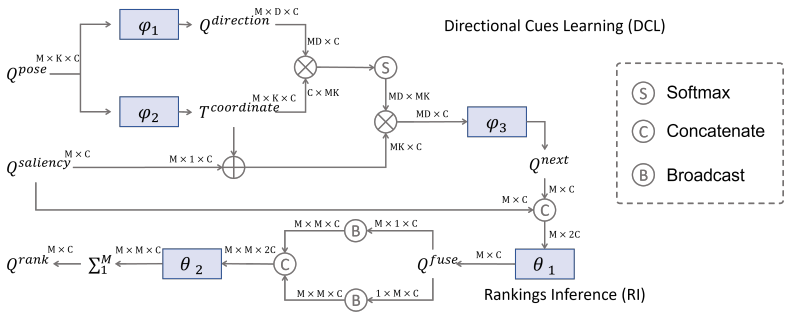

Fig. 4: Pose-Driven Ranking (PDR) Module. We transform the pose knowledge to directional queries $Q^{direction}$, and coordinate tokens $T^{coordinate}$, for modeling the attention shift process by predicting the next object/region that human attention is likely to shift towards. We then combine Q^{next} and $Q^{saliency}$ to model the inter-query ranking relations. $\varphi_1, \varphi_2, \varphi_3, \theta_1, \theta_2$ are 3-layer MLPs. $\sum_1^M$ means that we average the matrix with shape of $M \times M \times C$ along the second axis to reduce it to $M \times C$.

Rankings inference concatenates the salient object queries with Q^{next} for further modeling inter-query ranking relations:

$$Q^{fuse} = \theta_1([Q^{saliency}, Q^{next}]), \tag{15}$$

where θ_1 is a 3-layer MLP for combining interaction information and directional cues, and reducing the dimensionality to C. As ranking is related to all other queries, we model inter-query ranking relations in a pairwise manner, by concatenating every two queries and feeding them to a MLP to model their ranking relation as:

$$pairs_{i,j} = \theta_2([Q_i^{fuse}, Q_j^{fuse}]), \tag{16}$$

where $pairs \in R^{M \times M \times C}$ denotes the ranking relation between every two queries. We collect all pair relations belonging to the same instance, and average them to get the ranking-aware Q^{rank}:

$$Q_i^{rank} = \frac{1}{M} \sum_{j=1}^{j=M} pairs_{i,j}. \tag{17}$$

Finally, we predict the final ranking by treating the ranking process as a classification problem, following Fang *et al*. [12]. In addition, to facilitate the prediction of the next fixated region/object, we add a mask head on Q^{next} to force it to fixate on the next salient object instance:

$$maps_i^{next} = \sigma(\text{linear}(Q_i^{next}) \cdot feat_3), \tag{18}$$

where $maps_i^{next}$ should be matched to the $i+1$ salient instance mask, and we expect the last map of $maps^{next}$ to be the background, suggesting that humans will shift their attention to the background after visiting all salient items.

3.4 Training Strategy

PoseSOR uncovers the human pose cues for salient object ranking. We train PoseSOR in an end-to-end manner. The overall loss function is formulated as:

$$\ell_{posesor} = \ell_{masks} + \ell_{ranks} + \ell_{poses} + \ell_{auxiliary}, \tag{19}$$

where ℓ_{masks} adopts the binary cross-entropy loss and the dice loss [37] for salient instance masks and the next fixated region/object maps learning. ℓ_{ranks} adopts the cross-entropy loss for ranking prediction. ℓ_{poses} adopts ℓ_1 loss for joint coordinate regression, and employs a heatmap loss [35, 41, 42] for fast convergence. $\ell_{auxiliary}$ is used for training the Transformer decoder in Figure 2. The Transformer decoder generates N queries, each of which predicts the saliency likelihood, a bounding box, and a coarse instance mask. Following [4, 5, 47], we use the Hungarian algorithm to match these N predictions to GT or no object. Thus, the auxiliary loss can be formulated as:

