Fine-Grained Perception in Panoramic Scenes: A Novel Task, Dataset, and Method for Object Importance Ranking

Jia Song[1,2,3], **Chenglizhao Chen**[1,2,3*], **Xu Yu**[1,2,3], **Shanchen Pang**[1,2,3]

[1]Qingdao Institute of Software, China University of Petroleum (East China)
[2]College of Computer Science and Technology, China University of Petroleum (East China)
[3]Shandong Key Laboratory of Intelligent Oil & Gas Industry Software

Abstract

Existing Salient Object Ranking (SOR) aims to infer ranking of salient objects based on their saliency degree. However, it tends to only focus on salient objects while neglecting non-salient ones. This coarse-grained ranking limits the performance of downstream tasks. For instance, in image retrieval tasks, focusing solely on the relationship between salient objects is insufficient for achieving fine-grained scene analysis, which may result in retrieved results that do not satisfy user requirements. High-quality retrieval requires fine-grained analysis, making it essential to rank non-salient objects. Based on this need, we propose a new task: Fine-grained Object Importance Ranking in 360° Scenes (FOIR-360), which focus on predicting the relative importance of "ALL objects" at the instance-level. Our task takes into account all objects, allowing us to refine the original "coarse-grained" to a "fine-grained" level. Currently, the main challenge for this new task is the lack of supervised data for model training or even for model testing. Therefore, we propose a novel weakly supervised method to address the shortage of datasets. Furthermore, to the best of our knowledge, there is no existing suitable annotation protocol for this new task. The main reason is that annotating fine-grained rankings is extremely difficult, especially in panoramic scenes that contain numerous instances where even humans are unable to determine which one is more important than others. As the first attempt, we introduce a new annotation protocol designed to highlight the ranking of objects that are non-salient yet still important. Based on this protocol, we construct the first fine-grained 360Rank dataset. In summary, all these new task, weakly supervised method, annotation protocol, and dataset have the potential to drive advancements in the field.

Code — https://github.com/noname965/FOIR-360

Introduction

Compared to conventional 2D images, 360° images (a.k.a. panoramic images) offer a more comprehensive perspective by capturing a full range of views (Duan et al. 2024), which is critical for various real-world applications (Kiran et al. 2021; DeSouza and Kak 2002). Panoramic images typically comprise wider scenes and more objects, posing challenges in processing the entire image efficiently. Existing tools (see Fig. 1a), *e.g.*, fixation prediction (FP) (Xie et al. 2024), salient object detection (SOD) (Ma et al. 2020) and salient object ranking (SOR) (Qiao et al. 2024) can localize which objects are worthy of more attention in the panoramic images, thereby optimizing computational resources. Unfortunately, these state-of-the-art (SOTA) tools can only localize the most salient objects while ignoring those less-

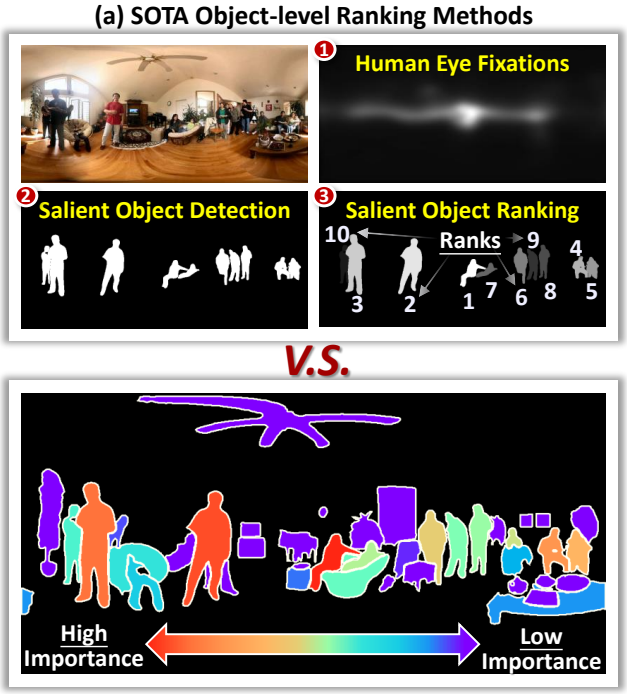

Figure 1: Motivation demonstration of our new task. Current SOTA object-level ranking methods (a) often rank the most salient objects while overlooking those less-salient ones, limiting the performance of downstream tasks (*e.g.*, image retrieval, see details in the Fig. 2). Therefore, we introduce a novel task, FOIR-360 (b), which is more fine-grained than SOR, as it places no limits on the number of important objects, thereby aligning more closely with real human visual.

*Corresponding Author: Chenglizhao Chen, cclz123@163.com

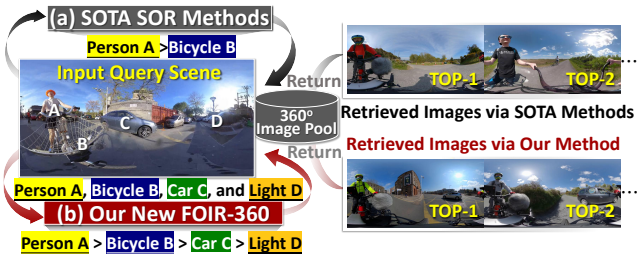

Figure 2: Application of importance ranking to image retrieval. (a) The existing SOTA methods ignore non-salient objects (*e.g.*, car and street light), leading to retrieved images that fail to satisfy the user requirements. In contrast, our approach (b) can return more fine-grained matching results, better fulfilling user needs.

salient ones. However, in practical applications, these objects, which are ignored by existing methods, often hold significant importance.

