TCGNet: Type-Correlation Guidance for Salient Object Detection

Yi Liu, Ling Zhou, Gengshen Wu, *Member, IEEE*, Shoukun Xu, and Jungong Han, *Senior Member, IEEE*

***Abstract*— Contrast and part-whole relations induced by deep neural networks like Convolutional Neural Networks (CNNs) and Capsule Networks (CapsNets) have been known as two types of semantic cues for deep salient object detection. However, few works pay attention to their complementary properties in the context of saliency prediction. In this paper, we probe into this issue and propose a Type-Correlation Guidance Network (TCGNet) for salient object detection. Specifically, a Multi-Type Cue Correlation (MTCC) covering CNNs and CapsNets is designed to extract the contrast and part-whole relational semantics, respectively. Using MTCC, two correlation matrices containing complementary information are computed with these two types of semantics. In return, these correlation matrices are used to guide the learning of the above semantics to generate better saliency cues. Besides, a Type Interaction Attention (TIA) is developed to interact semantics from CNNs and CapsNets for the aim of saliency prediction. Experiments and analysis on five benchmarks show the superiority of the proposed approach. Codes has been released on https://github.com/liuyi1989/TCGNet.**

***Index Terms*— Salient object detection, part-object relationship, capsule network.**

I. Introduction

THE task of salient object detection imitates the human visual perception to automatically identify and segment attractive regions or objects. It can help capture the informative regions that contain the main scene semantics. On account of its power, salient object detection has served for main scene understandings, including autonomous driving perception [1], [2], [3], image retrieval [4], video segmentation [5], image cropping [6], semantic segmentation [7], and object recognition [8]. For example, in an autonomous driving vision system, salient object detection can rapidly allocate the attention on the important objects for scene parsing [1], [2], [9], [10], [11]. The earlier salient object detection methods mostly extract the hand-crafted features to mine the contrast regions [12]. The development of deep learning has greatly broken the bottleneck [13] of hand-crafted approaches and will continue to bring steady progress.

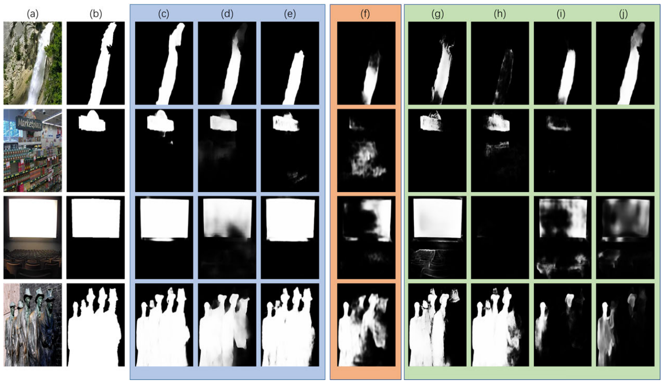


Fig. 1. Problem statements. (a) Image; (b) GT; (c) Ours; (d) POCINet [18]; (e) PWHCNet [19]; (f) TSPOANet [20]; (g) ITSD [14]; (h) MINet [15]; (i) GateNet [16]; (j) EGNet [17]. The green, orange, and blue domains include the contrast methods, the part-whole relational method, and the methods combining contrast and part-whole relations, respectively.

The deep learning based salient object detection methods mostly rely on deep neural networks, especially Convolutional Neural Networks (CNNs), to extract the discriminative features and find the salient regions with high contrast over their surroundings [13]. They have a genius for capturing the object details. However, these methods compute the salient regions within an image individually without considering the inter-region relations, thereby damaging the object's wholeness. For this reason, the performance of previous contrast-based salient object detectors heavily compromises when handling real-world complex structures. As can be seen from Fig. 1, the deep contrast saliency methods, including ITSD [14], MINet [15], GateNet [16], EGNet [17], cannot produce good results on complicated scenes, and mostly fail to achieve the object completeness.

Alternative to the contrast pipeline of deep salient object detectors, our previous works et al. [20], [21] put forward the pipeline of part-whole relations for salient object detection

Manuscript received 21 May 2023; revised 23 October 2023; accepted 10 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62001341; in part by the National Natural Science Foundation of Jiangsu Province under Grant BK20221379; in part by the State Key Laboratory of Reliability and Intelligence of Electrical Equipment under Grant EERI KF2022005; in part by the CNPC-CZU Innovation Alliance, Changzhou University, under Grant CCIA2023-01; and in part by the Changzhou Leading Innovative Talent Introduction & Cultivation Project 20221460. The Associate Editor for this article was Q. Ye. *(Corresponding authors: Shoukun Xu; Jungong Han.)*

Yi Liu is with the School of Computer Science and Artificial Intelligence, the Aliyun School of Big Data, the School of Software, and the CNPC-CZU Innovation Alliance, Changzhou University, Changzhou, Jiangsu 213164, China (e-mail: liuyi0089@gmail.com).

Ling Zhou and Shoukun Xu are with the School of Computer Science and Artificial Intelligence, the Aliyun School of Big Data, and the School of Software, Changzhou University, Changzhou, Jiangsu 213000, China (e-mail: 20497132@smail.cczu.edu.cn; jpuxsk@163.com).

