# RINet: Relative Importance-Aware Network for Fixation Prediction

Yingjie Song, Zhi Liu, *Senior Member, IEEE*, Gongyang Li, Dan Zeng, *Senior Member, IEEE*, Tianhong Zhang, Lihua Xu, and Jijun Wang

***Abstract*—Fixation prediction aims to simulate human visual selection mechanism and estimate the visual saliency degree of regions in a scene. In semantically rich scenes, there are generally multiple salient regions. This condition requires a fixation prediction model to understand the relative importance relationship of multiple salient regions, that is, to identify which region is more important. In practice, existing fixation prediction models implicitly explore the relative importance relationship in the end-to-end training process while they do not work well. In this article, we propose a novel Relative Importance-aware Network (RINet) to explicitly explore the modeling of relative importance in fixation prediction. RINet perceives multi-scale local and global relative importance through the Hierarchical Relative Importance Enhancement (HRIE) module. Within a single scale subspace, on the one hand, HRIE module regards the similarity matrix as the local relative importance map to weight the input feature. On the other hand, HRIE module integrates a set of local relative importance maps into one map, defined as the global relative importance map, to grasp global relative importance. Moreover, we propose a Complexity-Relevant Focal (CRF) loss for network training. As such, we can progressively emphasize learning difficult samples for better handling the complicated scenarios, further improving the performance. The ablation studies confirm the contributions of key components of our RINet, and extensive experiments on five datasets demonstrate our RINet is superior to 28 relevant state-of-the-art models. Our code and results are available at: https://github.com/Mango321321/RINet.**

***Index Terms*—Fixation prediction, relative importance, self-attention mechanism, complexity-relevant focal loss.**

## I. Introduction

HUMAN eyes receive a huge amount of visual stimulation every second and analyze them quickly. It is the human visual attention mechanism that can efficiently select the most important area in the visual field, so that the brain can handle the mass data. This mechanism serves as an information bottleneck to filter the valuable content [1]. In the field of computer vision, the researchers have spent many efforts on understanding and simulating this mechanism, resulting in the topic of fixation prediction (FP). Encouragingly, the existing FP works widely facilitate many meaningful applications, such as salient object detection [2], [3], visual quality assessment [4]–[6], video understanding [7]–[10], object segmentation [11]–[13] and so on.

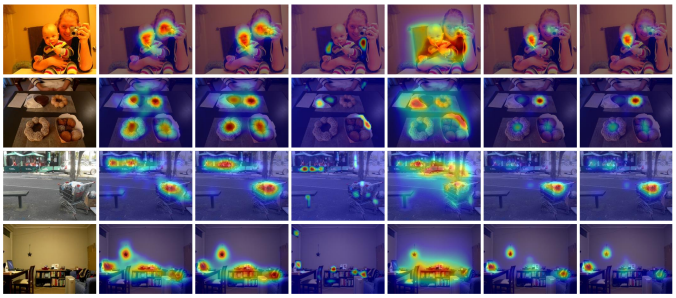



Fig. 1. Visual examples of proposed RINet and several state-of-the-art models. GT is short for ground truth, which is the fixation density map generated from the eye movement data of human observers.

Inspired by biological evidence, most traditional models extract hand-crafted features to capture low-level cues. However, they may not be able to handle the complex scenes because of unadaptable parameters for various situations. In the past few years, thanks to the emergence of high-quality datasets and benchmarks [14], [15], the deep learning-based approaches have boosted the accuracy of FP significantly as shown in Fig. 1 (ITTI [16] and GBVS [17] are traditional models, while SAM [18] and DINet [19] are deep learning-based models). These deep learning-based models go beyond the limitations of hand-crafted features and gather semantic information better.

Among the deep learning-based models, DINet is an excellent model that utilizes multi-scale contextual features to learn semantic information, and works well on simple scenes, *e.g.*, only one object appears in the image. However, the multi-scale contextual features extracted by DINet cannot fully understand the relative importance of multiple objects within a single image. SAM suffers from the same problem. Visually, as shown in the 2nd row of Fig. 1, DINet and SAM are able to pick out four salient regions correctly while only clearly highlighting one of them. It can be concluded that even the state-of-the-art FP models still cannot handle the relative importance of image regions in semantically rich scenarios [19]. Existing works implicitly include the relative importance relationship

This work was supported in part by the National Natural Science Foundation of China under Grants 62171269 and 82171544, in part by the Science and Technology Commission of Shanghai Municipality under Grant 21S31903100 and in part by the China Postdoctoral Science Foundation under Grant 2022M722037. *(Corresponding author: Zhi Liu.)*

Yingjie Song, Zhi Liu, Gongyang Li, and Dan Zeng are with Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (email: jie0222@shu.edu.cn; liuzhisjtu@163.com; ligongyang@shu.edu.cn; dzeng@shu.edu.cn).

Tianhong Zhang, Lihua Xu, and Jijun Wang are with Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiaotong University School of Medicine, Shanghai 200030, China (e-mail: zhang_tianhong@126.com; dr_xulihua@163.com; dr_wangjijun@126.com).

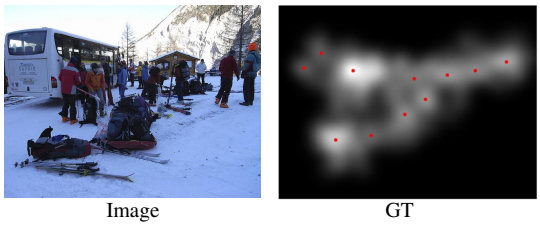


Image GT

Fig. 2. Quantization of input image's complexity degree. The red points in GT stand for local maximum. GT is the ground truth.

during the end-to-end training procedure, which is obviously sub-optimal. As a result, understanding the relative importance relationship in semantically rich scenes is one possible direction to realize the next improvement.

Inspired by the above observations, in this article, we focus on making model explicitly understand the relative importance relationship. Therefore, we propose the novel Relative Importance-aware Network (RINet) to achieve it from the following two aspects.

Firstly, a novel Hierarchical Relative Importance Enhancement (HRIE) module is proposed to explicitly build the relative importance relationship. A ticklish question is how to compare the image region, further to get the relative importance relationship. Motivated by the self-attention mechanism [20], we compute the patch embeddings to represent image regions, and regard the similarity matrix as the importance map to re-weight the input. Because the similarity matrix takes one patch as reference, it can be considered as the local relative importance map. We go through all patches as reference so as to harvest a set of local relative importance maps. To get a single explicit relative importance relationship of an image, these local relative importance maps are integrated into one global map by a well-designed operation called global importance scoring (GIS). We further expand the above operations from spatial domain to channel domain. In a word, we explicitly build the local and global (local-global in short) relative importance maps. The other ticklish question is the scale variation of different salient regions, which is an obstacle during relative importance comparison. Broadly speaking, deep layers with a large receptive field can acquire semantic information and handle large salient regions, while they tend to ignore detailed information and small salient regions. The shallow features are the opposite. Consequently, we model the local-global relative importance relationship on features in a hierarchical manner to solve the scale problem.

Secondly, we design the Complexity-Relevant Focal (CRF) loss to boost grasping relative importance in complicated scenes. Intuitively, the difficulty to understand global relative importance increases with the complexity of input. To explore the association between input complexity and models' performance, we quantize the complexity using the number of local maximum as shown in Fig. 2. We experimentally find that the complexity degree and evaluation metrics of fixation prediction performance are negatively correlated, which indicates the images with more salient regions are harder to grasp. Therefore, we put forward a complexity-relevant focusing factor to adjust the contribution of complicated scenes. In addition, we find emphasizing the complicated scenes directly is slightly too large to converge. Consequently, the emphasis is designed to grow progressively as a warming-up procedure.

With the help of the HRIE module and CRF loss, our RINet can understand the relative importance relationship better. As shown in Fig. 1, RINet learns the most approximate relative importance relationship compared to ground truths. Our main contributions are summarized as follows:

- We propose a novel relative importance-aware network (RINet) to explicitly explore modeling of relative importance for fixation prediction. In RINet, the local-global relative importance map is generated by a well-designed HRIE module, which distinguishes features from low level to high level.
- We design a complexity-relevant focal loss to properly emphasize the complicated samples. It adjusts the contribution of the complicated ones by a complexity-relevant focusing factor, which can be easily extended to other loss functions used in FP with negligible additional computation cost.
- Comprehensive experiments on five challenging datasets prove that the proposed RINet achieves superior performance compared with 28 state-of-the-art models, and demonstrate the effectiveness of the proposed HRIE module and CRF loss.

## II. Related Work

### A. Fixation Prediction

*1) Traditional Fixation Prediction Models:* Fixation prediction is also known as saliency prediction. It can be generally classified into two categories, namely scene-driven approaches (*a.k.a* bottom-up saliency) and expectation-driven approaches (*a.k.a* top-down saliency) [21]. Since our model aims at predicting free-viewing fixation over an image, we mainly focus on the related works of scene-driven models.

The scene-driven models have a long history of non-deep learning algorithms. The origin of early works dates back to the seminal work of feature integration theory [22], which suggests that human brains process multiple preliminary features simultaneously and combine them to guide human attention. Rooted in this work, subsequently, relevant researches mined the low-level cues by using hand-crafted features [23]. For example, Itti *et al.* [16] implemented a cognitive model based on the biologically computational architecture [1]. They decomposed the visual input into a set of preliminary feature maps, including intensity, color, and orientation. These feature maps were normalized and summed across Gaussian pyramid to yield the predicted fixation density map. Furthermore, some works [24]–[26] developed the cognitive models by taking more well-designed operations to process the low-level feature. Some other works [27], [28] considered the low-level feature processing from the view of information theory. They regarded the feature, which is rare statistically, as the most informative part of an image. In addition to the above models, there were several models processing the low-level features from other perspectives, such as Bayesian [29], [30], decision

theory [31], [32], spectral analysis [33], [34], and pattern classification [15], [35].

Although these traditional models are explainable in computational principles, their generalization capability is limited. The hand-crafted features based on low-level cues are difficult to handle complex scenes.

*2) Deep Learning-Based Fixation Prediction Models:* To improve the generalization capability, many deep learning-based models for FP have sprung up in recent years. They achieved the above goal by making models understand semantic information of an image. Kümmerer *et al.* [36], [37] first attempted to apply transfer learning to FP. They enabled the model to leverage the knowledge from existing CNNs trained in image classification task. Afterwards, many works proposed novel attempts in model structure to gather semantic information better.

Specifically, multi-scale feature fusion, generative adversarial network and long-range contextual information are three popular strategies. Here, we focus on the first one. For instance, Liu *et al.* [38] designed three convolutional neural networks (CNNs) of different scales and fused them with two fully connected layers, which were trained using multiresolution image patches centered on fixation and non-fixation locations. Kruthiventi *et al.* [39] employed a relatively deeper CNN (*i.e.*, VGG-16 [40]), and proposed a novel location-biased convolutional layer to fuse semantic information at multi-scale subspaces. In [41], Cornia *et al.* extracted features from different levels of VGG network into encoder and added a learnable center prior by learning a set of Gaussian parameters. Wang *et al.* [42] took three decoders to perceive the different scale subspaces of encoder network and integrate them into the predicted fixation density map. Yang *et al.* [19] further thought about computational efficiency of multi-scale features. They employed a dilated inception network where the parallel inception structure captured multi-scale features and the dilated convolutions significantly reduced the computation load. Reddy *et al.* [43] proposed a minimal FP model and briefly summarized four key components of FP models, including input features, multi-level integration, readout architecture, and loss functions. Wang *et al.* [2], [3] utilized the features from higher layers to model human fixation locations. Then they combined the fixation prior with the features from the lower layers to facilitate the task of salient object detection. In addition to multi-scale feature fusion, Pan *et al.* [44] and Che *et al.* [45] introduced generative adversarial network into FP. Liu *et al.* [46] and Lou *et al.* [47] emphasized long-range information during feature extraction by long short-term memory (LSTM) model and multi-head self-attention model, respectively.