For instance, as shown in Fig. 2, image retrieval is defined as searching for semantic similar target images in a large pool, given a query image. For similarity matching, the most important is to measure the similarity of both global and local information. However, not all objects hold equal importance, the existing SOTA methods (Fig. 2a) use a SOR method to guide the matching process, *e.g.*, they first determine which objects are salient in the current scene, and then return the matched images to the user based on the results of salient objects similarity metric. However, those SOTA methods only focus on the salient objects while neglecting non-salient ones. This coarse-grained ranking leads to retrieved images lack key details, failing to satisfy user requirements. Therefore, we propose a new task, namely **f**ine-grained **o**bject **i**mportance **r**anking in 360° scenes (FOIR-360), which focus on predicting the relative importance of all objects. By applying our new FOIR-360 method (see Fig. 2b), we can capture more detailed vision information, which substantially improve the performance of image retrieval. It is noteworthy that there exists a clear distinction between the term "saliency" and "importance"[1].

FOIR-360, as a new task, the main challenge is the lack of supervised data. Specially, large-scale FOIR-360 supervised datasets are difficult to obtain due to the annotation difficulty[2] and time-consuming. To address the above challenges, we propose a novel weakly-supervised approach that reduces the dependency on large-scale training data. This new method can automatically generate high-quality importance ranking pseudo-labels. Utilizing those pseudo-labels, a general FOIR network can be trained to achieve high precision and efficient inference.

[1]The former primarily refers to visual stimuli, closely consistent with the human eye fixations that can be recorded by eye tracker. The latter usually reflects to the 2nd thoughts, i.e., an object might be visually salient, but, after a deep thinking, we could have a conclusion that this object is not important. Hence, compared to the term "saliency", the term "importance" bias more towards the inter-object comparisons in semantics. So, an object could be very important yet staying low saliency, and vice versa.

[2]When numerous instances belonging to the same category are densely clustered together, humans also struggle to determine which instance is more important.

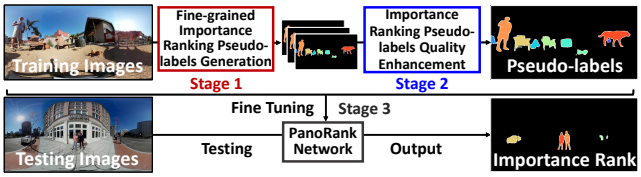

Figure 3: The overall pipeline of our proposed weakly-supervised approach.

Furthermore, this new task lacks a unified, standardized, and controllable method for dataset annotation. As the first attempt, we propose a new annotation protocol designed to highlight the ranking of salient and non-salient objects. Based on this annotation protocol, we have constructed the first fine-grained importance ranking dataset, 360Rank. Although the dataset is relatively small (only 500 images), it still serves as a test set to supports the development of the field. In summary, our main contributions are:

- We introduce a novel task, FOIR-360, aims at predicting the relative importance of all identifiable objects, which can benefit downstream tasks.
- We propose a novel weakly-supervised approach for the FOIR-360 task, effectively addressing challenges of data shortage and ranking ambiguity. Alongside this, we introduce a naive network, coined as PanoRank, to accelerate this process.
- For this new task, we introduce a new annotation protocol tailored for panoramic data. Built upon this annotation protocol, we have constructed a new 360Rank dataset.

Related Works

Fixation Prediction (FP). FP is intended to predict the focal points of viewers' attention. For 360° scenes, to address distortion problems, multi-projection fusion methods (Chao et al. 2018, 2020; Zou et al. 2023) and tailored network architectures (Esteves et al. 2018; Jiang et al. 2019; Cokelek et al. 2023) have been proposed. These methods effectively assess the relative importance of different areas but fail to segment salient objects with clear boundaries.

Salient Object Ranking (SOR). SOR aims to to segment and rank salient objects according to their degree of saliency. Islam *et al*. (Islam, Kalash, and Bruce 2018) first introduced the concept of relative saliency. Siris *et al*. (Siris et al. 2020) introduced the ASSR dataset, which defines saliency ranking based on the sequence of human attention shift. Liu *et al*. (Liu et al. 2021) constructed another IRSR dataset. However, these datasets that directly convert saliency maps into ranking labels tend to be unreliable. Because fixation data is collected through bottom-up, unconscious behavior, while SOR is a top-down, conscious process. Instead of using human eye-tracking data, our new dataset aims to simulate actual human visual importance perception by collecting raw ranking data based on in-depth thinking.