Gengshen Wu is with the Faculty of Data Science, City University of Macau, Macau 999078, China (e-mail: gengshenw@gmail.com).

Jungong Han is with the Department of Computer Science, University of Sheffield, S10 2TN Sheffield, U.K. (e-mail: jungonghan77@gmail.com).

Digital Object Identifier 10.1109/TITS.2023.3342811

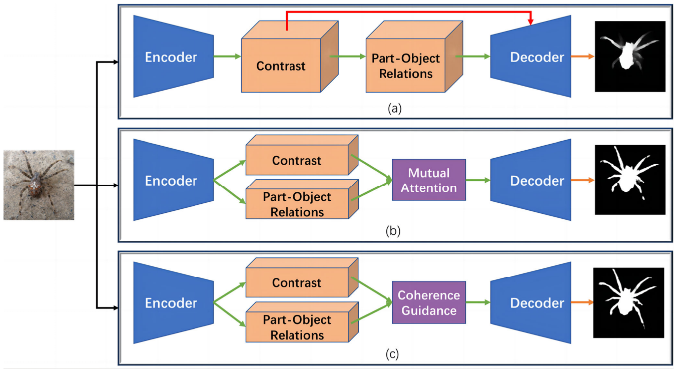


Fig. 2. Illustration for different interactions between contrast cues and part-whole relationships. (a) POCINet [18], (b) PWHCNet [19], (c) Ours. POCINet [18] combines these two cues in the decoder module. PWHCNet [19] adopts the attention mechanism to integrate them before the decoder module. Alternatively, we calculate the correlation to interact these two types of semantics.

endowed by the Capsule Networks (CapsNets) [22], which can well detect the object's wholeness. Later on, other attempts, *e.g.* ICON [23] and TSPORTNet [21], employ the CapsNets' semantics as guidance to learn primitive features for saliency inference. However, using part-whole relations alone will lead to the loss of object details in complicated scenes, which can be observed in Fig. 1 that the deep part-whole relational saliency method, *i.e.*, TSPOANet [20], misses some inner object details. In essence, contrast and part-whole relations capture different semantics for the same salient object, *e.g.*, contrast is able to highlight the local details while part-whole relations are good at capturing the object's wholeness. Clearly, these two saliency cues can complement to each other for better salient object detection. To this end, there are some attempts in the progress. As shown in Fig. 2, PWHCNet [19] proposes a mutual attention mechanism to integrate contrast and part-whole relations to detect the more accurate salient object. POCINet [18] combines these two cues in an upsampling module to tackle the problem of camouflaged object detection. Despite its preliminary success, it is still in its infancy, and there are still some issues to be solved. For instance, due to the lack of guidance after fusing contrast cues and part-whole relations, PWHCNet [19] cannot ensure the integrity of salient objects in the results. On the other hand, POCINet [18] always causes a blurry border between foreground and background because of the neglect of the relevance of these two cues.

In this paper, we delve into the coherence between these two types of cues, and propose a type-correlation aware mechanism to enhance their representation power, as shown in Fig. 2(c). Specifically, we propose a Multi-Type Cues Correlation (MTCC) module, in which the spatial correlation matrices, including width and height correlations, dare computed. In doing so, we intend to excavate the intrinsic relations of the two saliency cues. On top of that, these two correlation matrices are utilized to guide the intermediate contrast and part-whole relational saliency results to generate more accurate saliency priors, which are in return sreved as guidance to learn better contrast and part-whole relational features for saliency prediction. Besides, in the decoder, we develop a Type Interaction Attention (TIA) to interact the semantics of CNNs and CapsNets for saliency prediction, in which the CNNs map is activated to be interacted with the CapsNets map. Experiments indicate that our pipeline can adequately engage contrast and part-whole relations for the task of salient object detection, as can be shown in Fig. 1.

To sum up, the contributions of the paper can be described as follows:

- In this work, we propose a novel framework termed Type-Correlation Guidance Network (TCGNet) in the salient object detection task. The proposed network digs into the coherence between contrast and part-whole relational saliency cues, thus improving the detection performance via enhancing the representation power of two cues simultaneously. To the best of our knowledge, this is the first attempt to employ such type of coherence in deep salient object detection.
- A novel Multi-Type Cues Correlation (MTCC) module is proposed to obtain better saliency priors by integrating the correlations of contrast and part-whole relational cues from CNNs and CapsNets, thus improving the saliency prediction performance.
- A novel Type Interaction Attention (TIA) mechanism in the decoder is proposed to let the CNNs and CapsNets maps interact and guide the generation of saliency maps, which has been proven to improve performance of salient object detection.
- Extensive experiments on five datasets show that the proposed TCGNet can achieve superior performance, compared to the state-of-the-art baselines, which further consolidates our contributions.

The paper is organized as follows. Sec. II reviews the related works. Sec. III describes the details of the proposed method. Sec. IV evaluates and analyzes for understanding the proposed method. Sec. V concludes the paper.

II. Related Work

In this section, we will review the most related works, including CNNs for salient object detection and CapsNets for salient object detection.