The above-mentioned studies can deal with simple scenes. Nevertheless, as for the semantically rich scenes, they still cannot fully understand the relative importance of multiple salient regions. In this article, we pay attention to making the model explicitly understand the relative importance relationship, and propose an effective solution named RINet. Concretely, we present the HRIE module to learn the local-global relative importance relationship at multiple scale subspaces. Moreover, the CRF loss is designed to place extra emphasis on relatively complex scenarios.

### *B. Attention Mechanism*

The human vision system can distinguish the informative regions from their neighbors naturally. The attention mechanism imitates this inherent characteristic of human vision by adaptively weighting features according to the importance of the input [48].

The development of attention mechanism can be coarsely divided into two stages. As for the first stage, training an additional CNN branch to re-weight the input feature is the main characteristic. SENet [49] proposed a squeeze-and-excitation (SE) block to emphasize important channels. The squeeze operation extracted the global spatial information and the follow-up excitation operation exploited fully-connected layers to capture channel-wise attention vector. In [50], Park *et al.* designed two parallel branches, *i.e.*, channel attention and spatial attention, which were carried out simultaneously and integrated into one matrix to weight the input. Differently, CBAM [51] stacked channel attention and spatial attention in series. The channel and spatial attention maps were multiplied to the input feature sequentially. Li *et al.* [52] applied attention mechanism to feature fusion, which suggested high-level features could serve as guidance to select low-level features. The second stage is the self-attention era. The pioneering work [20] introduced self-attention to computer vision from the field of natural language processing [53]. It proposed the non-local block which took the similarity matrix as importance map to weight the input feature. Following [53], many subsequent works made progresses in different aspects including accuracy improvement [54] and computational complexity reduction [55], [56].

In our work, we consider the similarity matrix in self-attention mechanism [20] as the local relative importance map. Furthermore, we propose the GIS operation to obtain global relative importance maps on both spatial and channel dimensions. Given different scales of salient regions, we enhance the relative importance in a hierarchical manner. In this way, the relative importance relationship is modeled explicitly, which helps the model handle the semantically rich scenes and pushes the FP task to achieve further improvement.

### *C. Discriminative Loss Function*

There has been much interest in designing loss functions to discriminate the special samples and put emphasis on these samples when training. For example, Hastie *et al.* [57] down-weighted the loss of examples with large errors to reduce the disturbance of outliers. In [58], the class imbalance was relieved by emphasizing samples of minority classes. To force the network to learn the small separation borders, the distances to the border of the nearest cell and the second nearest cell were taken into loss function as part of weight parameter. The representative focal loss [59] discriminated the hard samples during training. It down-weighted the contribution of easy samples to total loss. Based on focal loss, Ridnik *et al.* [60] and Li *et al.* [61] solved the issue of positive sample scarcity and long-tailed data distribution further, respectively. Ridnik *et al.*

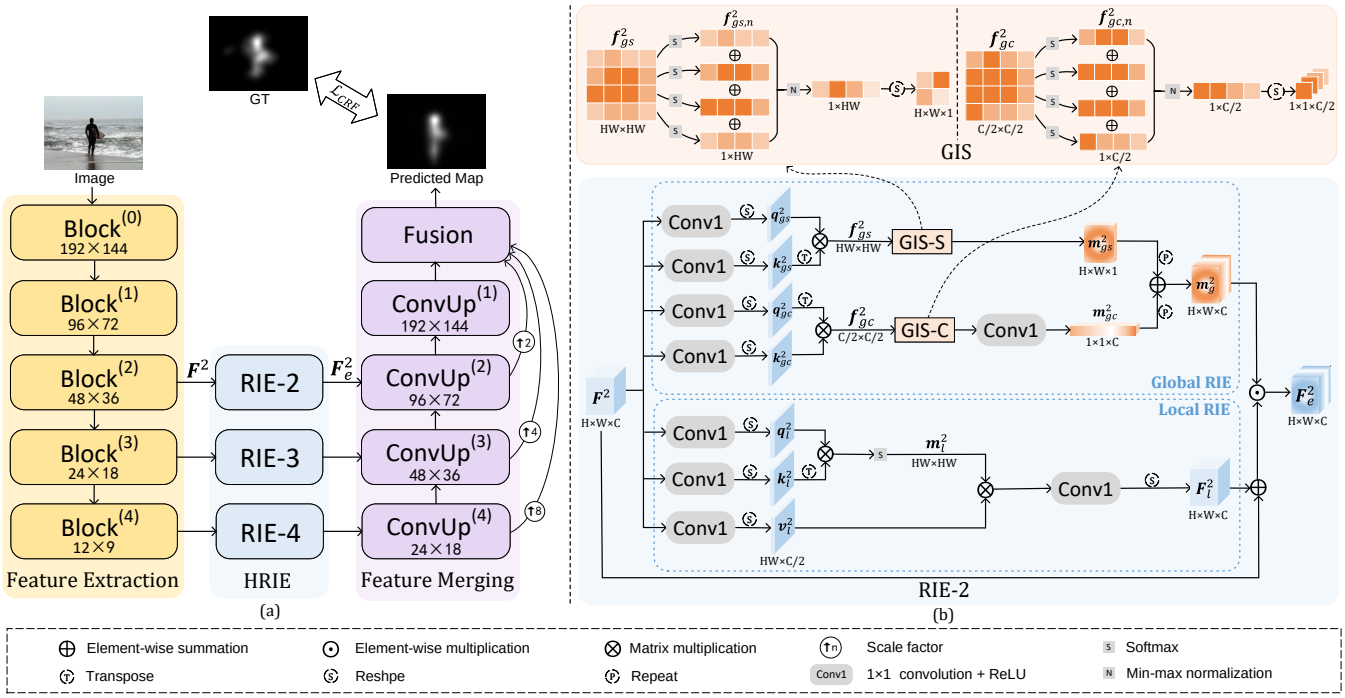



Fig. 3. (a) Pipeline of the proposed RINet. Our model contains three stages: feature extraction, hierarchical relative importance enhancement (HRIE) and feature merging. Firstly, the feature extraction stage encodes the features from an input image. Then, the relative importance of basic features at different scales is enhanced by HRIE. Finally, the output features of HRIE are fused to predict the fixation density map in feature merging stage. (b) The details of one branch of HRIE, *i.e.*, RIE-2, where the input of Global Importance Scoring (GIS) operation is simplified to $H = W = 2$ or $C/2 = 4$ for understanding.

suggested the discrimination of easy samples and hard samples in focal loss was insufficient. Therefore, they proposed an additional asymmetric scheme to adjust the contribution of easy samples. Li *et al.* thought focal loss only worked well on the foreground-background imbalance problem under the category-balanced distribution but cannot handle the long-tailed situation. They adopted a category-relevant focusing factor to address the positive-negative imbalance of different categories separately. Besides, Li *et al.* [62], [63] changed the sharp one-hot label to soft quality label on the basis of focal loss to merge the classification branch and classification quality prediction branch.

Inspired by the above, we introduce the focal loss into FP task, and propose a CRF loss to discriminate the hard samples and emphasize them. We discover the more complex the input image, the worse the prediction performance, where the complexity is quantified by the number of local maximum. Therefore, we regard the complex scene as hard samples and design a complexity-relevant focusing factor in our CRF loss to assist the model in understanding the relative importance of hard samples.

## III. Methodology

In this section, we illustrate the proposed Relative Importance-aware Network (RINet). In Sec. III-A, we present the overall architecture of our model. In Sec. III-B, we elaborate on the details of Hierarchical Relative Importance Enhancement (HRIE) module. In Sec. III-C, we describe the complexity-relevant focal (CRF) loss. Finally, the implementation details are provided in Sec. III-D.

### A. Architecture Overview

An extraction-merging structure is adopted in the proposed RINet. In Fig. 3(a), we can see that RINet consists of three stages, including feature extraction, hierarchical relative importance enhancement, and feature merging.

*1) Features Extraction:* CNNs are able to extract multi-level features from the image where the shallow layers capture low-level texture cues and deep layers encode high-level semantic information. We exploit the fully-convolutional part of DenseNet [64] as the backbone, that is, the last global average pooling layer and fully connected layer are removed. The backbone is divided into five stages and each stage is denoted as $\text{Block}^{(i)}$ in which $i \in \{0, 1, 2, 3, 4\}$ stands for the index of block. We extract the feature map from the last convolution layer of each block and denote them as $\mathbf{F}^i \in \mathbb{R}^{H \times W \times C}$ where $H$, $W$ and $C$ are the dimension of the height, width and channel, respectively. The input size is 384×288 and the output size of every block is shown in Fig. 3(a).

*2) Hierarchical Relative Importance Enhancement:* In particular, we employ three parallel branches named RIE-$i$ to enhance features hierarchically. These branches generate the local-global relative importance maps at different scale subspaces. We believe that the semantic information contained in three deep layers is enough to understand saliency of an image. More shallow layers bring undesired details and cause a lot of extra computational cost, which is proved by the experiments in Sec. IV-C. The outputs after enhancement are denoted as $\mathbf{F}_e^i \in \mathbb{R}^{H \times W \times C}$ and the details of HRIE are described in Sec. III-B.

*3) Feature Merging:* The feature merging stage is designed with respect to the features extraction stage, that is, the feature merging block ConvUp$^{(i)}$ matches the corresponding Block$^{(i)}$ ($i \in \{1,2,3,4\}$). ConvUp$^{(i)}$ block consists of concatenation, one convolutional layer and one upsampling layer. The upsampling layer uses the bilinear upsampling algorithm to restore $2\times$ resolution. Notably, there is no concatenation operation in ConvUp$^{(4)}$ because ConvUp$^{(4)}$ is the deepest layer. Given ConvUp$^{(i)}$ only merges the two adjacent blocks, we further adopt a Fusion block to fuse the multi-level features. The outputs of ConvUp$^{(2)}$, ConvUp$^{(3)}$ and ConvUp$^{(4)}$ are upsampled $2\times$, $4\times$ and $8\times$, respectively, to match the output size of ConvUp$^{(1)}$. Actually, the Fusion block has two more convolutional layers than ConvUp$^{(i)}$, facilitating the generation of fixation density map.

### B. Hierarchical Relative Importance Enhancement

Convolutional operation calculates the information within a fixed window so that it is the inherent characteristic that convolution fails to handle the salient regions of various sizes. We propose the HRIE module with a hierarchical structure to compensate limitation of convolution. The details of HRIE are shown in Fig. 3(b). It is comprised of two key components: Relative Importance Enhancement (RIE) block and Global Importance Scoring (GIS) operation. In the following, we elaborate on these two components.

*1) Relative Importance Enhancement Block:* To obtain the relative importance relationship of multiple image regions, we have to compare different image regions. We design a local relative importance enhancement (local RIE) branch to measure the importance of different image regions with one region as reference. To obtain an entire explicit relative importance map, we further propose a global relative importance enhancement (global RIE) branch to fuse local relative importance maps into one map.