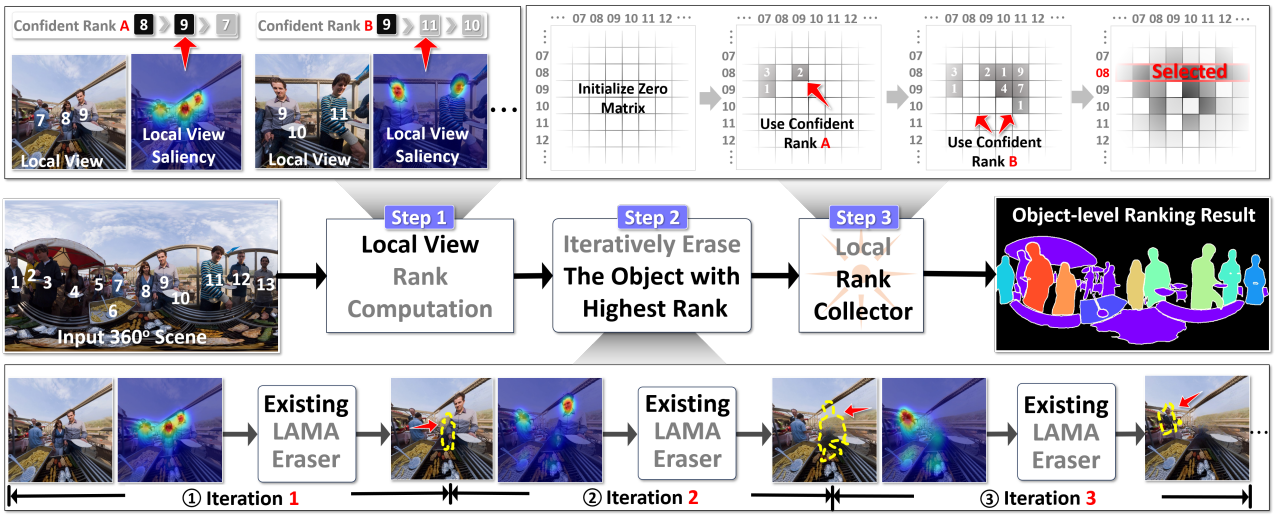

Figure 4: The overall pipeline of our proposed fine-grained importance ranking pseudo-labels generation. It consists of three main steps: 1) local view rank computation, 2) iteratively erase the object with the highest rank, and 3) local rank collector. Moving from left to right, the goal of Step1 is to compute the importance ranking of each instance within a local view. Step2 aims to sequentially remove the most important instance determined in the Step1. The Step3 then aggregates the ranks from all local views to form a comprehensive ranking.

Method

To address FOIR-360 data shortage problem, we propose a weakly-supervised approach. The core idea is to automatically generate high-quality importance ranking pseudo-labels. Based on these pseudo-labels, we train a novel fine-grained ranking model, which will eventually achieve fine-grained importance ranking. To achieve this, our method consists of three stages, the overview can be found in Fig. 3. Stage1 focuses on generating initial pseudo-labels, Stage2 improves their quality by aggregating multiple pseudo-labels, and in Stage3, PanoRank network is trained using these high-quality pseudo-labels. Notice that the first stage is the key technical innovations of this paper.

Stage 1: Fine-grained Importance Ranking Pseudo-labels Generation

Given a panoramic image, this stage aims to automatically generate fine-grained importance ranking pseudo-labels. To achieve it, we use off-the-shelf 2D saliency models to automatically assign importance values to each object.

Existing methods (Lin, Guan, and Lau 2022) generally use the saliency values within each object to infer their importance ranking. But this naive direct conversion from **s**aliency to **r**anking (Sal-to-Rank) has multiple limitations.

Firstly, current SOTA 2D saliency models cannot perform well in ERP-format panoramic images due to inherent visual distortion. **Secondly**, existing SOTA 2D saliency prediction results is not always reliable. Because they are primarily designed for sparse scenes with few objects, whereas the target task FOIR-360 is a relatively dense task. This differentiation leads to current saliency models failing to comprehensively cover all objects in panoramic scenes.

To address these issues, we propose a novel method that include a novel divide-and-conquer strategy and a new object erasing method. As shown in Fig. 4, this stage consists of three steps. The key idea is to divide the original problem into local rankings under multiple 2D local view (Step1: local view rank computation). By fully leveraging existing 2D saliency models, we employ an iterative erasing strategy to gradually transform **unreliable saliency evaluations** into **reliable instance-level importance ranking** (Step2: iteratively erase the object with the highest rank). Finally, we conquer the importance rankings from each local-view to achieve global importance ranking (Step3: local rank collector). This divide-and-conquer strategy is not sensitive to the visual distortions in panoramic images.

Local View Rank Computation. As mentioned above, ERP images suffer from heavy vision distortion, which may impair the semantic information embedded in 2D saliency maps. To optimally leverage the semantic information, rank computation must be performed within a 2D local view. To achieve this, we first perform a spherical-to-plane projection. Conventional projection methods, such as cubemap projection (CMP) and tangent projection (TP), often divide a single complete object into multiple parts. Therefore, in order to preserve the integrity and continuity of the object, we propose an object-centered projection strategy. Given an ERP image $\mathbf{I} \in \mathbb{R}^{H \times W}$, object centered-based projection can be represented as:

$$\text{LocalView}_i = \text{Proj}(x_i, y_i, \mathbf{I}), \tag{1}$$

where LocalView_i represents distortion-free local view (see Fig. 4 Step1: Local View). Proj represents rectilinear perspective projection. (x_i, y_i) represent the coordinates i-th

object's center points,and it can be formulated as Eq. 2.

$$\{(x_t, y_t)\}_{t=1}^{n} \leftarrow \mathcal{F}_{mask \to cp}(\underbrace{\underline{\mathbf{OM}}}_{\overbrace{\{\mathrm{OM}_1, \mathrm{OM}_2, \ldots, \mathrm{OM}_n\} = \mathrm{OS}(\mathbf{I})}^{\Uparrow}}), \tag{2}$$

where (x_t, y_t) denotes the center coordinates of the t-th object, n represents the total number of objects, and $\mathcal{F}$ denotes the mapping function from the mask to the center point. OM represents n objects returned by a segmentation tool, here, we employ MaskDino (Li et al. 2023).

Based on the above process, we can obtain a series of overlapping local views that are distortion-free. To determine the local rank, we perform straightforward computation using pixel-wise saliency predictions (see Fig. 4 Step1: Local View Saliency). Taking the local view image as inputs, the ranking scores are calculated as follows:

$$\begin{aligned} &\underbrace{\mathrm{Salmap} \leftarrow \mathrm{2DSal}(\mathrm{LocalView})}_{\Downarrow} \\ &\mathrm{IMSore}_i = \mathrm{Sum}(\mathrm{NewObj}_i \odot \xi(\overline{\mathrm{Salmap}})), \end{aligned} \tag{3}$$

where $\mathrm{IMScore}_i$ represents the importance score for the i-th object mask. Sum computes the sum of pixels within the mask. NewObj_i denotes the processed object mask for the i-th object. ξ is the min-max normalization function, and $\mathrm{2DSal}$ represents existing SOTA 2D saliency method. The operator $\odot$ denotes element-wise multiplication.