A. CNNs for Salient Object Detection

Deep CNNs have achieved a significant breakthrough in the task of salient object detection [24], [25], [26], [27], [28], [29]. At the beginning, an Multi-Layer Perceptron (MLP) is employed to predict foreground and background. Zhao et al. [30] used two pathways to extract local and global context, which was fed into an MLP for foreground and background classification. Wang et al. [31] relied on an MLP to predict saliency scores from deep segment-level features. Later on, the Fully Convolutional Network (FCN) is adopted to solve the problem of salient object detection. Luo et al. [26] combined deep local and global cues for saliency detection. Wang et al. [32] recurrently refined the saliency prediction from heuristic calculation or prediction of previous time step. Wang et al. [33] used a stage-wise manner to implement the

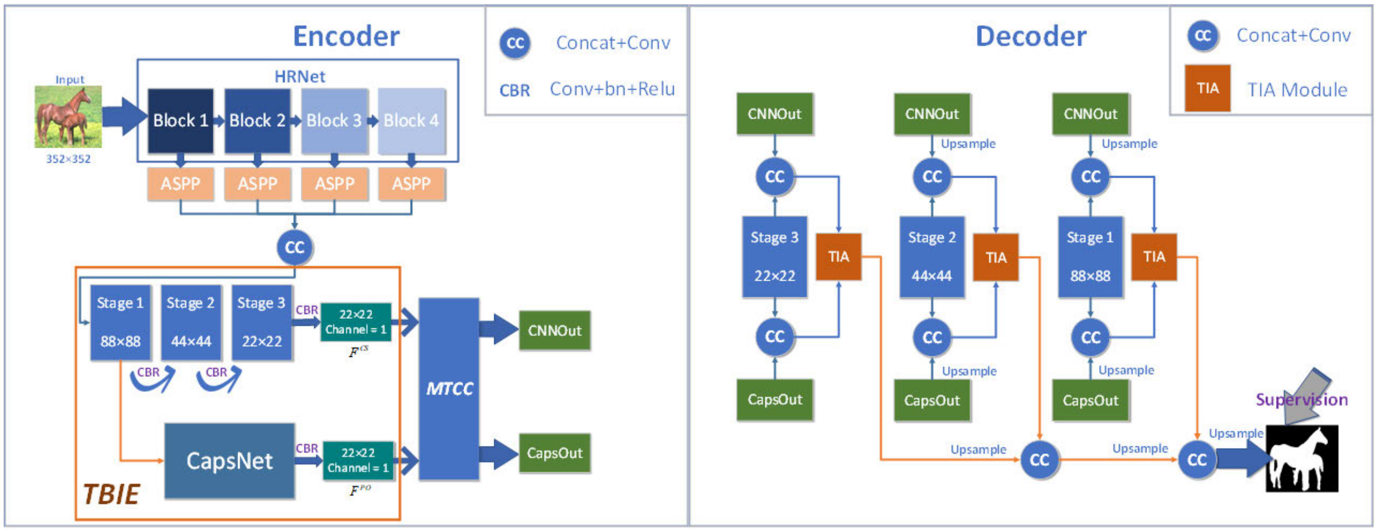


Fig. 3. Overview of our type-correlation guidance network (TCGNet), which consists of an encoder and a decoder. In the encoder, given a three-channel RGB image, it is fed into an HRNet to extract multi-scale features, which are further to learn rich context information with an ASPP module. After that, feature maps are fed into a TBIE module to get contrast cues and part-object relationships, which will be subject to correlation analysis in the MTCC module. Besides, in the decoder, the obtained contrast cues and part-object relationships will be integrated using a TIA module. Finally, a layer-by-layer manner is employed for upsampling the saliency cues to get the output saliency map.

coarse-to-fine saliency refinement. Liu and Han [34] combined shallower features using recurrent layers with intermediate deep supervision for saliency prediction. Su et al. [35] solved the selectivity-invariance dilemma problem of the salient object detection task with multiple branches. At the stage of success of MLP and FCN for salient object detection, they are combined into a unified framework for salient object detection with the aim of producing edge-preserving detection using multi-scale context. Tang and Hu [36] integrated pixel-level generated by FCN and super-pixel-level saliency for the final saliency prediction. Feng et al. [28] detected object boundaries via an attentive feedback network for better salient object detection. Recently, besides the need for different networks, many works attempt to solve real-life requirements for salient object detection, *e.g.*, real time and semantics understanding. For instance, Zhou et al. [14] constructed a two-branch decoder and interact with them to generate fast saliency. Liu et al. [37] explored the potential of pooling for real-time salient object detection. Cheng et al. [38] analyzed the semantic information of CNNs based salient object detection models. Wang et al. [39] identified the salient object with the guidance of human fixation. Wang et al. [40] exploited the pyramid attention to focus on the salient regions and salient edges detection to refine the object boundaries. Wang et al. [41] learned top-down and bottom-up inference for saliency prediction. Ke et al. [42] designed a contour-saliency network with the purpose of enhancing the edge quality of the salient object. Wang et al. [43] adopted the boundary sensibility, content integrity, iterative refinement, and frequency decomposition to enhance the performance for salient object detection. Lee et al. [44] excluded multi-decoder structures and minimized the learning parameters usage for a computationally efficient salient object detector using the attention guided tracing modules. Wu et al. [45] explored the high-level feature learning for locating salient objects via an intuitive extreme downsampling technique. Ma et al. [46] improved the performance of salient object detection with broader receptive fields. Jiao et al. [47] devised collaborative content-dependent networks to find the discriminative objects with a global context. More reviews about CNNs based salient object detection can be referred to [13].