As for local RIE branch of RIE-$i$ ($i \in \{2,3,4\}$), we first apply the convolutional operation to reduce the channel dimension by half to economize on computational cost, and then apply reshape operation to get the feature embeddings which refer to queries $\boldsymbol{q}_l^i$, keys $\boldsymbol{k}_l^i$ and values $\boldsymbol{v}_l^i$. The size of $\boldsymbol{q}_l^i$, $\boldsymbol{k}_l^i$ and $\boldsymbol{v}_l^i$ are $\mathbb{R}^{HW\times C/2}$, where $H$ is height of corresponding input feature $\mathbf{F}^i$, $W$ is width and $C$ is channel dimension. Furthermore, the bottleneck structure makes the feature compressive. This process can be defined as:

$$\begin{aligned} \boldsymbol{q}_l^i &= rs(conv(\mathbf{F}^i; \mathbf{W}_{l,q}^i)), i \in \{2,3,4\}, \\ \boldsymbol{k}_l^i &= rs(conv(\mathbf{F}^i; \mathbf{W}_{l,k}^i)), i \in \{2,3,4\}, \\ \boldsymbol{v}_l^i &= rs(conv(\mathbf{F}^i; \mathbf{W}_{l,v}^i)), i \in \{2,3,4\}, \end{aligned} \tag{1}$$

where $rs(\cdot)$ represents reshape operation, and $conv(*; \mathbf{W}_{l,*}^i)$ means the convolution layer with parameters $\mathbf{W}_{l,*}^i$. We employ the matrix multiplication between $\boldsymbol{q}_l^i$ and $\boldsymbol{k}_l^i$ to obtain the local relative importance map $\boldsymbol{m}_l^i \in \mathbb{R}^{HW\times HW}$ and exploit softmax function to normalize it as follows:

$$\boldsymbol{m}_l^i = Softmax(\boldsymbol{q}_l^i \otimes (\boldsymbol{k}_l^i)^T), \tag{2}$$

where $\otimes$ indicates matrix multiplication and $Softmax(\cdot)$ is the softmax function. After $\boldsymbol{v}_l^i$ is weighted by the local relative importance map $\boldsymbol{m}_l^i$, the following convolution layer restores its channel from $C/2$ to $C$. Specifically, the feature is processed by local RIE as:

$$\mathbf{F}_l^i = rs(conv((\boldsymbol{m}_l^i \otimes \boldsymbol{v}_l^i); \mathbf{W}_l^i)), \tag{3}$$

As for global RIE, the global relative importance is built in both spatial domain and channel domain. In spatial domain, we use the same way as local RIE, transforming extracted feature to queries $\boldsymbol{q}_{gs}^i$ and keys $\boldsymbol{k}_{gs}^i$. Afterwards, the well-designed operation GIS transforms the local relative importance map to the global relative importance map $\boldsymbol{m}_{gs}^i \in \mathbb{R}^{H\times W\times 1}$ as:

$$\boldsymbol{m}_{gs}^i = GIS(\boldsymbol{q}_{gs}^i \otimes (\boldsymbol{k}_{gs}^i)^T), \tag{4}$$

where $GIS(\cdot)$ denotes the global importance scoring operation. In channel domain, analogously, global RIE computes feature embeddings using the same way as the spatial domain, generating queries $\boldsymbol{q}_{gc}^i \in \mathbb{R}^{HW\times C/2}$ and keys $\boldsymbol{k}_{gc}^i \in \mathbb{R}^{HW\times C/2}$. Based on GIS operation, the global RIE in channel domain is summarized as:

$$\boldsymbol{m}_{gc}^i = conv(GIS((\boldsymbol{q}_{gc}^i)^T \otimes \boldsymbol{k}_{gc}^i); \mathbf{W}_{gc}^i), \tag{5}$$

Referring to BAM [50], we join the global relative importance in spatial domain and channel domain into a weight matrix with the same size as $\mathbf{F}^i$. Concretely, the spatial global relative importance map $\boldsymbol{m}_{gs}^i$ repeats $C$ times along the channel dimension, while the channel global relative importance map $\boldsymbol{m}_{gc}^i$ repeats $H \times W$ times along the spatial dimension. The final weight matrix is calculated as follows:

$$\boldsymbol{m}_g^i = rp(\boldsymbol{m}_{gs}^i) \oplus rp(\boldsymbol{m}_{gc}^i), \tag{6}$$

where $rp(\cdot)$ represents repeat operation and $\oplus$ is element-wise summation. The relative importance of input features is enhanced locally and globally as:

$$\mathbf{F}_e^i = (\mathbf{F}^i \oplus \mathbf{F}_l^i) \odot \boldsymbol{m}_g^i, \tag{7}$$

where $\odot$ is element-wise multiplication. Because the local RIE branch has extracted information from $\mathbf{F}^i$ to $\boldsymbol{v}_l^i$ and weighted $\boldsymbol{v}_l^i$ with the local relative importance maps $\boldsymbol{m}_l^i$, the weighted feature $\mathbf{F}_l^i$ is added with $\mathbf{F}^i$. The global RIE branch only measures the relative importance in spatial and channel domains to generate the weight matrix $\boldsymbol{m}_g^i$, so that the global relative importance map $\boldsymbol{m}_g^i$ is multiplied with $\mathbf{F}^i$.

*2) Global Importance Scoring:* To obtain an entire explicit relative importance map rather than that with one patch embedding as reference, we use GIS operation to fuse multiple local relative importance maps into a global relative importance map. In spatial domain, the GIS operation is denoted as GIS-S. The result of matrix multiplication between $\boldsymbol{q}_{gs}^i$ and $(\boldsymbol{k}_{gs}^i)^T$ is defined as $\boldsymbol{f}_{gs}^i \in \mathbb{R}^{HW\times HW}$. As shown in the top part of Fig. 3(b), we utilize softmax function to normalize every row of $\boldsymbol{f}_{gs}^i$ and define the $n$-th row after softmax as $\boldsymbol{f}_{gs,n}^i \in \mathbb{R}^{1\times HW}$. $\boldsymbol{f}_{gs,n}^i$ measures the similarity between the $n$-th feature embedding $\boldsymbol{q}_{gs,n}^i \in \mathbb{R}^{1\times C/2}$ (*i.e.*, the $n$-th row of $\boldsymbol{q}_{gs}^i$) and all feature embeddings (*i.e.*, all columns of $(\boldsymbol{k}_{gs}^i)^T$). Thus, $\boldsymbol{f}_{gs,n}^i$ can be regarded as the local relative importance map with the $n$-th feature embedding as reference. Next, we add multiple local relative importance maps element by

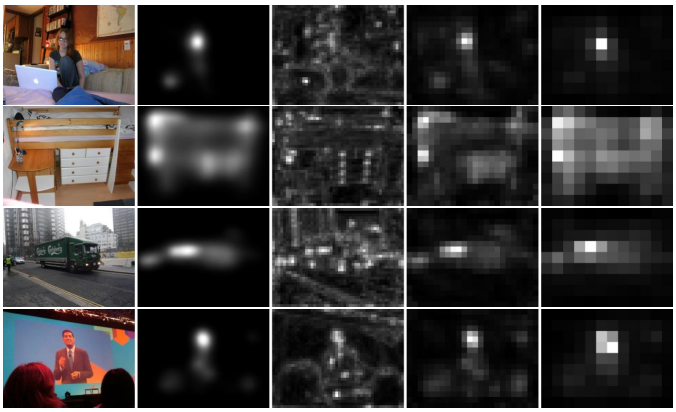



Fig. 4. Visualization of $m^i_{gs}$ where $i \in \{2, 3, 4\}$.

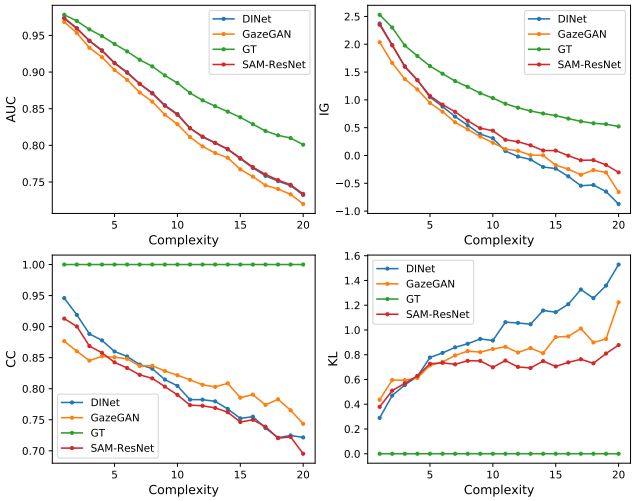



Fig. 5. The statistics about complexity and evaluation metrics on SALICON validation set.

element and rescale them by min-max normalization. Finally, the result is reshaped to restore its original spatial size. The GIS-S operation can be formulated as follows:

$$\boldsymbol{m}^i_{gs} = rs(norm(\boldsymbol{f}^i_{gs,1} \oplus \boldsymbol{f}^i_{gs,2} \oplus \cdots \oplus \boldsymbol{f}^i_{gs,HW})), \quad (8)$$

where $\boldsymbol{m}^i_{gs} \in [0,1]^{H\times W\times 1}$ and $norm(\cdot)$ is min-max normalization. We visualize $\boldsymbol{m}^i_{gs}$ in Fig. 4. It can be seen, within a single scale subspace, the relative importance relationship is modeled explicitly by one map. In channel domain, the GIS operation is denoted as GIS-C. Similarly, the global relative importance map $\boldsymbol{m}^i_{gc}$ is generated as follows:

$$\boldsymbol{m}^i_{gc} = conv(rs(norm(\boldsymbol{f}^i_{gc,1} \oplus \cdots \oplus \boldsymbol{f}^i_{gc,C/2})); \mathbf{W}^i_{gc}), \quad (9)$$

where $\boldsymbol{m}^i_{gc} \in [0,1]^{1\times 1\times C}$ whose channel is restored to $C$ through a convolutional layer, $\boldsymbol{f}^i_{gc}$ is the matrix multiplication result of $(\boldsymbol{q}^i_{gc})^T$ and $\boldsymbol{k}^i_{gc}$, and $\boldsymbol{f}^i_{gs,n}$ denotes the $n$-th row of $\boldsymbol{f}^i_{gc}$ after softmax. Consequently, the final weight matrix $\boldsymbol{m}^i_g \in [0,2]^{H\times W\times C}$ can both down-weight and up-weight the feature embeddings.

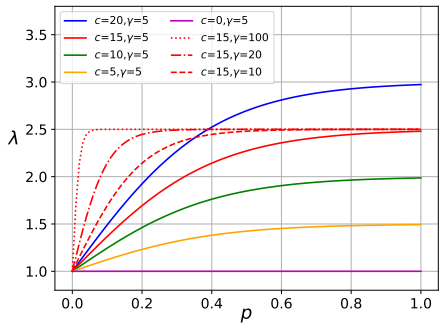



Fig. 6. The focusing factor $\lambda$ of different parameters $(c, \gamma)$ during training.

### C. Complexity-Relevant Focal Loss

The proposed CRF loss is derived from the focal loss [59] which intends to pay more attention to the difficult samples. Nevertheless, in FP, the definition of difficult samples is blurry. We assume that the difficulty of an image sample is related to the number of salient regions in the image, namely the local maximum number of ground truths. For the sake of simplification, the number of salient regions is defined as the complexity of an image. The relationship between performance and complexity is explored in Fig. 5. The statistical results based on several popular models support that the increase of complexity widens the gap between the predicted results and the ground truths, which validates our aforementioned hypothesis. Additionally, we find if the complexity is used to weight different samples directly, the model is hard to converge during training procedure. Learning from the warming-up method in [65], we design a complexity-relevant focusing factor $\lambda$ to progressively adjust the weighting process as:

$$\lambda = 1 + (\frac{2}{1+e^{-\gamma p}} - 1) \cdot \frac{c}{\alpha}, \quad (10)$$

where $\alpha$ is in charge of scaling and set to 10 in all experiments, $c$ is short for complexity, $p$ is the quotient of current iteration and total iterations which ranges from 0 to 1 during training, and $\gamma$ is responsible for tuning the weighting speed. As illustrated in Fig. 6, it can be seen that the focusing factor $\lambda$ ranges from 1 to $1 + \frac{c}{\alpha}$ gradually. As a result, the samples with higher $c$ (*i.e.*, the difficult samples) are emphasized by higher $\lambda$ progressively.