Iteratively Erase the Object with the Highest Rank. Based on Eq. 3, local-view rankings can be obtained through direct Sal-to-Rank conversation. However, these local-view rankings are not always reliable. *e.g.*, we find that the most salient object tend to be regarded as the most important. Yet, the correlation between saliency and importance gradually diminishes as the ranking decreases. This motivates us to focus on the most important object by selecting only the Top-1 mask. To obtain the importance ranking of all objects, we employ an iterative erasure strategy that gradually determines their ranking. That is, starting with the most important object, our method prompts the saliency model to sequentially determine new important objects by erasing the instances whose "importance degree" have been determined already in the previous iterations.

The iteratively erasing process is shown in Fig. 4 Step2, we first identify the most discriminative important object, such as "the woman in the middle" (see Fig. 4 Step2-①). Subsequently, by erasing this object and recalculating saliency assessment, we can reveal the next important object, such as "the man" as shown in Fig. 4 Step2-②. This process is repeated until all objects have been evaluated. The iterative process is time-consuming due to the large number of objects in the panoramic image. To reduce the computational cost, we attempt to simultaneously erasure all objects and then stitch them together, the implementation process is described below:

$$\mathrm{NewView} = \mathrm{Merge}(\underbrace{\underline{\mathrm{ViewNU}}}_{\overbrace{\mathrm{ViewNU} = \mathrm{Eraser}(\mathrm{View}, \mathcal{M})}^{\Uparrow}}, \mathrm{View} \odot (\mathcal{M} \setminus m_{top})), \tag{4}$$

where $\mathrm{NewView}$ represents the new image after erasing the Top-1 object, Merge denotes the operation of mapping the remaining objects back into the background window, ViewNU is the null background window after erasing all objects, View represents a local sub-window, $\mathcal{M}$ represents the complete mask of all objects, and m_{top} represents the selected Top-1 mask. Eraser denotes any image inpainting tool, and here we use LAMA (Suvorov et al. 2022). The operator $\odot$ denotes element-wise multiplication.

Furthermore, before performing the object erasing, we apply morphological dilation (for different expansion rates (ER) , see Table 4) to expand the edges of the segmentation mask, allowing the mask to better capture the object.

Algorithm 1: Local Rank Collector Algorithm

Input: Local rank list set $\mathcal{S}$, Number of objects N
Output: Global rank list

```
Initialize matrix C as an N × N zero matrix.
for O ∈ S do
    Calculate ranking weights W by Eq. 5.
    Convert the local rank list O and the weight vector
    w ∈ W to a set of tuples R(i, j, w).
    C ← UpdateMatrix(C, R).
end for
Complete matrix C via relation transmission.
Convert pairwise matrix C to global rank list
return global rank list L ∈ R^{1×N}
```

Local Rank Collector. After the aforementioned calculations, we can only obtain the local importance rankings within each local-view. These rankings are localized and not well-integrated, which makes it challenging to combine them into a global ranking. We utilize the overlapping regions between local-view as "transfer bridges" to globally propagate their corresponding instance-level importance rankings. As shown in Fig. 4 Step3, We first transform local rankings into pairwise importance rankings between objects. Then, through a voting mechanism, we calculate the win rate of each object in these pairwise rankings, serving as the basis for the global importance ranking. The implementation details can be found in Algorithm 1.

Stage 2: Importance Ranking Pseudo-labels Quality Enhancement

In the previous section, we have obtained fine-grained importance ranking pseudo-labels using single saliency model. However, this single label may introduce prejudice when evaluating the importance of objects, leading to lower accuracy. The saliency maps produced by different 2D saliency models potentially containing complementary information. To improve pseudo-label quality and avoid the bias of single label, we use the differences in importance values between objects as confidence weights to adaptively aggregate multiple pseudo-labels. The fusion process is as follows:

$$\mathcal{S} = \sum \underbrace{\underline{\omega_i(m_1, m_2)}}_{\overbrace{\omega(m_1, m_2) = |\mathrm{Sal}(\mathbf{I}, m_1) - \mathrm{Sal}(\mathbf{I}, m_2)|}^{\Uparrow}} \odot \mathrm{Sal}_i, \tag{5}$$

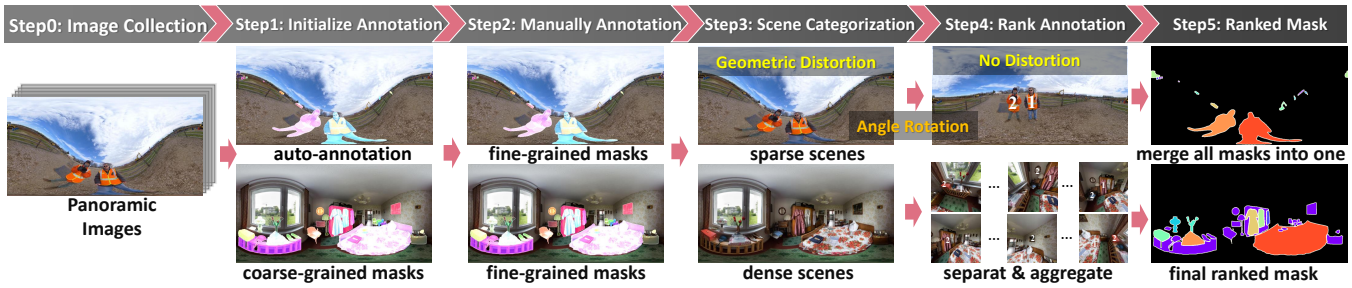

Figure 5: Overview of the importance ranking annotation process.