Different from these approaches that rely on the discriminative contrast cues of CNNs for salient object detection, our method involves the part-whole relations explored by CapsNets to augment the salient object detection performance.

B. CapsNets for Salient Object Detection

While CNNs-based salient object detection methods have achieved breakthrough performance, they still encounter issues. For example, CNNs mostly infer the saliency of each region separately, which will cause failure in object wholeness. To address this problem, our previous work [20] introduced CapsNets [22], which can capture the spatial structures between different object parts, for the task of saliency prediction, resulting in the task of part-whole visual saliency. Instead of directly using CapsNets for saliency prediction, we proposed a two-stream strategy to reduce the complexity of CapsNets to tackle the dense salient object prediction. Later, we consolidated our work with a correlation-aware capsule routing for network training. On top of that, several efforts are devoted to advocate CapsNets-based part-whole visual saliency [21], [48]. To solve the heavy computation of the part-whole relational saliency, Liu et al. [49] disentangled the horizontal and vertical capsule routing within the capsule routing algorithm for fast saliency prediction. Besides, a few works have been devoted to the complementary of CNNs and CapsNets. For example, [19] involved an attention mechanism to interact CNNs features and CapsNets features for better salient object detection. Reference [23] used part-whole verification to judge whether the part and whole objects are related. Reference [50] designed a multi-scale capsule-wise attention to aggregate features and generate fine-grained prediction maps. Liu et al. [18] integrated the CapsNets semantics and CNNs features to detect the eye-attracting objects in the concealed scene.

Different from the existing saliency detection methods involving CapsNets and CNNs, we design a new interactive mechanism for these two types of information. Specifically, we compute the correlations between contrast cues from CNNs and part-whole relational cues from CapsNets as the saliency priors to learn better saliency cues. To this end, we develop a novel attention mechanism to involve these two-type semantics to infer the saliency.

III. Methodology

In this section, we will describe the details of the proposed method.

A. Overview

The overview of the proposed TCGNet is illustrated in Fig. 3. Given a three-channel RGB image, it is fed into an HRNet [51] backbone to extract multi-scale and high-resolution features, which are further fed into an ASPP module [52] with different dilation rates (1, 6, 12 and 18) to capture rich context information. Then we fuse all of the feature maps to obtain the integrated feature maps (88 × 88 × 128), which are rich in both fine details and semantic knowledge. Afterwards, the feature maps are sent into the Two-Branch Information Extraction (TBIE) module to grab contrast cues and part-whole relationships, which will be subject to correlation analysis in the Multi-Type Cues Correlation (MTCC) module. In our decoder, the guided contrast cues and part-whole relationships will be integrated in a Type Integration Attention (TIA). Finally, a layer-by-layer manner is employed for upsampling the decoded saliency cues to achieve the final output results.

B. Two-Branch Information Extraction (TBIE)

Fig. 4 details the architecture of TBIE, which is composed of two parallel branches, including the CNNs branch for contrast cues extraction and the CapsNets branch for part-whole relational cues exploration.

1) CNNs Branch: CNNs branch is composed of three stages with the same structure, each of which contains one convolution layer and ReLU. Each stage of CNN branch is formulated as follows

$$\mathbf{F}_{out} = ReLU(BN(Conv(\mathbf{F}_{in}))), \tag{1}$$

where $\mathbf{F}_{in}$ and $\mathbf{F}_{out}$ represent the input and output of the convolution stage, respectively. $Conv$ means the convolution operation. The convolutions in the stage 1 and stage 2 adopt a 3 × 3 convolution kernel with the stride of 2, while the convolution in Stage 3 uses a 1 × 1 convolution kernel with the stride of 1. Additionally, BN and $ReLU$ mean Batchnorm and ReLU operation, respectively. Finally, we get the contrast saliency prediction $\mathbf{F}^{CS}$ (22 × 22 × 1) via a 1 × 1 convolution operation on the output feature maps of stage 3.

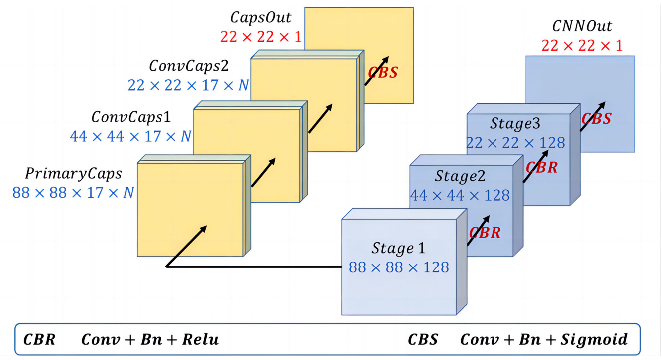


Fig. 4. Framework of the TBIE module. Stage 1 is obtained by fusing all the multi-scale feature maps. In the CNN branch, Stage 2 and Stage 3, which are achieved successively by Stage 1 through the same convolution operation, generate saliency map with contrast cues by one convolution with a kernel size of 1 × 1. In the CapsNets branch, Stage 1 is sent into one PrimaryCaps layers to get capsule features, which will be fed into two ConvCaps layer to generate features with part-whole relationships.