$$\ell_{auxiliary} = \ell_{\hat{y}} + \ell_{bbox} + \ell_{cmask}, \tag{20}$$

where $\ell_{\hat{y}}$ adopts the binary cross-entropy loss for the likelihood prediction. ℓ_{bbox} adopts ℓ_1 loss and GIoU loss [40] for the bounding box regression. ℓ_{cmask} adopts the binary cross-entropy loss and the dice loss for the coarse masks learning.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on two popular SOR benchmarks: ASSR [44] and IRSR [31]. ASSR contains 7,646 images for training, 1,436 images for validation, and 2,418 images for testing. It is the first SOR benchmark that contains both instance-level salient object masks and saliency rankings. IRSR [31] comprises 6,059 training images and 2,929 testing images. These two benchmarks are constructed from MS-COCO [29], and the images in both benchmarks depict a variety of human scenarios and everyday objects, presenting great challenges for the SOR task. We further collect human pose annotations from MS-COCO [29] dataset for both SOR benchmarks to train our PoseSOR.

We use Segmentation-Aware SOR (SA-SOR) [31], Salient Object Ranking (SOR score) [44] and Mean Absolute Error (MAE) for evaluation, following previous works [31, 46, 47]. SA-SOR computes the Pearson correlation between the predicted rankings and human labels, and incorporates a penalty for the missed salient instances. It ranges from -1.0 to 1.0, with a positive/negative value indicating a positive/negative correlation. Higher SA-SOR scores are better. SOR is the Spearman's rank correlation between the predicted rankings and human labels. SOR scores are normalized to a range of 0.0 to 1.0, with higher values being better, without penalizing for false negatives. MAE measures the salient object segmentation quality, with lower values being better.

Table 1: Quantitative Comparison. SOD: Salient Object Detection task. SID: Salient Instance Detection task. INS: INstance Segmentation task. SOR: Salient Object Ranking task. Best results are marked in **bold** and second-best results are underlined.

Method	Year-Pub	Task	ASSR Test Set (2418)			IRSR Test Set (2929)		
			SA-SOR↑	SOR↑	MAE↓	SA-SOR↑	SOR↑	MAE↓
VST [32]	2021-ICCV	SOD	0.422	0.643	9.99	0.183	0.571	8.75
MENet [49]	2023-CVPR	SOD	0.369	0.627	9.60	0.162	0.558	8.25
S4Net [11]	2019-CVPR	SID	0.451	0.649	14.4	0.224	0.611	12.1
QueryInst [13]	2021-ICCV	INS	0.596	0.865	8.52	0.538	0.816	7.13
Mask2Former [5]	2022-CVPR	INS	0.635	0.867	7.31	0.521	0.799	7.14
RSDNet [18]	2018-CVPR	SOR	0.386	0.692	18.2	0.326	0.663	18.5
ASRNet [44]	2020-CVPR	SOR	0.590	0.770	9.39	0.346	0.681	9.44
PPA [12]	2021-ICCV	SOR	0.635	0.863	8.52	0.521	0.797	8.08
IRSR [31]	2021-TPAMI	SOR	0.650	0.854	9.73	0.543	0.815	7.79
OCOR [47]	2022-CVPR	SOR	0.541	**0.873**	10.2	0.504	**0.820**	8.45
PSR [46]	2023-ACMMM	SOR	0.644	0.815	9.59	0.454	0.752	8.07
Ours	2024	SOR	**0.673**	0.871	**7.23**	**0.568**	0.817	**6.29**

4.2 Implementation Details

We use Swin Transformer [34], pretrained on ImageNet [23], and FPN [28] as our feature extractor. We set $N = 100$, $K = 17$, $C = 256$, $L = 3$, $\tau = 0.1$, and $D = 8$. The output size of $MaxPool$ in Eq. 11 is set to 4×4, *i.e.*, 16 scene tokens. During training, we compute the mask loss on $12,544$, *i.e.*, 112×112, randomly sampled points instead of the whole mask to reduce training memory and improve training efficiency, as suggested by Cheng *et al.* [5]. We adopt the AdamW optimizer with a weight decay of 1e-4 to optimize the model for 50k iterations, with a batch size of 32. We set the initial learning rate to 5e-5 and decay it to 5e-6 after 30k iterations. We use random flip and random resize, such that each side is at least 704 and at most 800 pixels, for data augmentation. It takes roughly 20 hours to train PoseSOR using 4 A100 GPUs. During inference, we resize the input image to 768×768 and feed it to PoseSOR for prediction.