Pearson's Correlation Coefficient (CC) and Kullback-Leibler divergence (KL) are broadly-used evaluation metrics in FP. The former rates how correlated two variables are and the latter measures the difference between two probability distributions. They are computed as:

$$KL(\mathbf{P}, \mathbf{G}) = \sum_i \mathbf{G}_i \log\left(\epsilon + \frac{\mathbf{G}_i}{\epsilon + \mathbf{P}_i}\right), \quad (11)$$

$$CC(\mathbf{P}, \mathbf{G}) = \frac{cov(\mathbf{P}, \mathbf{G})}{sd(\mathbf{P}) \cdot sd(\mathbf{G})}, \quad (12)$$

where $\mathbf{P}$ is the predicted map, $\mathbf{G}$ is the ground truth, $\epsilon$ is a regularization constant, $sd(\cdot)$ represents standard deviation, and $cov(\cdot)$ refers to covariance. We propose the CRF loss based on these two metrics in this article. Typically, both

KL and CC are always positive values. The lower the KL score, the better the performance. Thus we multiply KL with the focusing factor $\lambda$ to increase the contribution of difficult samples. CC is the opposite, so we divide CC by $\lambda$ to achieve the same goal. Therefore, the CRF loss is defined as:

$$\mathcal{L}_{CRF}(\mathbf{P},\mathbf{G}) = \lambda \cdot KL(\mathbf{P},\mathbf{G}) - \frac{CC(\mathbf{P},\mathbf{G})}{\lambda}, \qquad (13)$$

The core idea of CRF loss is to gradually increase the contribution of difficult samples. The CRF loss can be easily extended to other loss functions in FP with tiny additional computational cost.

### D. Implementation Details

The proposed RINet is implemented by PyTorch [66] with one single NVIDIA TITAN Xp GPU. During training procedure, we initialize the backbone of RINet, *i.e.*, DenseNet-161, with the pre-trained weights on ImageNet [67]. The remaining weights are initialized by the default setting of PyTorch. We first train our model on the SALICON training set while monitoring whether it converges on the SALICON validation set. Then we fine-tune our model on the MIT1003 dataset [15] with the same evaluation protocol in [18], [19], [46], [68], that is, 903 images randomly selected from MIT1003 are used for training and the remaining 100 images are for validation. We use the Adam optimizer [69] to train our RINet and set the batch size to 8. The initial learning rate of training and fine-tuning are all set to $10^{-4}$, which will be divided by 10 after every four epochs. We train our model on SALICON for 10 epochs and fine-tune it on MIT1003 for five epochs. The input image size of the model is set to 384×288 with zero padding to keep the original aspect ratio. The weight $\gamma$ and $\alpha$ in Eq. 10 are fixed as 5 and 10, respectively.

## IV. Exprerimental Results

In this section, we present our experimental results and analyze them. In Sec. IV-A, the popular datasets and evaluation metrics in FP are introduced briefly. In Sec. IV-B, we compare our model with existing state-of-the-art models on five datasets quantitatively and qualitatively. In Sec. IV-C, the influence of main components of RINet is analyzed in detail. In Sec. IV-D, we explore how the parameters in CRF loss affect the performance.

### A. Datasets and Evaluation Metrics

*1) Datasets:* **SALICON** [14] is the largest public dataset in FP which includes 10,000 training images, 5,000 validation images and 5,000 test images. The size of each image is 480×640. This dataset uses mouse movements to simulate eye movements. The ground truth labels of test set are all held out. Researchers can evaluate their models on the SALICON challenge website[1].

**MIT1003** [15] contains 1,003 images and the eye movements are collected from 15 subjects. Usually, the FP models are trained on the SALICON dataset firstly and finetuned on the MIT1003 dataset subsequently.

**MIT300** [70] collects 300 natural images and the subjects are 39 observers. The ground truth labels are all held out. As a result, for evaluation, the researchers email their results to MIT Saliency Benchmark[2].

**TORONTO** [71] collects 120 images from indoor and outdoor scenes which are highly varied and natural. The eye movement data is collected from 20 subjects.

**PASCAL-S** [72] extends 850 existing images from PASCAL 2010 dataset[3] with eye fixations. A free-viewing task is given to 8 subjects to get the eye movement data.

**DUT-OMRON** [73] consists of 5,168 images and eye movement data collected from 5 subjects, which is a new large-scale challenging dataset in FP.

*2) Evaluation Metrics:* The evaluation metrics used in our experiments include two categories, *i.e.*, location-based and distribution-based metrics. The location-based metrics include Area Under ROC Curve (AUC), shuffled AUC (sAUC), Normalized Scanpath Saliency (NSS), and Information Gain (IG). The distribution-based metrics involve Similarity (SIM), CC, and KL. More detailed descriptions of these metrics can be found in [74]. For KL, lower value represents better performance. For the other metrics, larger values are regarded to be better.

### B. Comparison with the State-of-the-Art Models

*1) Quantitative Comparison:* We compare the proposed RINet with 28 state-of-the-art FP models, including ITTI [16], GBVS [17], AIM [27], CAS [75], SUN [29], SAM-ResNet [18], MSI-Net [76], GazeGAN [45], DINet [19], TranSalNet [47], UNISAL [77], SimpleNet [43], EML-NET [78], CEDNS [79], SalED [68], SalFBNet [80], DeepGaze I [36], DeepGaze II [37], ICF [37], DeepGaze II-E [81], CASNet II [82], DVA [42], SalGAN [44], MLNet [41], eDN [83], ACSalNet [84], FastSal [85], SATSal [86]. For SALICON and MIT300 benchmarks, we submit the results of our model trained on SALICON training set to SALICON benchmark and the results of our model fine-tuned on MIT1003 dataset to MIT300 benchmark. The results of other models are obtained from their corresponding articles or the two benchmarks. For TORONTO, PASCAL-S and DUT-OMRON datasets, since there are no public benchmarks, we conduct the test by ourselves. Because these three datasets are all natural eye movement data rather than mouse movement data as SALICON, we utilize our model fine-tuned on the MIT1003 dataset to test them. For ITTI, GBVS, AIM, CAS and SUN, we directly adopt the codes provided by pysaliency[4]. For MSI-Net, GazeGAN, UNISAL and SalFBNet, we use the publicly available models that have already been trained on the SALICON dataset and then fine-tuned on the MIT1003 dataset. Because the released models of SAM-ResNet, DINet, TranSalNet, and FastSal are only trained on the SALICON

[1]http://salicon.net/challenge-2017/

[2]https://saliency.tuebingen.ai/

[3]http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[4]https://github.com/matthias-k/pysaliency

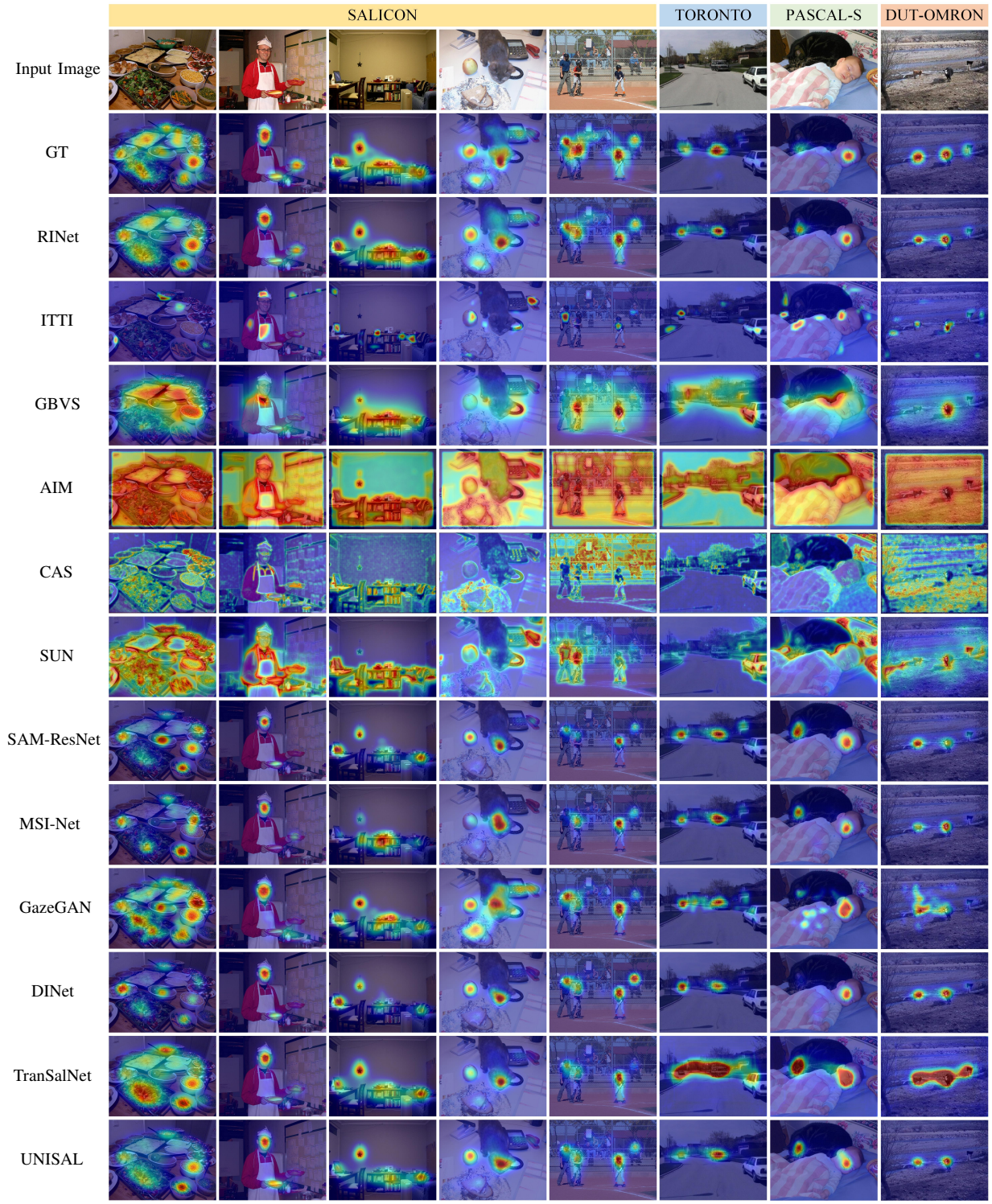



Fig. 7. Visual comparisons with start-of-the-art FP models. Eight examples are presented in a column-wise manner. The first five columns are from SALICON validation set. The last three columns are from TORONTO, PASCAL-S and DUT-OMRON datasets, respectively.

dataset, for a fair comparison, we fine-tune them on the MIT1003 dataset and then test on TORONTO, PASCAL-S and DUT-OMRON datasets. As for ACSalNet, we train it on SALICON dataset and then fine-tune it on the MIT1003 dataset.