2) CapsNets Branch: CapsNets branch is purposed to enhance the object wholeness of feature maps contained by backbone. This is implemented by CapsNets [22]. To be specific, we design one Primary Capsule (*PrimaryCaps*) layer, one Convolutional Capsule (*ConvCaps*) layer, and one Class Capsule (*ClassCaps*) layer to implement CapsNets.[1] Each layer contains 8 types of capsules. The activation of the *ClassCaps* output is used as the capsule features (22 × 22 × 8 × 1), which is further computed by a convolution to learn part-whole relational saliency prediction F^{PO} (22 × 22 × 1).

C. Multi-Type Cues Correlation (MTCC)

Using the TBIE module, we obtain the contrast and part-whole relational cues with different saliency priors, which prefer to learn the object details and object wholeness, respectively. Therefore, the relation of these two saliency cues will benefit the task of salient object detection. To this end, Fig. 5 designs an MTCC module involving the correlations of these two types of saliency predictions to improve their saliency priors. In the following, we will illustrate the details.

1) Details of MTCC: Suppose $\mathbf{I}_1$ and $\mathbf{I}_2$ with the same spatial resolution $W \times H$ and the channel size of 1 represent the two types of saliency prior, including contrast cues and part-whole relational cues. The spatial correlation for these two-type cues, including horizontal correlation and vertical correlation, is chosen to measure the coherence between these two types of saliency priors. The type correlation algorithm consists of three steps, including spatial correlation computation, correlation guidance, and self-attention.

Step 1: Spatial correlation computation.

Spatial correlation contains horizontal correlation and vertical correlation. The horizontal correlation can be computed by

$$\mathbf{SC}_H \in \Re^{W \times W} = \mathbf{I}_1 \times \mathbf{I}_2^T, \tag{2}$$

[1] *PrimaryCaps*, *ConvCaps*, and *ClassCaps* layers can be referred to TSPOANet [20].

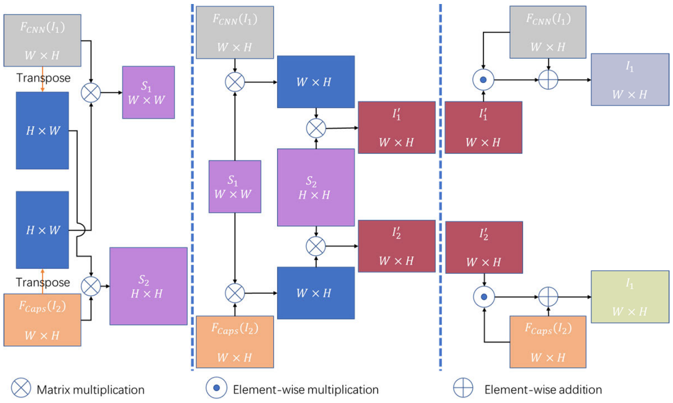


Fig. 5. Framework of the proposed MTCC module. The maps got from TBIE first obtains horizontal- and vertical-level features through matrix multiplication, which are used to guide the original feature map through the element-wise multiplication and addition.

where $(\cdot)^T$ represents the transpose operation. Similarly, the vertical correlation can be computed by

$$\mathbf{SC}_V \in \Re^{H \times H} = \mathbf{I}_1^T \times \mathbf{I}_2. \tag{3}$$

$\mathbf{SC}_H$ and $\mathbf{SC}_V$ reveal the spatial correlations between two types of saliency priors implicitly along the horizontal dimension and vertical dimension, respectively.

Step 2: Correlation guidance.

The spatial correlation $\mathbf{SC}_H$ and $\mathbf{SC}_V$ can be employed to guide the saliency priors $\mathbf{I}_1$ and $\mathbf{I}_2$ to improve their saliency properties. To this end, the guidance can be achieved by the following computation, *i.e.*,

$$\begin{aligned} \mathbf{I}_1{}' \in \Re^{W \times H} &= \mathbf{SC}_H \times \mathbf{I}_1 \times \mathbf{SC}_V, \\ \mathbf{I}_2{}' \in \Re^{W \times H} &= \mathbf{SC}_H \times \mathbf{I}_2 \times \mathbf{SC}_V. \end{aligned} \tag{4}$$

With regard to $\mathbf{I}_1$ and $\mathbf{I}_2$, $\mathbf{I}_1'$ and $\mathbf{I}_2'$ are strengthened in saliency extracting with the guidance of two-type correlation guidance.

Step 3: Self-attention.