4.3 Quantitative Results

To fully evaluate our approach, we compare it with 11 other related methods, covering salient object detection methods [32, 49], salient instance detection methods [11], instance segmentation methods [5, 13] and salient object ranking methods [12, 18, 31, 44, 46, 47]. We re-train all these methods on both ASSR and IRSR benchmarks, to ensure a fair comparison. For salient object/instance detection methods, we compute their saliency rankings based on the average saliency intensity, following Islam *et al.* [18]. For instance segmentation methods, we cast the rank labels to be the class labels.

Table 1 shows the quantitative results. We can see that our approach outperforms all compared methods with nearly all metrics on both benchmarks. In

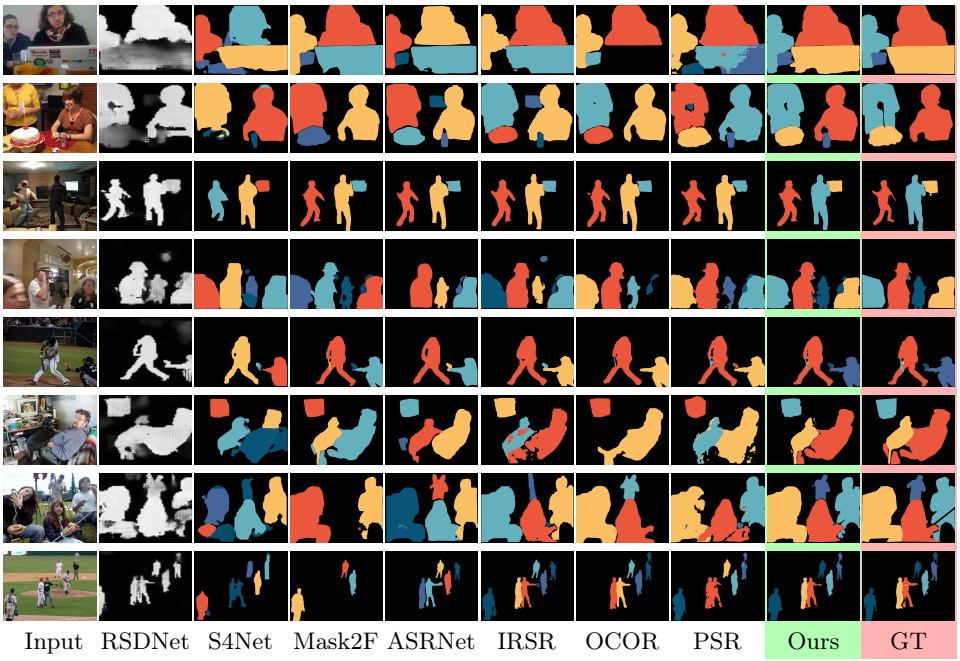

Fig. 5: Qualitative Comparison. Our method generally produces more accurate results that align with human labels (GT). Salient instances are colorized using varying color temperatures, ranging from warm to cold, to indicate the saliency ranking order.

particular, our SA-SOR score surpasses the second best by a clear margin of 3.5% on ASSR and 4.6% on IRSR benchmarks. In addition, our MAE score surpasses the second best by 11.8% on the IRSR benchmark. We also observe that the other state-of-the-art methods exhibit biases towards certain metrics. For example, while OCOR [47] performs slightly better than ours in the SOR metric, our SA-SOR and MAE scores are much better than those of OCOR. We find that OCOR often misses some salient objects. However, SOR does not penalize under-detection, while both SA-SOR and MAE do. In contrast, our method achieves consistent competitive results in all metrics on both benchmarks.

4.4 Qualitative Results

We also evaluate our method qualitatively. As shown in Figure 5, our method generally produces more accurate results that align with human labels (GT), particularly in complex scenes. For example, in the 1*st* row, our method initially focuses on the man, then shifts attention to the laptop that the man is interacting in, and finally turns to the woman and the other objects. Our predictions are in line with the GT, while the SOTA methods, *e.g.*, PSR [46], OCOR [47] and IRSR [31], tend to assign more attention to the persons and visit the laptop

Table 2: Analysis on the PAI module. ✓: human pose labels are used for training.