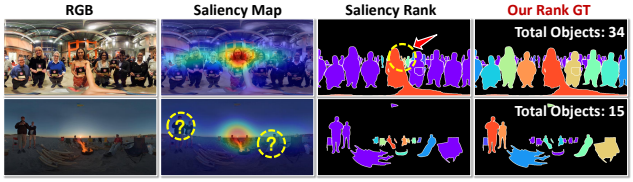

Figure 6: Comparison of saliency maps-based saliency rank *v.s.* our new importance rank GT.

where $\mathcal{S}$ represents the final saliency map, $\omega(m_1, m_2)$ denotes saliency-related weights, Sal_i denotes different saliency models, here, we utilize three saliency models, *i.e*, TranSalNet, RInet, and GSGNet. $\mathbf{I}$ denotes ERP images, m_1 and m_2 represent the Top-1 and Top-2 ranked object masks, respectively. $\odot$ denotes element-wise multiplication.

Through the above two stages, we can automatically generate high-quality training data pairs, which have the potential to enhance the performance of SOTA ranking methods.

Stage 3: PanoRank Network Fine Tuning

Our weakly-supervised method for generating fine-grained importance ranking pseudo-labels is offline and time-consuming. To accelerate the process, we propose a new PanoRank network, which is trained using those high-quality pseudo-labels to achieve end-to-end fast predictions. Notice that network is not the main focus of this paper, PanoRank is utilized as a naive implementation to accelerate our new FOIR-360 task.

We utilize Mask2former (Cheng et al. 2022) as the foundational architecture of the PanoRank, we replace the encoder with PanoSwin (Ling et al. 2023) to extract distortion-free panoramic features. In addition, we introduce an extra ranking branch specifically designed to compute importance ranking independently. Following SeqRank (Guan and Lau 2024), we design an REM model, which sequentially predicts the current most important object by masking objects that have already been visited. Additional details of the SeqRank network can be found in the *supplementary material*.

360Rank Dataset

To validate the effectiveness of our weakly-supervised method, we introduce a new annotation protocol to guide the annotation process for fine-grained importance ranking,

Figure 7: The category statistics of our 360Rank dataset, which consists of 13 superclasses and 148 subclasses.

detailed in Fig. 5. Based on this novel annotation protocol, we have established a new fine-grained importance ranking dataset (360Rank). Fig 7 shows detailed category statistics.

Image Collection and Mask Annotation

Image Collection. Due to data copyright restrictions, we decided not to collect new data. Instead, we directly combine four existing datasets, (*i.e.*, PANDORA (Xu et al. 2022), AOI (Xu et al. 2021), ODI-SOD (Wu et al. 2023), and PanopticVideo (Wang et al. 2024)) as our raw data sources. To ensure data diversity, we select 500 panoramic images with indoor and outdoor scenes. **Instance-level Mask Annotation.** Panoramic images often contain numerous small and complex-shaped instances, making manually annotating all segmentation masks are time-consuming and labor-intensive. To accelerate this process, we employ a semi-automatic approach for instance-level mask annotation. We first utilized the existing instance segmentation model MaskDino (Li et al. 2023) to perform automatic segmentation (see Fig.5 Step1). Then, we employ SAM (Kirillov et al. 2023) to manually refine or supplement the instances that were not accurately segmented in the Step1 (see Fig.5 Step2). Finally, 7,000 instance-level segmentation masks were selected for subsequent ranking annotation.

		2D SOR					2D Saliency							360 Saliency		
	Metric	IRSR [2021]	OCOR [2022]	PSR [2023]	Seqrank [2024]	QAGNet [2024]	CaslNet [2018]	Unisal [2020]	Emalnet [2020]	TranSal [2022]	TempSal [2023]	RInet [2023]	GSGNet [2024]	ATSal [2021]	EPS [2023]	OURS
Our 360Rank	SA-SOR	0.566	0.503	0.578	0.594	0.650	0.612	0.615	0.612	0.613	0.608	0.616	0.616	0.568	0.616	**0.658**
	SOR	0.664	0.782	0.736	0.762	0.778	0.857	0.864	0.856	0.857	0.863	0.862	0.865	0.816	0.859	**0.877**
	MAE	0.067	0.061	0.066	0.056	0.056	0.045	0.044	0.044	0.045	0.044	0.044	0.044	0.047	0.046	**0.041**
ASSR	SA-SOR	0.625	0.572	0.735	0.692	**0.779**	0.749	0.758	0.757	0.764	0.765	0.765	0.767	--	--	0.770
	SOR	0.734	0.780	0.824	0.824	0.836	0.886	0.883	0.876	0.885	0.884	0.882	**0.888**	--	--	0.887
	MAE	0.091	0.082	0.074	**0.073**	0.049	0.147	0.140	0.136	0.139	0.139	0.141	0.140	--	--	0.139
SalMon	SA-SOR	0.439	0.318	**0.538**	0.391	0.419	0.533	0.524	0.521	0.524	0.524	0.530	0.534	--	--	**0.538**
	SOR	0.708	0.700	0.780	0.704	0.679	**0.819**	0.811	0.812	0.814	0.799	0.814	0.818	--	--	**0.819**
	MAE	0.053	0.051	0.045	**0.044**	0.038	0.046	0.046	0.046	0.044	0.045	0.046	0.045	--	--	0.045

Table 1: Quantitative comparisons between our method with other SOTA models on our newly proposed 360Rank dataset and the 2D SOR datasets. We have included the corresponding qualitative comparisons in the *supplementary material*.