To improve the saliency property of two-type saliency predictions, *i.e.*, $\mathbf{I}_1$ and $\mathbf{I}_2$, they are self-attended by their correlation-aware saliency priors, which can be formulated as

$$\begin{aligned} \mathbf{I}_1 &= \mathbf{I}_1 + \mathbf{I}_1 \odot \mathbf{I}_1{}', \\ \mathbf{I}_2 &= \mathbf{I}_2 + \mathbf{I}_2 \odot \mathbf{I}_2{}'. \end{aligned} \tag{5}$$

$\mathbf{I}_1'$ and $\mathbf{I}_2'$ improve the salient property by involving the spatial correlation between two types of saliency predictions, including $\mathbf{I}_1$ and $\mathbf{I}_2$.

Using Eq. (2), Eq. (3), Eq. (4), and Eq. (5), we can obtain the type-correlation based saliency predictions $\mathbf{F}_{CS}'$ and $\mathbf{F}_{PO}'$ from the contrast saliency prediction $\mathbf{F}_{CS}$ and the part-whole relational saliency prediction $\mathbf{F}_{PO}$, respectively.

2) Difference to SCMC [19] and POGU [18]: As shown in Fig. 5, our MTCC differs from SCMC [19] and POGU [18] by computing the mutual-type correlation as *guidance* for two-type information integration rather than using the self-type spatial correlation in SCMC [19] or the direct concatenation of two cues in POGU [18]. Compared to SCMC [19] and POGU [18], experiments in Sec. IV-C.2 further demonstrate that the correlation-guided integration strategy in the proposed MTCC module can efficiently achieve the complementary of two types of semantics for more accurate saliency inference.

3) Difference to PWHCNet [19]: The difference of our MTCC and PWHCNet [19] lies in two folds.

First, our work focuses on the decision-level integration, which is completely different from the feature-level integration of PWHCNet [19]. As shown in Fig. 5 of [19], the inputs for PWHCNet [19] are multiple channels of features maps from contrast cues and part-whole relational cues. In contrast, as shown in Fig. 5 of this paper, the inputs for our work are the detection results with one channel inferred from contrast cues and part-whole relational cues. It is obvious that PWHCNet [19] focuses on the feature-level integration of these two types of cues, while our work focuses on the decision-level integration of these two types of cues. Compared with the feature-level integration, our decision-level integration has two advantages: i) The feature-level integration of PWHCNet [19] is an intermediate-level integration, while our decision-level integration is a high-level integration, which helps our work to get more accurate saliency cues exploration when integrating two types of cues; ii) The feature-level integration of PWHCNet [19] integrates high-dimension feature maps, which inevitably generates heavy computation, while our work integrating two one-channel decision results, which will be computational efficient, as will be verified in Table III.

Secondly, our MTCC implements the mutual correlation for inter-type integration, which is ignored by the self-type attention of PWHCNet [19]. As shown in the Fig. 5 of [19], when integrating these two types of cues, PWHCNet [19] computes two self-attentions, including self-channel attention for one type of cues and self-spatial attention for the other type of cues. In essence, since self-channel attention and self-spatial attention are implemented within one individual type of cues, named self-type attention, there is no inherent interaction between these two types of cues. Differently, in our MTCC, when integrating these two types of cues, we compute two dimensions of interactive maps, including horizontal interactive map and vertical interactive map, which are computed by the mutual correlation between the contrast inference map and part-whole relational inference map. These two interactive maps are used to transform the contrast inference map and part-whole relational inference map to new versions by matrix multiplication, which have inherent interaction between these two types of cues. Therefore, PWHCNet [19] cannot explore the internal interaction between contrast cues and part-whole relational cues, while our MTCC indeed catches the inherent interaction between these two types of cues.

D. Type Interaction Attention (TIA)

1) Motivation: On top of the MTCC module, two saliency predictions $\mathbf{F}_{CS}'$ and $\mathbf{F}_{PO}'$ corresponding to contrast saliency and part-whole relational saliency, respectively, are shown in Fig. 6(d) and (e). It is obvious that $\mathbf{F}_{CS}'$ in Fig. 6(d) tend to be integrated, while the salient objects detected by $\mathbf{F}_{PO}'$ in Fig. 6(e) are more pronounced. But $\mathbf{F}_{CS}'$ mainly focuses

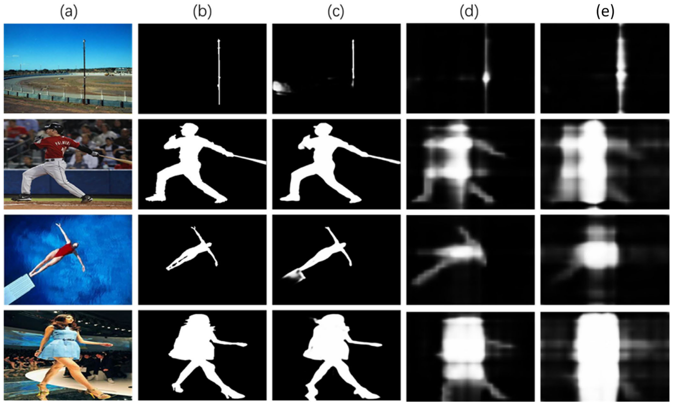


Fig. 6. Illustration for TIA. Comparison of saliency maps before and after TIA. (a) Image; (b) Ground truth; (c) Saliency maps after TIA; (d) CNN-based saliency maps ($\mathbf{F}'_{\mathbf{CS}}$) before TIA; (e) CapsNets-based saliency maps ($\mathbf{F}'_{\mathbf{PO}}$) before TIA.