Tab. I reports the quantitative performance comparison on SALICON benchmark. It can be observed that our model is superior to other models on five metrics of IG, CC, AUC, SIM and KL, and ranks third on sAUC. Our model is not very good at NSS. The reason is that NSS calculates the average of the normalized saliency values at the location of fixation [19]. If a model obtains a higher NSS, the predicted map will be more like a discrete fixation map rather than a fixation density map, which means it is difficult to achieve a promising result on

TABLE I
QUANTITATIVE PERFORMANCE COMPARISON ON SALICON BENCHMARK. ↑ AND ↓ STAND FOR LARGER VALUE AND SMALLER VALUE ARE BETTER, RESPECTIVELY. THE BEST THREE RESULTS ARE MARKED IN RED, BLUE AND GREEN, RESPECTIVELY.

| Models | SALICON | | | | | | |
|---|---|---|---|---|---|---|---|
| | sAUC↑ | IG↑ | NSS↑ | CC↑ | AUC↑ | SIM↑ | KL↓ |
| GazeGAN [45] | 0.736 | 0.720 | 1.899 | 0.879 | 0.864 | 0.773 | 0.376 |
| SimpleNet [43] | 0.743 | 0.880 | 1.960 | **0.907** | **0.869** | 0.793 | **0.201** |
| EML-NET [78] | **0.746** | 0.736 | **2.050** | 0.886 | **0.866** | 0.780 | 0.520 |
| MSI-Net [76] | 0.736 | 0.793 | 1.931 | 0.889 | 0.865 | 0.784 | 0.307 |
| SAM-ResNet [18] | 0.741 | 0.538 | 1.990 | 0.899 | 0.865 | 0.793 | 0.610 |
| CEDNS [79] | 0.744 | 0.845 | **2.050** | 0.840 | 0.863 | 0.732 | - |
| DINet [19] | 0.739 | 0.195 | 1.959 | 0.902 | 0.862 | 0.795 | 0.864 |
| SalED [68] | 0.745 | **0.909** | 1.984 | **0.910** | **0.869** | **0.801** | **0.190** |
| SalFBNet [80] | 0.740 | 0.839 | 1.952 | 0.892 | **0.868** | 0.772 | 0.236 |
| TranSalNet [47] | **0.747** | - | **2.014** | **0.907** | **0.868** | **0.803** | 0.373 |
| DeepGaze II-E [81] | **0.767** | 0.766 | **1.996** | 0.872 | **0.869** | 0.733 | 0.285 |
| ACSalNet [84] | 0.744 | **0.890** | 1.981 | 0.905 | **0.868** | **0.798** | 0.232 |
| FastSal [85] | 0.732 | 0.770 | 1.845 | 0.874 | 0.863 | 0.768 | 0.288 |
| RINet | **0.746** | **0.913** | 1.982 | **0.911** | **0.869** | **0.803** | **0.189** |

TABLE II
QUANTITATIVE PERFORMANCE COMPARISON ON MIT300 BENCHMARK. THE UPPER PART IS FIVE CLASSICAL MODELS AND THE BOTTOM PART IS 18 DEEP LEARNING-BASED MODELS. THE BEST THREE RESULTS ARE MARKED IN RED, BLUE AND GREEN, RESPECTIVELY.

| Models | MIT300 | | | | | |
|---|---|---|---|---|---|---|
| | AUC↑ | sAUC↑ | NSS↑ | CC↑ | KL↓ | SIM↑ |
| ITTI [16] | 0.543 | 0.536 | 0.408 | 0.131 | 1.496 | 0.338 |
| GBVS [17] | 0.806 | 0.630 | 1.246 | 0.479 | 0.888 | 0.484 |
| AIM [27] | 0.762 | 0.665 | 0.882 | 0.342 | 1.248 | 0.410 |
| CAS [75] | 0.758 | 0.640 | 1.019 | 0.385 | 1.072 | 0.432 |
| SUN [29] | 0.694 | 0.626 | 0.762 | 0.277 | 1.282 | 0.393 |
| CASNet II [82] | 0.855 | 0.740 | 1.986 | 0.705 | 0.586 | 0.581 |
| SAM-ResNet [18] | 0.853 | 0.740 | 2.063 | 0.690 | 1.171 | 0.612 |
| DVA [42] | 0.843 | 0.726 | 1.931 | 0.663 | 0.629 | 0.585 |
| SalGAN [44] | 0.850 | 0.735 | 1.862 | 0.674 | 0.757 | 0.593 |
| DeepGaze I [36] | 0.843 | 0.723 | 1.723 | 0.614 | 0.668 | 0.572 |
| DeepGaze II [37] | 0.873 | 0.776 | 2.337 | 0.770 | 0.424 | 0.664 |
| ICF [37] | 0.833 | 0.696 | 1.613 | 0.588 | 0.708 | 0.558 |
| MLNet [41] | 0.839 | 0.740 | 1.975 | 0.663 | 0.801 | 0.582 |
| eDN [83] | 0.817 | 0.618 | 1.140 | 0.452 | 1.137 | 0.411 |
| GazeGAN [45] | 0.861 | 0.732 | 2.212 | 0.758 | 1.339 | 0.649 |
| EML-NET [78] | **0.876** | 0.747 | **2.488** | 0.789 | 0.844 | 0.676 |
| MSI-Net [76] | 0.874 | 0.779 | 2.305 | 0.779 | 0.423 | 0.670 |
| UNISAL [77] | **0.877** | **0.784** | 2.369 | 0.785 | **0.415** | 0.675 |
| SalFBNet [80] | **0.877** | **0.786** | **2.470** | **0.814** | **0.415** | **0.693** |
| TranSalNet [47] | 0.873 | 0.747 | **2.413** | **0.807** | 1.014 | **0.690** |
| ACSalNet [84] | 0.873 | 0.747 | 2.397 | 0.806 | **0.409** | 0.673 |
| SATSal [86] | 0.851 | 0.703 | 1.947 | 0.703 | 0.854 | 0.614 |
| FastSal [85] | 0.868 | 0.770 | 2.191 | 0.751 | 0.467 | 0.646 |
| RINet | **0.880** | **0.791** | 2.412 | **0.813** | **0.384** | **0.694** |

both NSS and other metrics. It can be seen that EML-NET and CEDNS get the highest NSS scores but obtain low scores on other metrics.

The results of MIT300 benchmark are reported in Tab. II. Our model still shows competitive performance on all metrics except NSS, which is consistent with the ones on SALICON benchmark. RINet outperforms all compared models on AUC, sAUC, SIM and KL. Specifically, our model outperforms the second best model by 6.11% on KL (0.409 → 0.384). As for the CC metric, the score RINet gets is very close to the best one.

We further evaluate our model on TORONTO, PASCAL-S and DUT-OMRON datasets. The results are reported in Tab. III. Our model achieves the best performance on TORONTO and gains competitive performance on PASCAL-S and DUT-OMRON datasets. Besides, our model consistently outperforms all other models on IG and KL. On TORONTO dataset, our model outperforms the second best model (ACSalNet) by a large margin such as 6.87% (1.004 → 1.073) on IG, and 11.40% (0.544 → 0.482) on KL. In terms of PASCAL-S dataset, compared to ACSalNet, our model improves by 5.23% (1.280 → 1.347) and 6.20% (0.726 → 0.681) on IG and KL, respectively. On the DUT-OMRON dataset, our model ranks first on IG and KL, and performs almost similar to the best model on CC and AUC. Among the models compared, TranSalNet, SalFBNet, FastSal and ACSalNet are the latest models so far, which further demonstrates the excellence of our model.

*2) Visual Comparison:* We conduct visual comparisons with the 11 representative FP models in Fig. 7 including five classical models (ITTI, GBVS, AIM, CAS, and SUN) and six deep learning-based models (SAM-ResNet, MSI-Net, GazeGAN, DINet, TranSalNet, and UNISAL). It can be observed that deep learning-based models surpass classical models overall. Furthermore, among deep learning-based models, our model can capture the relative importance relationship among multiple salient regions in a better way than competitors. For instance, as shown in the second column of Fig. 7, all the compared models highlight the face as a salient region, but ignore or make a mistaken emphasis on the salient region of hands. In contrast, RINet acquires the correct relative importance relationship between face and hands.

*3) Efficiency Comparison:* We also report the input size, model parameters and average processing time (APT) per image in Tab. III. The model parameters are measured by the online available codes or borrowed from their original articles. The APT is tested in our experimental platform. As for model parameters and APT, our model is at a medium level of computational efficiency. Although the HRIE module needs most of the computational consumption, it indeed realizes a substantial improvement in performance for RINet. Combining with the quantitative results, visual presentations and efficiency comparison, we can conclude that our model is very competitive in FP task.

### C. Ablation Analysis

In this section, we provide comprehensive ablation studies on SALICON validation set. Firstly, the contribution of each key component in our model is evaluated. Then we investigate the rationality of HRIE structure.

*1) The Contributions of Main Components:* We quantitatively evaluate the contributions of main components of our models, namely, HRIE module and CRF loss in Tab. IV. Firstly, the baseline is constructed, which directly passes the features extracted from backbone to the feature merging stage and is trained with $KL - CC$ as loss function simply. Next, the HRIE module is added over the baseline to evaluate the effectiveness of HRIE. Obviously, the HRIE module brings

TABLE III
QUANTITATIVE PERFORMANCE COMPARISON ON TORONTO, PASCAL-S AND DUT-OMRON DATASETS. WE ALSO REPORT THE INPUT SIZE, MODEL PARAMETERS (PARAMS) AND AVERAGE PROCESSING TIME (APT) PER IMAGE. THE UPPER PART IS FIVE CLASSICAL MODELS AND THE BOTTOM PART IS NINE DEEP LEARNING-BASED MODELS. * MEANS THAT WE FINE-TUNE THE MODEL ON THE MIT1003 DATASET. † INDICATES WE TRAIN THE MODEL ON SALICON DATASET FIRSTLY AND THEN FINE-TUNE IT ON THE MIT1003 DATASET. THE BEST THREE RESULTS ARE MARKED IN RED, BLUE AND GREEN, RESPECTIVELY.