ID	Methods	Pose	SA-SOR↑	SOR↑	MAE↓
1	PoseSOR	✓	**0.673**	0.860	**7.35**
2	PoseSOR		0.665	0.856	7.70
3	w/o scene tokens	✓	0.670	**0.861**	7.49
4	w/o $feat^{ins}$	✓	0.669	0.858	7.35
5	w/o aggregation	✓	0.667	0.858	7.61

either at the very end or completely ignore it. In the $2nd$ row, the woman in red is looking at the cake while the woman in yellow is standing next to her. Our method correctly predicts the visiting order of this event, while the SOTA methods fail to predict reasonable saliency rankings. In the $4th$ row, the man on the phone is facing to the right while two women are passing in front of the camera. The SOTA methods, such as OCOR [47] and PSR [46], initially fixate on the man, then shift to the woman on the left, who is behind the man, and finally focus on the right women. In contrast, our method makes a more natural shift in attention from the man to the woman on the right, who is in front of the man, then turns to the woman on the left, who is in front of the camera, and finally reaches the woman in the middle, who is walking away from the camera. In addition, our method can handle complex scenarios involving a group of people, as shown in the last two rows of Figure 5.

In summary, our PoseSOR, which is equipped with the human pose-aware ability, can produce more favorable results than SOTA methods, which mainly focus on explicit visual saliency cues.

4.5 Ablation Study

We further conduct ablation experiments to evaluate the effectiveness of each component on the ASSR benchmark. We adopt a computation-friendly training setting of 512×512 resolution and a batch size of 16. Each experiment requires roughly 15 hours on an A100.

Analysis on the PAI Module. As shown in Table 2, the SA-SOR score drops from 0.673 (ID1) to 0.665 (ID2, which does not use human pose labels during training), demonstrating the advantages of incorporating human pose knowledge. We can also see that the aggregation operation (ID5), which allows the instance-level interaction stage to work in a pose-aware manner, has a more significant effect on the performance than the scene tokens (ID3) and the instance-specific features (ID4), while our full model (ID1) achieves the best performance.

Analysis on Human Quantity Impact. We further investigate whether human pose knowledge has advantages in scenarios where there are no human participants. Table 3 shows that human pose knowledge significantly enhances SOR performance when the scenarios contain one or more humans (ID7), but it does not yield benefits when no humans are involved (ID6). Additionally, we analyze

Table 3: Analysis on Human Quantity Impact. "w/ Pose" means that we train PoseSOR with pose labels. "w/o Pose" means that we do not use pose labels.

ID	Num. of Humans	Percentages	SA-SOR↑		MAE↓	
			w/ Pose	w/o Pose	w/ Pose	w/o Pose
6	0	38.5%	0.584	**0.596**	**8.84**	9.11
7	≥ 1	61.5%	**0.728**	0.709	**6.42**	6.81
8	1	21.6%	**0.722**	0.717	**6.81**	7.21
9	2	13.9%	**0.745**	0.719	**6.17**	6.73
10	≥ 3	26.1%	**0.724**	0.696	**6.24**	6.53

Table 4: Analysis on the PDR module. DCL: Directional Cues Learning part. RI: Rankings Inference part. A ✓indicates that the corresponding part is used.

ID	DCL	RI	SA-SOR↑	SOR↑	MAE↓	ID	Num.	SA-SOR↑	SOR↑	MAE↓
11			0.666	0.856	**7.34**	14	$D = 8$	**0.673**	**0.860**	**7.35**
12	✓		0.661	**0.861**	7.46	15	$D = 4$	0.673	0.860	7.44
13	✓	✓	**0.673**	0.860	7.35	16	$D = 1$	0.670	0.852	7.54

the performance gains in situations involving different numbers of human participants (ID8-ID10). In summary, incorporating pose knowledge generally leads to better performance, especially when humans are present, and the performance gains are more pronounced in the scenarios with multiple humans (ID9-ID10) compared to the scenes containing only one person (ID8).