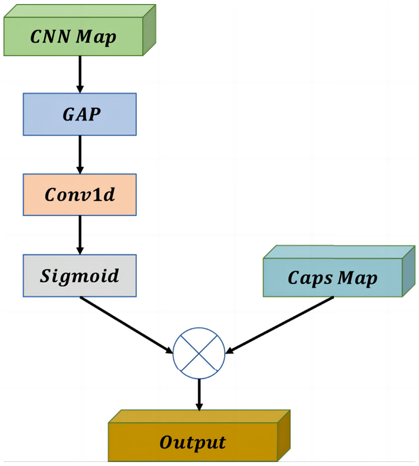


Fig. 7. TIA module. CNNs Map and CapsNets Map represent the saliency maps of contrast cues and part-whole relationships obtained from the MTCC module, respectively. GAP, Conv1d and Sigmoid are the operations of global average pooling, 1D convolution and activation function, respectively.

on contrast cuts, while part-whole relationships dominate in $\mathbf{F}'_{PO}$ after guidance after the MTCC module. So $\mathbf{F}'_{CS}$ and $\mathbf{F}'_{PO}$ can complement to each other. Subsequently, we aim to utilize attention mechanisms to fuse these two feature maps to enhance both types of information, which is achieved by TIA. As shown in Fig. 6(c), the salient objects can be identified and segment uniformly and wholly using TIA.

2) Details: By the MTCC module, we have obtained two complementary feature predictions of contrast cues and part-whole relationships, *i.e.*, $\mathbf{F}'_{CS}$ and $\mathbf{F}'_{PO}$. In the following, we will interact these two types of information for better saliency cues capture. Here, we design an attention mechanism, named TIA, to this end. Specifically, as shown in Fig. 7, CNNs saliency prediction map, *i.e.*, $\mathbf{F}'_{CS}$ is followed by a global average pooling, a convolution with kernel of 1×1, and the Sigmoid activation function, to generate an attention, which is used to guide the CapsNets prediction map, *i.e.*, $\mathbf{F}'_{PO}$. The details of TIA can be formulated as

$$\alpha = \sigma(Conv_31D(GAP(\mathbf{F}'_{CS}))), \tag{6}$$

$$\mathbf{M}_O = \alpha \times \mathbf{F}'_{PO}, \tag{7}$$

where $GAP(\cdot)$ means global average pooling, $Conv_31D(\cdot)$ means 1D convolution with the kernel size of 3 and σ represents the Sigmoid activation function. $\mathbf{F}'_{CS}$ and $\mathbf{F}'_{PO}$ represent two types of feature prediction maps corresponding to CNNs and CapsNets, respectively. $\mathbf{M}_O$ means the output map of TIA.

As shown in Fig. 3, different-scale outputs of TIA are integrated and upsampled stage-wisely from deep to shallow for the final saliency prediction.

3) Difference to ECA [53]: The difference between our TIA and ECA [53] can be concluded as the following two folds. First, ECA [53] absorbs multi-channel feature maps[2] as inputs, while our TIA treats two types of saliency prediction maps as inputs. Secondly, ECA [53] enhances feature representation by a self-attention mechanism, while our TIA manipulates the interactive attention with respect to two types of semantics, including CNNs map and CapsNets map, for saliency prediction (See Fig. 7). Quantitative and visual advantages of our TIA over ECA [53] can be found in Sec. IV-C.

4) Difference to PWHCNet [19]: When interacting contrast cues and part-whole relational cues, PWHCNet [19] adopts the CapsNets cues as the self-spatial activation to attend the CNNs cues, while our TIA activates the CNNs cues to generate the self-channel activation, which is used to attend the CapsNets cues. Due to the fact that the channel-wise activation is efficient over the spatial-wise activation in terms of computational efficiency, our work achieves a computationally efficient mechanism over PWHCNet [19], as can be seen from Table III.

E. Loss Function

In this work, we adopt both weighted BCE loss (L_{wbce}) [54] and weighted IoU loss (L_{wiou}) [54] as our loss function to train the network, *i.e.*,

$$Loss = L_{wbce} + L_{wiou}. \tag{8}$$

L_{wbce} and L_{wiou} can be calculated as follows:

$$L^s_{wbce} = -\frac{\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{W}(1+\gamma\alpha_{ij})\sum\limits_{l=0}^{1}\mathbf{1}(g^s_{ij}=l)log\mathbf{Pr}(p^s_{ij}=l|\Psi)}{\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{W}\gamma\alpha_{ij}}, \tag{9}$$

$$L^s_{wiou} = 1 - \frac{\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{W}(g^s_{ij}\times p^s_{ij})\times(1+\gamma\alpha^s_{ij})}{\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{W}(g^s_{ij}+p^s_{ij}-g^s_{ij}\times p^s_{ij})\times(1+\gamma\alpha^s_{ij})}, \tag{10}$$

where γ is a hyperparameter. $\mathbf{1}(\cdot)$ is a calibration function. p^s_{ij} and g^s_{ij} represent the saliency value of the location of each

[2] In this paper, "feature maps" means intermediate features with multiple channels. In contrast, "saliency prediction map" means one-channel inference map.