| Models | Input Size | Params (M) | APT (ms) | TORONTO | | | | | | PASCAL-S | | | | DUT-OMRON | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | IG↑ | NSS↑ | CC↑ | AUC↑ | SIM↑ | KL↓ | IG↑ | CC↑ | AUC↑ | KL↓ | IG↑ | CC↑ | AUC↑ | KL↓ |
| ITTI [16] | - | - | - | -0.067 | 0.720 | 0.206 | 0.574 | 0.336 | 1.462 | -0.002 | 0.181 | 0.623 | 1.880 | 0.159 | 0.215 | 0.656 | 2.026 |
| GBVS [17] | - | - | - | 0.532 | 1.519 | 0.569 | 0.832 | 0.486 | 0.849 | 0.740 | 0.508 | 0.876 | 1.140 | 0.834 | 0.471 | 0.888 | 1.347 |
| AIM [27] | - | - | - | 0.102 | 0.835 | 0.312 | 0.768 | 0.368 | 1.444 | 0.187 | 0.297 | 0.810 | 1.679 | 0.309 | 0.282 | 0.845 | 1.904 |
| CAS [75] | - | - | - | 0.372 | 1.268 | 0.448 | 0.783 | 0.437 | 1.023 | 0.480 | 0.375 | 0.807 | 1.407 | 0.646 | 0.378 | 0.837 | 1.562 |
| SUN [29] | - | - | - | 0.015 | 0.723 | 0.242 | 0.688 | 0.357 | 1.458 | 0.091 | 0.218 | 0.702 | 1.789 | 0.296 | 0.254 | 0.764 | 1.923 |
| SAM-ResNet* [18] | 320×240 | 70.09 | 91 | -0.570 | 2.317 | 0.772 | 0.872 | 0.645 | 2.267 | 0.606 | **0.739** | 0.912 | 1.478 | 1.261 | 0.685 | 0.927 | 1.338 |
| MSI-Net [76] | 360×360 | 24.93 | 23 | 0.402 | 2.305 | 0.766 | **0.881** | 0.640 | 1.264 | 1.009 | 0.723 | **0.918** | 1.067 | 1.432 | 0.674 | **0.929** | 1.207 |
| GazeGAN [45] | 640×480 | 230.48 | 45 | -1.895 | 2.015 | 0.675 | 0.790 | 0.558 | 3.845 | -0.930 | 0.627 | 0.846 | 3.357 | -0.251 | 0.587 | 0.864 | 3.558 |
| DINet* [19] | 480×320 | 27.04 | 29 | -1.430 | **2.392** | **0.793** | 0.869 | 0.638 | 3.346 | -0.145 | **0.733** | 0.909 | 2.494 | 0.622 | **0.691** | 0.922 | 2.500 |
| TranSalNet* [47] | 384×288 | 76.56 | 43 | 0.649 | 2.005 | 0.744 | 0.879 | 0.509 | **0.753** | 0.966 | 0.721 | **0.918** | 0.955 | 1.047 | 0.644 | **0.930** | 1.196 |
| UNISAL [77] | 384×288 | 3.71 | 20 | 0.022 | **2.337** | **0.782** | **0.880** | **0.649** | 1.631 | 0.763 | 0.728 | **0.917** | 1.355 | **1.447** | **0.688** | **0.931** | 1.193 |
| SalFBNet [80] | 384×224 | 17.78 | 31 | 0.394 | 2.325 | 0.762 | 0.869 | 0.638 | 1.320 | 0.928 | 0.711 | 0.909 | 1.190 | 1.372 | 0.645 | 0.921 | 1.414 |
| FastSal* [85] | 256×192 | 2.57 | 8 | **0.659** | 1.788 | 0.684 | 0.857 | 0.590 | 0.783 | **1.045** | 0.650 | 0.904 | **0.873** | 1.182 | 0.591 | 0.912 | **1.096** |
| ACSalNet† [84] | 256×192 | 47.35 | 15 | **1.004** | 2.323 | **0.782** | 0.874 | **0.650** | **0.544** | **1.280** | **0.737** | 0.914 | **0.726** | **1.595** | **0.689** | 0.927 | **0.879** |
| RINet | 384×288 | 70.52 | 37 | **1.073** | **2.421** | **0.800** | **0.883** | **0.654** | **0.482** | **1.347** | 0.732 | **0.919** | **0.681** | **1.647** | **0.689** | **0.930** | **0.863** |

TABLE IV
ABLATION STUDY ON MAIN COMPONENTS OF THE PROPOSED MODEL ON SALICON VALIDATION SET AND TORONTO DATASET. THE BEST RESULT OF EACH METRIC IS SHOWN IN BOLD.

| | Models | IG ↑ | NSS ↑ | CC ↑ | AUC ↑ | SIM ↑ | KL ↓ |
|---|---|---|---|---|---|---|---|
| SALICON | Baseline | 0.9858 | 1.9343 | 0.9077 | 0.8718 | 0.7999 | 0.2007 |
| | + HRIE | 0.9941 | 1.9381 | 0.9103 | 0.8728 | 0.8032 | 0.1908 |
| | + CRF loss | 0.9909 | **1.9384** | 0.9088 | 0.8723 | 0.8014 | 0.1937 |
| | Ours | **0.9947** | 1.9370 | **0.9115** | **0.8730** | **0.8044** | **0.1867** |
| TORONTO | Baseline | 0.9665 | 2.2986 | 0.7783 | 0.8714 | 0.6256 | 0.5548 |
| | + HRIE | 1.0307 | 2.3635 | 0.7862 | 0.8821 | 0.6468 | 0.5145 |
| | + CRF loss | 1.0207 | 2.3183 | 0.7749 | 0.8821 | 0.6438 | 0.5302 |
| | Ours | **1.0729** | **2.4214** | **0.7998** | **0.8831** | **0.6541** | **0.4825** |

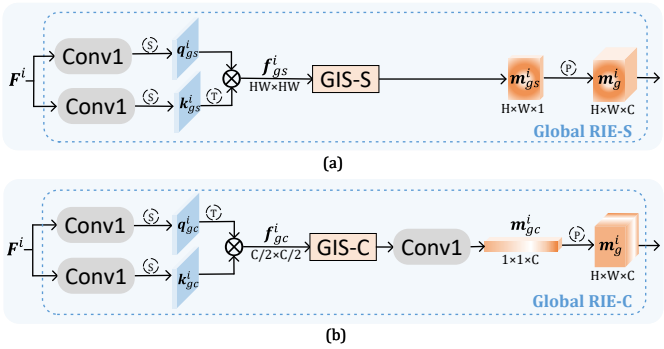



Fig. 8. Two variants of Global RIE module: (a) Global RIE-S. (b) Global RIE-C.

significant performance improvement on all evaluation metrics. In particular, on SALICON validation set, the percentage gain of HRIE reaches 0.84% (0.9858 → 0.9941) on IG and 4.93% (0.2007 → 0.1908) on KL. On TORONTO dataset, HRIE brings the percentage gain of 6.64% (0.9665 → 1.0307) on IG and 7.26% (0.5548 → 0.5145) on KL. This confirms our HRIE performs more efficiently than direct operation. In terms of CRF loss, we measure its contribution by training the baseline with the proposed CRF loss. As shown in Tab. IV, CRF loss also optimizes most metrics on both SALICON validation set and TORONTO dataset. In contrast, the contribution of HRIE is more obvious than CRF loss. Finally, we apply both HRIE and CRF loss to the baseline. The combination of HRIE and CRF loss further optimizes the performance. For example, the score of KL is promoted by 2.15% (0.1908 → 0.1867) compared to only adopting HRIE and 3.61% (0.1937 → 0.1867) compared to only adopting CRF loss on SALICON validation set. As for the results on TORONTO dataset, they are consistent with the ones on SALICON validation set. Overall, both HRIE and CRF loss play important roles in the proposed RINet. The combination of HRIE and CRF loss can better boost the performance of our model.

In addition, to demonstrate the effectiveness of HRIE, we visualize the intermediate feature maps $m^i_{gs}$ $(i \in \{2,3,4\})$ in Fig. 4. As mentioned in Sec. I, the scale variation of different salient regions is an obstacle for relative importance comparison. Fig. 4 shows HRIE can effectively handle the scale variation problem. The last three columns of Fig. 4 are the global relative importance map in spatial domain produced by RIE-2, RIE-3 and RIE-4, respectively, which show the deeper features recognize large salient regions, while the shallower features preserve the detailed information of small salient regions.

*2) Ablation Study on HRIE Module:* To validate the rationality of HRIE structure, we conduct in-depth ablation studies about global RIE, local RIE and hierarchical structure.

The RIE-$i$ block consists of a local RIE branch and a global RIE branch. To explore the effectiveness of local RIE and global RIE, we offer two variants of RIE-$i$ which remove local RIE and global RIE, respectively, denoted as *w/o* Local RIE and *w/o* Global RIE in Tab. V. The absence of Local RIE and Global RIE leads to performance degradation on all metrics except NSS, such as 1.29% (0.1867 → 0.1891) and 3.59% (0.1867 → 0.1934) on KL, respectively. The local RIE calculates local importance map for a certain feature embedding. It can be regarded that the local RIE uses the local importance map as customized global semantic information to update

this feature embedding, so as to benefit the performance. The global RIE normalizes and merges all local importance maps to score how important each feature embedding is. The results prove that emphasizing the important feature embeddings is also helpful. The global RIE branch is constructed in both spatial and channel domain. To assess the spatial domain and channel domain of global RIE, we provide two variants, namely Global RIE-S and Global RIE-C. Global RIE-S simply reserves the spatial domain by replacing the Global RIE branch with the variant in Fig. 8(a). Global RIE-C only keeps the channel domain by replacing the Global RIE branch with the variant in Fig. 8(b). As shown in Tab. V, the performance of both variants is degraded (*e.g.* KL: 0.1867 $\rightarrow$ 0.1889 of Global RIE-S and 0.1867 $\rightarrow$ 0.1893 of Global RIE-C), which proves that emphasizing the important feature embeddings in both spatial and channel domains is necessary.

The GIS mainly includes three operations, *i.e.* the Softmax function, the element-wise summation and min-max normalization. We design the ablation study to test whether these operations are reasonable. As for the Softmax function, we remove this operation in the variant GIS *w/o* Softmax in Tab. V. If without the Softmax function, the local importance maps with larger values will contribute more to the global importance map. In fact, only the value comparison within a certain local importance map is meaningful. Paying more attention to local importance maps with higher values is meaningless, and even impairs the performance. The results show the absence of Softmax function weakens our model. We design the Softmax function to equalize the contribution of each local importance map. As for the element-wise summation, it is utilized to fuse local importance maps. Theoretically, both element-wise summation and element-wise multiplication are suitable for fusion. However, the values of local importance maps after Softmax are generally less than $10^{-2}$. Multiplication of hundreds of such small values exceeds the lower limit of 32-bit float data, even 64-bit double data, which means it is infeasible actually. As a result, we choose the element-wise summation to merge local importance maps. As for the min-max normalization, we provide the model without this operation in the variant GIS *w/o* Min-max. If discarding min-max normalization, the higher value of H$\times$W, the higher value of global importance maps. The consequence is that the shallower features become more important, which does not obey Fig. 4. We use the min-max normalization to rescale global importance maps to a unified range. The experimental results of GIS *w/o* Min-max in Tab. V prove that min-max normalization is necessary.

To study the influence of hierarchical structure, we modify the number of RIE-$i$ and offer three variants as reported in the bottom part of Tab. V. By comparing the results of RIE-$3, 4$, RIE-$4$ and our model (*i.e.* RIE-$2, 3, 4$), we find that addition of the number of RIE-$i$ blocks brings continuous performance improvements (*e.g.*, IG: 0.9896 $\rightarrow$ 0.9939 $\rightarrow$ 0.9947, SIM: 0.8021 $\rightarrow$ 0.8022 $\rightarrow$ 0.8044). The reason behind this is that the semantic information and detailed information captured by the deeper block and shallower block, respectively, are complementary. For example, as depicted in the first row of Fig. 4, $m_{gs}^4$ focuses on the face while ignores the logo of laptop. $m_{gs}^2$ and $m_{gs}^3$ compensate for the negligence. Nonetheless, this trend disappears in the variant of RIE-$1, 2, 3, 4$ (*e.g.*, IG: 0.9947 $\rightarrow$ 0.9898, SIM: 0.8044 $\rightarrow$ 0.8007). This is because the shallower blocks may cause undesired details [47]. Consequently, we exclude RIE-$1$ from our model.