Annotation Protocol

In panoramic scenes, saliency maps-based methods are commonly used for ranking annotation. These methods determine saliency ranking by calculating the average or maximum pixel values of saliency within the mask. However, 360° saliency map commonly suffers from a central bias issue (*e.g.*, see yellow circle in the first row of Fig. 6). A large number of objects in the "equator region" are assigned higher importance, making it difficult to effectively distinguish the levels of importance among these instances. Besides, 360° saliency maps fail to cover all objects, leading to inaccurate ranking, especially for those objects with lower ranks (as shown in the second row of Fig. 6).

Furthermore, as we have mentioned before, there is a significant gap between saliency values and importance rankings. Therefore, instead of using human eye fixation data, we manually annotate ranking based on the sequence of mouse clicks, reflecting in-depth thinking. ERP images often cause visual distortion, making objects appear unnaturally stretched and enlarged, leading to misjudgments of their importance. To minimize the impact of visual distortions, we attempt to project the complete ERP image into multiple sub-viewports for separate annotation. Since object distribution varies across scenes, we classify scenes into sparse and dense categories based on the following formula:

$$\mathrm{Dens} = \mathrm{M} * \sum_{i=1}^{\mathrm{n}} (1/(\mathrm{CenDis}_i + \mathrm{ObjSize}_i)), \qquad (6)$$

where Dens represents the density score of an image, n represents the total number of masks. CenDis_i denotes the average distance from the centers of the other masks to the center of the i-th mask, $\mathrm{ObjSize}_i$ represents the pixel sum of the i-th mask, and M is a constant of 10,000. * denotes multiplication between the variables. $\mathrm{Dens}>10$ is categorized as a dense scenes (as shown in Fig. 5 Step3).

Sparse Data Ranking Annotation. For sparse scenes, we directly labeled importance ranking on the ERP image. To address distortion issues, we adjust the panoramic images to a more optimal ERP projection angle (see Fig. 5 Step4). **Dense Data Ranking Annotation.** Fig. 5 Step4 shows the dense data annotation process. First, the ERP image is divided into multiple sub-windows. Then, the importance ranking of all objects within each sub-window is annotated. After collecting importance rankings from all sub-windows, Algorithm 1 is used to aggregate local rankings into a global ranking.

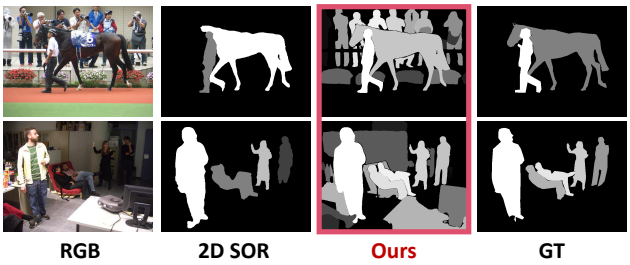

Figure 8: Visual comparisons between our approach *v.s.* the 2D SOR method (SenqRank). Our method achieves more fine-grained ranking results while maintaining accuracy.

Experiments

Experimental Setup

Evaluation Metrics. We utilize widely-used evaluation metrics for fair comparison, namely salient object ranking (SOR) (Siris et al. 2020), segmentation-aware SOR (SA-SOR) (Liu et al. 2021), and mean absolute error (MAE) (Zhang et al. 2017). **Implementation details.** Our implementation is based on MMDetection (Chen et al. 1906). We conduct our experiments on NVIDIA GTX 3090 GPU, and the batch size is set to 4.

Comparison With SOTA Methods

Evaluation on our 360Rank Dataset. Comparison with SOR Methods. We compare the performance of our method with 5 SOTA 2D SOR methods on our 360Rank dataset. These methods including IRSR (Liu et al. 2021), OCOR (Tian et al. 2022), PSR (Sun et al. 2023), Seqrank (Guan and Lau 2024), and QAGNet (Deng et al. 2024). The quatitative comparison results in Table 1 demonstrate that our method outperforms other SOTA models. Specifically, compared with the QAGNet approach, our method achieve 11.5% improvements in SOR, This is because 2D SOR models fail to precisely predict the impor-

NO.	360Rank					Metrics		
	LVRC	LIE	LRC	PLE	PR	SA-SOR	SOR	MAE
①	✓					0.597	0.848	0.048
②	✓	✓				0.603	0.860	0.044
③	✓	✓	✓			0.628	0.873	0.043
④	✓	✓	✓	✓		0.646	0.865	0.044
⑤	✓	✓	✓	✓	✓	**0.658**	**0.877**	**0.041**

Table 2: Ablation study on each component. **LVRC**: local view rank computation; **LIE**: iteratively erase the object with the highest rank; **LRC**: local rank collector; **PLE**: importance ranking pseudo-labels quality enhancement; **PR**: PanoRank network.

	Metrics	IRSR	IRSR*	PSR	PSR*	Seqrank	Seqrank*
2D SOR	SA-SOR	0.566	0.591	0.578	0.605	0.594	0.621
	SOR	0.663	0.722	0.736	0.761	0.762	0.769
	MAE	0.067	0.065	0.066	0.066	0.056	0.055
	Metrics	TranSal	TranSal*	ATSal	ATSal*	GSGNet	GSGNet*
Saliency	SA-SOR	0.613	0.626	0.613	0.626	0.616	0.630
	SOR	0.856	0.856	0.858	0.865	0.863	0.867
	MAE	0.045	0.044	0.045	0.044	0.044	0.044

Table 3: SOTA methods *v.s.* their improved results using our iterative erasing method that are highlighted in *.

tance ranking of all objects, especially those ranked lower. In contrast, our method can rank all objects, achieving fine-grained importance ranking.