Analysis on the PDR Module. The PDR module contains two parts, *i.e.*, directional cues learning and rankings inference. Table 4 (ID11-ID13) shows that enabling both parts can lead to more competitive results on all metrics. We suppose that a robust prediction favors both the awareness of the upcoming attention shift and the modeling of pairwise ranking relations. We also study the impact of the different number of shifting directions (ID14-ID16). Results suggest that a larger number of shifting directions leads to better performance.

5 Conclusion

In this work, we have proposed to explore human pose cues to acquire a deeper understanding of high-level interactions between humans and their surroundings for the SOR task. Extensive qualitative and quantitative results demonstrate the superiority of our method, and the advantages of incorporating human pose knowledge, especially in scenarios where multiple humans are present. Our method does present some limitations. For example, our PoseSOR can only detect coarse pose predictions due to the limited pose data during training. This may hinder the integration of pose knowledge for the SOR task. We aim to address this issue in the future by transferring pose knowledge from existing human pose models or by training it with more pose data.

Acknowledgements

This work is in part supported by a GRF grant from the Research Grants Council of Hong Kong (RGC Ref.: 11205620).

References

1. Aberman, K., He, J., Gandelsman, Y., Mosseri, I., Jacobs, D.E., Kohlhoff, K., Pritch, Y., Rubinstein, M.: Deep saliency prior for reducing visual distraction. In: CVPR (2022)
2. Achanta, R., Hemami, S.S., Estrada, F.J., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR (2009)
3. Bardi, L., Di Giorgio, E., Lunghi, M., Troje, N.F., Simion, F.: Walking direction triggers visuo-spatial orienting in 6-month-old infants and adults: An eye tracking study. Cognition **141**, 112–120 (2015)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022)
6. Cheng, M., Warrell, J., Lin, W., Zheng, S., Vineet, V., Crook, N.T.: Efficient salient region detection with soft image abstraction. In: ICCV (2013)
7. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual review of neuroscience **18**(1), 193–222 (1995)
8. Driver IV, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., Baron-Cohen, S.: Gaze perception triggers reflexive visuospatial orienting. Visual cognition **6**(5), 509–540 (1999)
9. Du, L., Li, L., Wei, D., Mao, J.: Saliency-guided single shot multibox detector for target detection in sar images. IEEE Transactions on Geoscience and Remote Sensing **58**(5), 3366–3376 (2019)
10. Fan, L., Wang, W., Huang, S., Tang, X., Zhu, S.C.: Understanding human gaze communication by spatio-temporal graph reasoning. In: ICCV (2019)
11. Fan, R., Cheng, M.M., Hou, Q., Mu, T.J., Wang, J., Hu, S.M.: S4net: Single stage salient-instance segmentation. In: CVPR (2019)
12. Fang, H., Zhang, D., Zhang, Y., Chen, M., Li, J., Hu, Y., Cai, D., He, X.: Salient object ranking with position-preserved attention. In: ICCV (2021)
13. Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Instances as queries. In: ICCV (2021)
14. Gervais, W.M., Reed, C.L., Beall, P.M., Roberts, R.J.: Implied body action directs spatial attention. Attention, Perception, & Psychophysics **72**, 1437–1443 (2010)
15. Guan, H., Lau, R.W.: Seqrank: Sequential ranking of salient objects. Proceedings of the AAAI Conference on Artificial Intelligence **38**(3), 1941–1949 (2024)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
17. He, S., Jiao, J., Zhang, X., Han, G., Lau, R.W.: Delving into salient object subitizing and detection. In: ICCV (2017)
18. Islam, M.A., Kalash, M., Bruce, N.D.: Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In: CVPR (2018)
19. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research **40**(10-12), 1489–1506 (2000)

20. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.: Saliency detection via absorbing markov chain. In: ICCV (2013)
21. Johnston, W.A., Dark, V.J.: Selective attention. Annual review of psychology **37**(1), 43–75 (1986)
22. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human neurobiology **4**(4), 219–227 (1985)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
24. Langton, S.R., Bruce, V.: You must see the point: automatic processing of cues to the direction of social attention. Journal of Experimental Psychology: Human Perception and Performance **26**(2), 747 (2000)
25. Li, G., Xie, Y., Lin, L., Yu, Y.: Instance-level salient object segmentation. In: CVPR (2017)
26. Li, H., Zhang, D., Liu, N., Cheng, L., Dai, Y., Zhang, C., Wang, X., Han, J.: Boosting low-data instance segmentation by unsupervised pre-training with saliency prompt. In: CVPR (2023)
27. Li, X., Lu, H., Zhang, L., Ruan, X., Yang, M.: Saliency detection via dense and sparse reconstruction. In: ICCV (2013)
28. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
30. Liu, N., Han, J.: A deep spatial contextual long-term recurrent convolutional network for saliency detection. IEEE Transactions on Image Processing **27**(7), 3264–3274 (2018)
31. Liu, N., Li, L., Zhao, W., Han, J., Shao, L.: Instance-level relative saliency ranking with graph reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8321–8337 (2021)
32. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: ICCV (2021)
33. Liu, N., Zhao, W., Shao, L., Han, J.: Scg: Saliency and contour guided salient instance segmentation. IEEE Transactions on Image Processing **30**, 5862–5874 (2021)
34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
35. Mao, W., Tian, Z., Wang, X., Shen, C.: Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In: CVPR (2021)
36. Miangoleh, S.M.H., Bylinskii, Z., Kee, E., Shechtman, E., Aksoy, Y.: Realistic saliency guided image enhancement. In: CVPR (2023)
37. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
38. Neisser, U.: Cognitive psychology: Classic edition. Psychology press (2014)
39. Pei, J., Cheng, T., Tang, H., Chen, C.: Transformer-based efficient salient instance segmentation networks with orientative query. IEEE Transactions on Multimedia (2022)
40. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression (2019)

41. Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: CVPR (2022)
42. Shi, D., Wei, X., Yu, X., Tan, W., Ren, Y., Pu, S.: Inspose: instance-aware networks for single-stage multi-person pose estimation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3079–3087 (2021)
43. Shi, J., Weng, X., He, S., Jiang, Y.: Biological motion cues trigger reflexive attentional orienting. Cognition **117**(3), 348–354 (2010)
44. Siris, A., Jiao, J., Tam, G.K., Xie, X., Lau, R.W.: Inferring attention shift ranks of objects for image saliency. In: CVPR (2020)
45. Siris, A., Jiao, J., Tam, G.K., Xie, X., Lau, R.W.: Scene context-aware salient object detection. In: ICCV (2021)
46. Sun, C., Xu, Y., Pei, J., Fang, H., Tang, H.: Partitioned saliency ranking with dense pyramid transformers. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1874–1883 (2023)
47. Tian, X., Xu, K., Yang, X., Du, L., Yin, B., Lau, R.W.: Bi-directional object-context prioritization learning for saliency ranking. In: CVPR (2022)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
49. Wang, Y., Wang, R., Fan, X., Wang, T., He, X.: Pixels, regions, and objects: Multiple enhancement for salient object detection. In: CVPR (2023)
50. Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., Tian, Q.: Label decoupling framework for salient object detection. In: CVPR (2020)
51. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: ECCV (2012)
52. Wu, Y.H., Liu, Y., Zhang, L., Cheng, M.M., Ren, B.: Edn: Salient object detection via extremely-downsampled network. IEEE Transactions on Image Processing **31**, 3125–3136 (2022)
53. Wu, Y.H., Liu, Y., Zhang, L., Gao, W., Cheng, M.M.: Regularized densely-connected pyramid network for salient instance segmentation. IEEE Transactions on Image Processing **30**, 3897–3907 (2021)
54. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR (2013)
55. Zhang, L., Wu, J., Wang, T., Borji, A., Wei, G., Lu, H.: A multistage refinement network for salient object detection. IEEE Transactions on Image Processing **29**, 3534–3545 (2020)
56. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: Capsal: Leveraging captioning to boost semantics for salient object detection. In: CVPR (2019)
57. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: CVPR (2014)