TABLE V
ABLATION STUDY ON HRIE MODULE ON SALICON VALIDATION SET. THE UPPER PART IS ABLATION ANALYSIS ABOUT LOCAL RIE AND GLOBAL RIE. THE MIDDLE PART IS ABLATION ANALYSIS ABOUT GIS OPERATION. THE BOTTOM PART IS ABLATION ANALYSIS ABOUT HIERARCHICAL STRUCTURE IN HRIE. THE BEST RESULT OF EACH COLUMN IS SHOWN IN BOLD.

| Models | IG ↑ | NSS ↑ | CC ↑ | AUC ↑ | SIM ↑ | KL ↓ |
|---|---|---|---|---|---|---|
| *w/o* Local RIE | 0.9945 | **1.9386** | 0.9097 | 0.8726 | 0.8024 | 0.1891 |
| *w/o* Global RIE | 0.9924 | 1.9372 | 0.9095 | 0.8726 | 0.8027 | 0.1934 |
| Global RIE-S | 0.9925 | 1.9345 | 0.9099 | 0.8726 | 0.8030 | 0.1889 |
| Global RIE-C | 0.9943 | 1.9381 | 0.9111 | 0.8728 | 0.8041 | 0.1893 |
| GIS *w/o* Softmax | 0.9935 | 1.9369 | 0.9095 | 0.8726 | 0.8028 | 0.1919 |
| GIS *w/o* Min-max | 0.9899 | 1.9315 | 0.9098 | 0.8724 | 0.8026 | 0.1908 |
| RIE-4 | 0.9896 | 1.9263 | 0.9095 | 0.8724 | 0.8021 | 0.1896 |
| RIE-3, 4 | 0.9939 | 1.9349 | 0.9092 | 0.8726 | 0.8022 | 0.1889 |
| RIE-1, 2, 3, 4 | 0.9898 | 1.9300 | 0.9079 | 0.8724 | 0.8007 | 0.1922 |
| Ours | **0.9947** | 1.9370 | **0.9115** | **0.8730** | **0.8044** | **0.1867** |

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT COMPONENTS OF LOSS FUNCTION ON SALICON VALIDATION SET. *w/* NSS REPRESENTS NSS IS INCLUDED AS PART OF LOSS FUNCTION DURING TRAINING. THE BEST RESULT OF EACH COLUMN IS SHOWN IN BOLD.

| Loss | IG↑ | NSS↑ | CC↑ | AUC↑ | SIM↑ | KL↓ |
|---|---|---|---|---|---|---|
| *w/* NSS | **1.0122** | **2.0106** | 0.8965 | **0.8736** | 0.7948 | 0.2019 |
| Ours | 0.9947 | 1.9370 | **0.9115** | 0.8730 | **0.8044** | **0.1867** |

## D. Loss Function Analysis

In this section, we offer numerous experiments to analyze the behavior of the loss function. Specifically, we observe 1) the influence of loss function components, 2) the influence of $\alpha$ and $\gamma$, and 3) the generalization of CRF loss.

*1) Influence of Loss Function Components:* To get a higher score on NSS, previous models usually exploit NSS as part of loss functions, such as EML-NET, CEDNS, SAM and TranSalNet. For further improving our performance on NSS, we include NSS as part of our loss function. The results in Tab. VI support that the addition of NSS can indeed improve location-based metrics (*i.e.*, IG, NSS and AUC), but it will lead to a sharp decline in the distribution-based metrics (*i.e.*, CC, SIM and KL). Our model trained without NSS as part of loss function achieves a better comprise between location-based metrics and distribution-based metrics.

*2) Influence of $\alpha$ and $\gamma$:* The CRF loss introduces two new hyperparameters, $\alpha$ in charge of scaling and $\gamma$ responsible for tuning the weighting speed. Our next attempt is to explore the influence of $\alpha$ and $\gamma$ on the CRF loss. Results for various $\alpha$ are shown in the upper part of Tab. VII. It can be seen $\alpha = 10$ works best. As for $\gamma$, when $\gamma$ increases, the warming-up stage is compressed as shown in Fig. 6. In other words, when $\gamma$ is larger than 100, the CRF loss approximates to

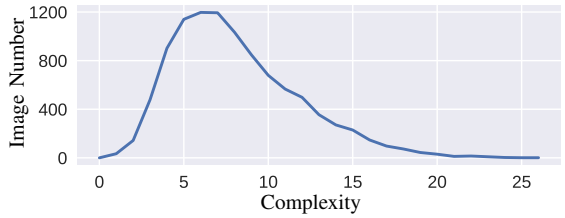



Fig. 9. The statistics about image number of different complexitys on SALICON training set.

TABLE VII
PERFORMANCE COMPARISON OF DIFFERENT $\alpha$ AND $\gamma$ ON SALICON VALIDATION SET. RESULTS FOR VARIOUS $\alpha$ ARE SHOWN IN THE UPPER PART WHILE RESULTS FOR VARIOUS $\gamma$ ARE SHOWN IN THE BOTTOM PART. THE BEST RESULTS OF $\alpha$ AND $\gamma$ ARE IN BOLD.

| Loss | IG↑ | NSS↑ | CC↑ | AUC↑ | SIM↑ | KL↓ |
|---|---|---|---|---|---|---|
| $\alpha$=2 | 0.9945 | 1.9325 | 0.9094 | 0.8728 | 0.8016 | 0.1882 |
| $\alpha$=8 | 0.9935 | 1.9360 | 0.9097 | 0.8727 | 0.8022 | 0.1904 |
| **$\alpha$=10** | **0.9947** | **1.9370** | **0.9115** | **0.8730** | **0.8044** | **0.1867** |
| $\alpha$=20 | 0.9920 | 1.9354 | 0.9099 | 0.8726 | 0.8026 | 0.1911 |
| $\alpha$=50 | 0.9913 | 1.9340 | 0.9088 | 0.8724 | 0.8013 | 0.1919 |
| **$\gamma$=5** | **0.9947** | **1.9370** | **0.9115** | **0.8730** | **0.8044** | **0.1867** |
| $\gamma$=10 | 0.9928 | 1.9336 | 0.9087 | 0.8725 | 0.8012 | 0.1887 |
| $\gamma$=100 | 0.9886 | 1.9325 | 0.9083 | 0.8723 | 0.8014 | 0.1964 |

TABLE VIII
PERFORMANCE COMPARISON BETWEEN ORIGINAL LOSS FUNCTION AND THE ONE OPTIMIZED BY OUR CRF LOSS ON SALICON VALIDATION SET, WHERE THE BETTER RESULT IS IN BOLD. * REPRESENTS TRAINING WITH THE LOSS FUNCTION OPTIMIZED BY CRF LOSS.

| Loss | IG↑ | NSS↑ | CC↑ | AUC↑ | SIM↑ | KL↓ |
|---|---|---|---|---|---|---|
| TVdist | 0.6754 | 1.9380 | 0.9088 | 0.8693 | 0.8029 | 0.6292 |
| TVdist* | **0.7332** | **1.9383** | **0.9103** | **0.8702** | **0.8054** | **0.5515** |
| SAM loss | 0.9954 | **1.9785** | 0.9056 | 0.8726 | 0.8003 | 0.2083 |
| SAM loss* | **0.9965** | 1.9599 | **0.9086** | **0.8728** | **0.8019** | **0.1944** |

weight the difficult samples directly. The training procedure always breaks down. Besides, we note that smaller $\gamma$ (*i.e.* a persistent warming-up stage) results in better performance. This is because, statistically, simpler samples predominate in the dataset as shown in Fig. 9. The focusing factor $\lambda$ treats all samples equally in the beginning so that the model tends to handle the simpler samples. Gradually, the focusing factor $\lambda$ changes emphasis to complicated scenes such that the model can master the knowledge from easy to difficult, which benefits the performance.

*3) Generalization of CRF loss:* CRF loss promotes the performance of RINet as shown in Tab. IV (Ours *vs.* + HRIE). To evaluate its generalization capability, we introduce other loss functions used in FP task into our RINet, including the total variation distance (TVdist) of DINet [19] and SAM loss [18] (*i.e.*, 10·KL−2·CC−NSS), and extend our CRF loss to them. As shown in Tab. VIII, we can observe the performance basically increases with the help of CRF loss, particularly KL and IG (*e.g.* KL: 12.35% on TVdist, 6.67% on SAM loss). This confirms our CRF loss can be generalized to other loss functions and promote the performance.

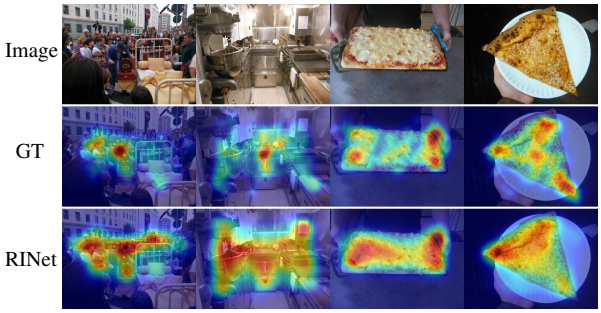



Fig. 10. Some failure cases of our RINet.

### E. Failure Cases and Analysis

In this section, we illustrate two types of failure cases of RINet in Fig. 10 and point out the possible directions of future work. The first two columns represent one type of failure cases. When there is no very salient region in the image, the fixation density map tends to have a center bias. Our model only grasps this knowledge implicitly during the training procedure, which might be not enough to handle the knowledge. Adding the center bias as an explicit prior like [36], [37] is a possible solution. The other way is utilizing position coding to let the feature get the position information. The second type of failure cases is shown in the last two columns. When a relatively large object fills the field of vision and there is no obvious difference between its local patches, our model fails to accurately highlight the salient regions. In this failure case, the corners of the object seem to attract attention. We think combining FP with the task of instance segmentation might be beneficial. Instance segmentation can point out how large the object is and where the boundary of the object is, which provides useful information to solve this problem.

### F. Applications and Discussion

In this article, we try to make the model understand the relative importance relationship better. This relationship plays an important role in not only the FP task, but also many tasks in the vision community. We have discussed the relative importance relationship in spatial and channel domains in Sec. III-B. The proposed RIE module can be employed directly as a plug-and-play module to handle the relative importance relationship in spatial and channel domains, so as to benefit the related tasks such as salient object detection [87], [88], co-saliency detection [89], [90] and ROI extraction in medical images [91], [92]. Beyond the spatial and channel domains, the RIE module can also be generalized to master the relative importance relationship in modality domain (*i.e.*, multi-modality magnetic resonance images [93], [94] and RGB-D/RGB-T images [95], [96]), time domain (*i.e.*, video [97]), and feature domain (*i.e.*, multi-level features [98]). We just need to adjust the GIS operation to compare the features of different modalities, different frames, or different levels of the network.

Besides supervised learning, relative importance is also beneficial to scribble-supervised learning. For instance, for the

task of scribble-supervised video object segmentation [99], our RIE module can highlight the more important regions as extra knowledge to make up for the coarse scribble annotation.

## V. Conclusion

In this article, we have proposed a novel and effective relative importance-aware network (RINet) for FP. The relative importance relationship is captured by the well-designed hierarchical relative importance enhancement (HRIE) module, which generates local-global relative importance maps at different scale subspaces and takes relative importance of both spatial and channel domains into account. Besides, samples of difficult scenes are emphasized by complexity-relevant focal (CRF) loss. The experimental results of loss function analysis show that CRF loss not only brings performance gains in the loss function used in this article, but also can be generalized to other loss functions for FP. Comprehensive experimental results, including quantitative comparison and visualization analysis, have proved the outstanding performance of RINet with respect to 28 state-of-the-art FP models.

## References


[1] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*, 1987, pp. 115–141.

[2] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proc. IEEE CVPR*, Jun. 2018, pp. 1711–1720.

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

[4] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Mar. 2016.

[5] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, and H. Sun, "Optimizing multistage discriminative dictionaries for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2035–2048, Oct. 2017.