Comparison with Saliency Prediction Methods. We compare our method with 9 saliency prediction methods, *i.e.*, CasNet (Fan et al. 2018), Unisal (Droste, Jiao, and Noble 2020), Emalnet (Jia and Bruce 2020), TranSalNet (Lou et al. 2022), TempSal (Aydemir et al. 2023), RInet (Song et al. 2023), GSGNet (Xie et al. 2024), EPS (Zou et al. 2023), and ATSal (Dahou et al. 2021). For saliency models, we calculate the total number of salient values within each instance to obtain the importance ranking. As shown in Table 1, our method also achieves the best performance, further demonstrating robustness of our method.

Evaluation on 2D SOR Datasets. To evaluate the generalization ability of our method, we further compared our iterative erasing method with 12 methods on two additional 2D SOR datasets, namely ASSR (Siris et al. 2020), and SalMon (Yildirim et al. 2020). We use RInet model as the baseline, our method achieves significant performance improvements (see Table 1). Our method performs slightly below the QAGNet on the ASSR dataset for two main reasons: first, unlike SOTA methods that are individually optimized for each SOR dataset, which may lead to overfitting. Our approach does not utilize any datasets tailored for ranking task. Second, our method focus on fine-grained prediction (see Fig. 8) which involves complex inter-object relationships, thereby increasing the challenge.

Component Evaluation

We validate the effectiveness of each component in the proposed method, quantitative results are listed in Table 2. **The effectiveness of LIE and LRC.** We take RInet model as baseline (see Table 2-①), by supplementing missing saliency values using the LIE and LRC methods, our method obtain a 3% performance improvement (see Table 2-②,③). **The effectiveness of PLE.** The data in Table 2-④ show that our multi-label aggregation method can integrate various complementary saliency cues generated by different saliency models, proving more accurate predictions than those single-label methods. **The Effectiveness of the Model Fine Tuning.** Table 2-⑤ shows that our proposed network fine-tuning strategy significantly improves the model's overall performance. It is noteworthy that although we introduced additional data, we did not use real ground truth, we train the model using high-quality pseudo-labels. **Generalization Ability.** We further evaluate the generalization ability of our method (see Table 3). For the 2D SOR models, we erase the TOP-5 objects in each iteration until no objects are recognizable. Our method achieve an average 4% improvement in the SA-SOR scores.

Method	SA-SOR	SOR
LAMA	**0.628**	**0.868**
MAEF	0.623	0.865
MAT	0.625	0.867
Base	0.616	0.862

A. Object Erasing Tools

Method	SA-SOR	SOR
Max	0.450	0.754
Mean	0.610	0.854
Sum	**0.616**	**0.862**
Mean-X	0.603	0.859

B. Sal-to-Rank Conversion

Method	SA-SOR	SOR
Base	0.616	0.862
ER=0	0.621	0.865
ER=10	**0.628**	**0.868**
ER=20	0.624	0.866

C. Expansion Rates (ER)

Table 4: Ablation studies on iterative erasure method.

Ablation Study

Different Object Erasing Tools. We conduct an in-depth comparison of three different object erasing tools, including LAMA (Suvorov et al. 2022), MAEF (Cao, Dong, and Fu 2022) and MAT (Li et al. 2022). From Table 4-A, LAMA achieve the best performance due to it can avoid adding unnecessary elements when reconstructing the original scene. **Different Sal-to-Rank Conversion.** We compare four common Sal-to-Rank methods, we calculate the average/maximum/sum pixel values within each object to assign importance ranking. Mean-X refers to the method proposed by (Song et al. 2024). From the data in Table 4-B, the maximum value achieved the best results because larger objects are typically more important. **Different Expansion Rates (ER).** We explore how various expansion rates ER = {0, 10, 20} affect erasure performance (see Table 4-C). Please refer to *supplementary material* for more details.

Conclusion

In this paper, we introduce a new task, named FOIR-360, which emphasizes the importance of non-salient objects that existing SOR tasks typically overlook. Meanwhile, to address the issue of data shortage, we propose a weakly-supervised approach, whose key technical innovations include a new object erasing strategy and a novel collection strategy. Besides, we propose a new annotation protocol that can flexibly annotate the ranking of important and non-important objects. Based on this annotation protocol, we establish a small-scale 360Rank dataset. Our contributions have the potential to drive advancements in the field.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China (No. 62172246), the Youth Innovation and Technology Support Plan of Colleges and Universities in Shandong Province (2021KJ062), and the Natural Science Foundation of Shandong Province (ZR2024YQ071).

References

Aydemir, B.; Hoffstetter, L.; Zhang, T.; Salzmann, M.; and Süsstrunk, S. 2023. Tempsal-uncovering temporal information for deep saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6461–6470.

Cao, C.; Dong, Q.; and Fu, Y. 2022. Learning prior feature and attention enhanced image inpainting. In *European Conference on Computer Vision*, 306–322.

Chao, F.-Y.; Zhang, L.; Hamidouche, W.; and Deforges, O. 2018. Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. In *IEEE International Conference on Multimedia & Expo Workshops*, 01–04.

Chao, F.-Y.; Zhang, L.; Hamidouche, W.; and Dforges, O. 2020. A Multi-FoV Viewport-Based Visual Saliency Model Using Adaptive Weighting Losses for. *IEEE Transactions on Multimedia*, 23: 1811–1826.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 1906. MMDetection: Open mmlab detection toolbox and benchmark. arXiv 2019. *arXiv preprint arXiv:1906.07155*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Cokelek, M.; Imamoglu, N.; Ozcinar, C.; Erdem, E.; and Erdem, A. 2023. Spherical Vision Transformer for 360-degree Video Saliency Prediction. arXiv:2308.13004.