[6] H. Kim and S. Lee, "Transition of visual attention assessment in stereoscopic images with evaluation of subjective visual quality and discomfort," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2198–2209, Oct. 2015.

[7] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Jun. 2013.

[8] Z. Wang, Z. Liu, G. Li, Y. Wang, T. Zhang, L. Xu, and J. Wang, "Spatio-temporal self-attention network for video saliency prediction," *IEEE Trans. Multimedia*, 2021.

[9] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1688–1698, Nov. 2018.

[10] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Jun. 2013.

[11] G. Li, Z. Liu, R. Shi, and W. Wei, "Constrained fixation point based segmentation via deep neural network," *Neurocomputing*, vol. 368, pp. 180–187, Nov. 2019.

[12] G. Li, Z. Liu, R. Shi, Z. Hu, W. Wei, Y. Wu, M. Huang, and H. Ling, "Personal fixations-based object segmentation with object localization and boundary preservation," *IEEE Trans. Image Process.*, vol. 30, pp. 1461–1475, Dec. 2020.

[13] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2017.

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

[15] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE ICCV*, Sept. 2009, pp. 2106–2113.

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

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NeurIPS*, vol. 19, 2006, pp. 545–552.

[18] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.

[19] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2163–2176, Aug. 2020.

[20] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE CVPR*, Jun. 2018, pp. 7794–7803.

[21] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[22] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[23] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 679–700, Feb. 2021.

[24] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, Oct. 1995.

[25] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[26] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. IEEE CVPR*, Jun. 2011, pp. 433–440.

[27] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. NeurIPS*, vol. 18, 2005, pp. 155–162.

[28] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. NeurIPS*, vol. 21, 2008, pp. 681–688.

[29] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 32–32, Dec. 2008.

[30] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Proc. NeurIPS*, vol. 18, 2005, pp. 547–554.

[31] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proc. NeurIPS*, vol. 17, 2004, pp. 481–488.

[32] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 989–1005, Jun. 2009.

[33] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.

[34] P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency," in *Proc. International Conference Advances in Neuro-Information Processing*, 2008, pp. 251–258.

[35] W. Kienzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann, "Center-surround patterns emerge as optimal predictors for human saccade targets," *J. Vis.*, vol. 9, no. 5, pp. 1–15, May 2009.

[36] M. Kümmerer, L. Theis, and M. Bethge, "Deep Gaze I: boosting saliency prediction with feature maps trained on imagenet," *arXiv preprint arXiv:1411.1045*, 2014.

[37] M. Kümmerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *Proc. IEEE ICCV*, Oct. 2017, pp. 4789–4798.

[38] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE CVPR*, Jun. 2015, pp. 362–370.

[39] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A fully convolutional neural network for predicting human eye fixations," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4446–4456, Sept. 2017.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[41] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. IEEE ICPR*, Dec. 2016, pp. 3488–3493.

[42] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2017.

[43] N. Reddy, S. Jain, P. Yarlagadda, and V. Gandhi, "Tidying deep saliency prediction architectures," in *Proc. IEEE IROS*, Oct. 2020, pp. 10 241–10 247.

[44] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[45] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is gaze influenced by image transformations? Dataset and model," *IEEE Trans. Image Process.*, vol. 29, pp. 2287–2300, Oct. 2019.

[46] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3264–3274, Mar. 2018.

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

[48] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *arXiv preprint arXiv:2111.07624*, 2021.

[49] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Apr. 2020.

[50] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[51] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Oct. 2018, pp. 3–19.

[52] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017, pp. 5998–6008.

[54] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3146–3154.

[55] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE ICCV*, Oct. 2019, pp. 603–612.

[56] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE ICCV*, Oct. 2021, pp. 10 012–10 022.

[57] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Nov. 2015, pp. 234–241.

[59] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, Oct. 2017, pp. 2980–2988.

[60] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proc. IEEE ICCV*, Oct. 2021, pp. 82–91.

[61] B. Li, Y. Yao, J. Tan, G. Zhang, F. Yu, J. Lu, and Y. Luo, "Equalized focal loss for dense long-tailed object detection," *arXiv preprint arXiv:2201.02593*, 2022.

[62] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. NeurIPS*, vol. 33, 2020, pp. 21 002–21 012.

[63] X. Li, W. Wang, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection," in *Proc. IEEE CVPR*, Jun. 2021, pp. 11 632–11 641.

[64] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 4700–4708.

[65] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*, 2015, pp. 1180–1189.

[66] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, vol. 32, 2019, pp. 8024–8035.

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NeurIPS*, vol. 25, 2012, pp. 1106–1114.

[68] Z. Wang, Z. Liu, W. Wei, and H. Duan, "SalED: Saliency prediction with a pithy encoder-decoder architecture sensing local and global information," *Image Vis. Comput.*, vol. 109, p. 104149, May 2021.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[70] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations." MIT Comput. Sci. Artif. Intell. Lab., Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 2012.

[71] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. NeurIPS*, vol. 18, 2005, pp. 155–162.

[72] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE CVPR*, Jun. 2014, pp. 280–287.

[73] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3166–3173.

[74] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2018.

[75] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2011.

[76] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," *Neural Networks*, vol. 129, pp. 261–270, Sept. 2020.

[77] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Proc. ECCV*, Oct. 2020, pp. 419–435.

[78] S. Jia and N. D. Bruce, "EML-NET: an expandable multi-layer network for saliency prediction," *Image Vis. Comput.*, vol. 95, p. 103887, Mar. 2020.

[79] F. Qi, C. Lin, G. Shi, and H. Li, "A convolutional encoder-decoder network with skip connections for saliency prediction," *IEEE Access*, vol. 7, pp. 60 428–60 438, May 2019.

[80] G. Ding, N. İmamoğlu, A. Caglayan, M. Murakawa, and R. Nakamura, "SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks," *Image Vis. Comput.*, vol. 120, p. 104395, Feb. 2022.

[81] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, "DeepGaze IIE: calibrated prediction in and out-of-domain for state-of-the-art saliency modeling," in *Proc. IEEE ICCV*, Oct. 2021, pp. 12 919–12 928.

[82] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proc. IEEE CVPR*, Jun. 2018, pp. 7521–7531.

[83] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE CVPR*, Jun. 2014, pp. 2798–2805.

[84] C. Qing, H. Zhu, X. Xing, D. Chen, and J. Jin, "Attentive and context-aware deep network for saliency prediction on omni-directional images," *Digit. Signal Process.*, vol. 120, p. 103289, Jan. 2022.

[85] F. Hu and K. McGuinness, "FastSal: a computationally efficient network for visual saliency prediction," in *Proc. IEEE ICPR*, Jan. 2021, pp. 9054–9061.

[86] M. Tliba, M. A. Kerkouri, B. Ghariba, A. Chetouani, A. Çöltekin, M. S. Shehata, and A. Bruno, "SATSal: A multi-level self-attention based architecture for visual saliency prediction," *IEEE Access*, vol. 10, pp. 20 701–20 713, Oct. 2022.

[87] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, Mar. 2020.

[88] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3528–3542, Mar. 2021.

[89] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, pp. 38:1–38:31, Jan. 2018.

[90] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, May 2017.

[91] C. Fang, D. Zhang, L. Wang, Y. Zhang, L. Cheng, and J. Han, "Cross-modality high-frequency transformer for MR image super-resolution," in *Proc. ACM MM*, Oct. 2022, pp. 1584–1592.

[92] Y. Yang, S. Wei, D. Zhang, Q. Yan, S. Zhao, and J. Han, "Hierarchical and global modality interaction for brain tumor segmentation," in *Proc. MICCAI*, vol. 12962, Jul. 2022, pp. 441–450.

[93] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, "Cross-modality deep feature learning for brain tumor segmentation," *Pattern Recognit.*, vol. 110, p. 107562, Feb. 2021.

[94] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Wang, and Y. Yu, "Exploring task structure for brain tumor segmentation from multi-modality MR images," *IEEE Trans. Image Process.*, vol. 29, pp. 9032–9043, Sep. 2020.

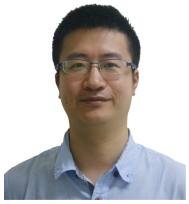
[95] N. Huang, Y. Yang, D. Zhang, Q. Zhang, and J. Han, "Employing bilinear fusion and saliency prior information for RGB-D salient object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1651–1664, Mar. 2021.
[96] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, Dec. 2019.
[97] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *Proc. IEEE CVPR*, Jun. 2018, pp. 9080–9089.
[98] G. Li, Z. Liu, D. Z. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 526–538, Jan. 2023.
[99] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, "Scribble-supervised video object segmentation," *IEEE CAA J. Autom. Sinica*, vol. 9, no. 2, pp. 339–353, Feb. 2022.

**Yingjie Song** received the B.E. degree from Shanghai University, Shanghai, China, in 2019. She is currently pursuing the Ph.D. degree at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. Her research interests include fixation prediction and autism spectrum disorder/schizophrenia identification.

**Zhi Liu** (M'07-SM'15) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 1999, 2002 and 2005, respectively. He is currently a Professor at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He is an Area Editor of *Signal Processing: Image Communication* and served as a Guest Editor for the special issue on *Recent Advances in Saliency Models, Applications and Evaluations* in *Signal Processing: Image Communication*.

**Gongyang Li** received the Ph.D. degree from Shanghai University, Shanghai, China, in 2022. From July 2021 to June 2022, he was a Visiting Ph.D. Student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently a Postdoctoral Fellow with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include image/video object segmentation, semantic segmentation, and saliency detection.

**Dan Zeng** (Senior Member, IEEE) received her Ph.D. degree in circuits and systems, and her B.S. degree in electronic science and technology, both from University of Science and Technology of China, Hefei. She is a full professor and the Dean of the Department of Communication Engineering and the Computer Vision and Pattern Recognition Lab at Shanghai University. Her main research interests include computer vision, multimedia analysis, and machine learning. She is serving as the Associate Editor of *IEEE Transactions on Multimedia*, the Associate Editor of the *IEEE Transactions on Circuits and Systems for Video Technology*, the TC Member of IEEE MSA and TC member of IEEE MMSP.



**Tianhong Zhang** MD. PhD. Senior Psychiatrist, Director of Early Identification and Intervention for Clinical High Risk of Psychosis, Clinical PI for SHARP (ShangHai At Risk for Psychosis) program, Shanghai Mental Health Centre, Master's tutor in Shanghai Jiaotong University School of Medicine. Adjunct Professor of Medical College, University of Ottawa, Secretary and Member of CSNP Schizophrenia Research Alliance, Member of Schizophrenia Collaborative Group of Psychiatric Society of Chinese Medical Association, BMC psychiatry, Psychiatry Research Editor, Frontier in Psychiatry Guest Editor.

**Lihua Xu** MD. PhD. Psychiatrist, Attending Doctor. Her research direction is the biomarker research of high risk syndrome of psychosis. To be responsible for the clinical evaluation, follow-up and data management of high risk syndrome of psychosis.

**Jijun Wang** MD. PhD. Chief Physician, Doctoral Supervisor of Shanghai Mental Health Center (mental health center affiliated to Shanghai Jiaotong University School of Medicine), director of brain film image eye movement research office, PI of Shanghai heavy mental disease laboratory, member of psychiatry basic and clinical branch of Chinese Neuroscience Society, member of EEG and EMG branch of Shanghai Medical Association, director of Youth Committee of Shanghai Overseas Chinese joint committee. As editor of the Journal of psychiatry, BMC psychiatry and reviewer of various international journals (Cochrane Database Syst. Rev., Biological Psychology, International Journal of Psychiatry, etc.).