Dahou, Y.; Tliba, M.; McGuinness, K.; and O'Connor, N. 2021. ATSal: an attention based architecture for saliency prediction in 360 videos. In *International Conference on Pattern Recognition*, 305–320.

Deng, B.; Song, S.; French, A. P.; Schluppeck, D.; and Pound, M. P. 2024. Advancing Saliency Ranking with Human Fixations: Dataset Models and Benchmarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 28348–28357.

DeSouza, G. N.; and Kak, A. C. 2002. Vision for mobile robot navigation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24: 237–267.

Droste, R.; Jiao, J.; and Noble, J. A. 2020. Unified Image and Video Saliency Modeling. In *European Conference on Computer Vision*, 419–435.

Duan, H.; Zhu, X.; Zhu, Y.; Min, X.; and Zhai, G. 2024. A Quick Review of Human Perception in Immersive Media. *IEEE Open Journal on Immersive Displays*, 41–50.

Esteves, C.; Allen-Blanchette, C.; Makadia, A.; and Daniilidis, K. 2018. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision*, 52–68.

Fan, S.; Shen, Z.; Jiang, M.; Koenig, B. L.; Xu, J.; Kankanhalli, M. S.; and Zhao, Q. 2018. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7521–7531.

Guan, H.; and Lau, R. W. 2024. SeqRank: Sequential Ranking of Salient Objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1941–1949.

Islam, M. A.; Kalash, M.; and Bruce, N. D. 2018. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7142–7150.

Jia, S.; and Bruce, N. D. 2020. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95: 103887.

Jiang, C. M.; Huang, J.; Kashinath, K.; Prabhat; Marcus, P.; and Niessner, M. 2019. Spherical CNNs on Unstructured Grids. arXiv:1901.02039.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23: 4909–4926.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision*, 4015–4026.

Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2023. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3041–3050.

Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10758–10768.

Lin, J.; Guan, H.; and Lau, R. W. H. 2022. Rethinking Video Salient Object Ranking. arXiv:2203.17257.

Ling, Z.; Xing, Z.; Zhou, X.; Cao, M.; and Zhou, G. 2023. Panoswin: a pano-style swin transformer for panorama understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 17755–17764.

Liu, N.; Li, L.; Zhao, W.; Han, J.; and Shao, L. 2021. Instance-level relative saliency ranking with graph reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8321–8337.

Lou, J.; Lin, H.; Marshall, D.; Saupe, D.; and Liu, H. 2022. TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494: 455–467.

Ma, G.; Li, S.; Chen, C.; Hao, A.; and Qin, H. 2020. Stagewise salient object detection in 360 omnidirectional image

via object-level semantical saliency ranking. *IEEE Transactions on Visualization and Computer Graphics*, 26(12): 3535–3545.

Qiao, M.; Xu, M.; Jiang, L.; Lei, P.; Wen, S.; Chen, Y.; and Sigal, L. 2024. HyperSOR: Context-aware Graph Hypernetwork for Salient Object Ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9): 5873–5889.

Siris, A.; Jiao, J.; Tam, G. K.; Xie, X.; and Lau, R. W. 2020. Inferring attention shift ranks of objects for image saliency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12133–12143.

Song, M.; Li, L.; Wu, D.; Song, W.; and Chen, C. 2024. Rethinking object saliency ranking: A novel whole-flow processing paradigm. *IEEE Transactions on Image Processing*, 33: 338–353.

Song, Y.; Liu, Z.; Li, G.; Zeng, D.; Zhang, T.; Xu, L.; and Wang, J. 2023. RINet: Relative importance-aware network for fixation prediction. *IEEE Transactions on Multimedia*, 25: 9263–9277.

Sun, C.; Xu, Y.; Pei, J.; Fang, H.; and Tang, H. 2023. Partitioned Saliency Ranking with Dense Pyramid Transformers. In *Proceedings of the ACM International Conference on Multimedia*, 1874–1883.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2149–2159.

Tian, X.; Xu, K.; Yang, X.; Du, L.; Yin, B.; and Lau, R. W. 2022. Bi-directional object-context prioritization learning for saliency ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5882–5891.

Wang, G.; Chen, C.; Hao, A.; Qin, H.; and Fan, D.-P. 2024. WinDB: HMD-free and Distortion-free Panoptic Video Fixation Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Wu, J.; Xia, C.; Yu, T.; and Li, J. 2023. View-aware Salient Object Detection for 360 Omnidirectional Image. *IEEE Transactions on Multimedia*, 25: 6471–6484.

Xie, J.; Liu, Z.; Li, G.; Lu, X.; and Chen, T. 2024. Global semantic-guided network for saliency prediction. *Knowledge-Based Systems*, 284: 111279.

Xu, H.; Zhao, Q.; Ma, Y.; Li, X.; Yuan, P.; Feng, B.; Yan, C.; and Dai, F. 2022. Pandora: A panoramic detection dataset for object with orientation. In *European Conference on Computer Vision*, 237–252.

Xu, M.; Yang, L.; Tao, X.; Duan, Y.; and Wang, Z. 2021. Saliency prediction on omnidirectional image with generative adversarial imitation learning. *IEEE Transactions on Image Processing*, 30: 2087–2102.

Yildirim, G.; Sen, D.; Kankanhalli, M.; and Süsstrunk, S. 2020. Evaluating salient object detection in natural images with multiple objects having multi-level saliency. *IET Image Processing*, 14(10): 2249–2262.

Zhang, P.; Wang, D.; Lu, H.; Wang, H.; and Ruan, X. 2017. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 202–211.

Zou, Z.; Ye, M.; Li, S.; Li, X.; and Dufaux, F. 2023. 360° image saliency prediction by embedding self-supervised proxy task. *IEEE Transactions on Broadcasting*, 69(3): 704–714.