TABLE I

QUANTITATIVE COMPARISON OF OUR TCGNET WITH OTHER SOD METHODS. RED, GREEN AND BLUE REPRESENT FOR THE TOP THREE METHODS, RESPECTIVELY

Model	ECSSD [55]				PASCAL-S [56]				DUTS [57]				HKU-IS [24]				DUT-OMRON [58]			
	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$
CNNs based methods																				
AFNet$_{19}$ [28]	0.908	0.947	0.913	0.042	0.815	0.894	0.848	0.072	0.792	0.910	0.867	0.046	0.889	0.950	0.906	0.036	0.739	0.861	0.826	0.057
BASNet$_{19}$ [59]	0.880	0.951	0.916	0.037	0.771	0.886	0.838	0.076	0.791	0.903	0.866	0.048	0.898	0.951	0.908	0.033	0.756	0.872	0.836	0.057
PoolNet$_{19}$ [60]	0.910	0.948	0.917	0.042	0.806	0.897	0.852	0.072	0.799	0.918	0.879	0.042	0.894	0.954	0.912	0.033	0.737	0.869	0.832	0.056
ASNet$_{19}$ [39]	0.875	0.954	0.915	0.047	0.786	0.908	0.861	0.070	0.728	0.893	0.843	0.061	0.871	0.955	0.905	0.041	-	-	-	-
PAGE$_{19}$ [40]	0.906	0.947	0.912	0.042	0.806	0.887	0.842	0.076	0.777	0.896	0.854	0.052	0.884	0.951	0.904	0.036	0.736	0.858	0.825	0.062
TDBU$_{19}$ [41]	0.880	0.954	0.918	0.041	0.775	0.899	0.850	0.071	0.767	0.914	0.865	0.048	0.880	0.955	0.908	0.037	0.739	0.885	0.837	0.061
MINet$_{20}$ [15]	0.924	0.957	0.925	0.034	0.829	0.903	0.856	0.064	0.823	0.917	0.875	0.039	0.906	0.955	0.914	0.030	0.741	0.856	0.822	0.057
ITSD$_{20}$ [14]	0.875	0.949	0.914	0.040	0.773	0.902	0.856	0.068	0.798	0.919	0.877	0.042	0.891	0.951	0.907	0.035	0.745	0.866	0.829	0.063
F3Net$_{20}$ [54]	0.925	0.955	0.924	0.033	0.835	0.904	0.861	0.062	0.840	0.927	0.888	0.036	0.910	0.958	0.917	0.028	0.766	0.872	0.839	0.053
PFSNet$_{21}$ [61]	0.932	0.959	0.930	0.031	0.837	0.907	0.860	0.063	0.846	0.931	0.892	0.036	0.918	0.962	0.924	0.026	0.774	0.878	0.843	0.055
LGSL$_{21}$ [62]	0.931	0.956	0.928	0.032	0.838	0.923	0.858	0.067	0.865	0.942	0.900	0.033	0.920	0.962	0.924	0.027	0.793	0.894	0.853	0.050
RCSBNet$_{22}$ [42]	0.927	0.956	0.922	0.034	0.848	0.906	0.860	0.059	0.856	0.925	0.881	0.035	0.924	0.959	0.919	0.027	0.779	0.866	0.835	0.049
TRACER$_{22}$ [44]	0.922	0.959	0.925	0.031	0.842	0.918	0.867	0.056	0.855	0.945	0.892	0.031	0.915	0.963	0.919	0.027	0.787	0.888	0.847	0.047
MEMNet$_{23}$ [43]	0.932	0.956	0.928	0.031	0.855	0.914	0.871	0.056	0.873	0.943	0.901	0.028	0.926	0.966	0.926	0.026	0.789	0.879	0.850	0.047
Transformer based methods																				
VST$_{21}$ [29]	0.920	0.964	0.932	0.033	0.829	0.918	0.871	0.061	0.818	0.939	0.896	0.037	0.900	0.967	0.929	0.029	0.756	0.888	0.850	0.058
CapsNets based methods																				
TSPOANet$_{19}$ [20]	0.887	0.920	0.868	0.052	0.812	0.877	0.814	0.075	0.797	0.888	0.820	0.048	0.880	0.930	0.866	0.040	0.703	0.826	0.769	0.064
TSPORTNet$_{21}$ [21]	0.914	0.946	0.913	0.041	0.820	0.891	0.850	0.071	0.809	0.912	0.871	0.043	0.901	0.954	0.909	0.032	0.744	0.860	0.823	0.058
PWHCNet$_{21}$ [19]	0.885	0.962	0.932	0.031	0.765	0.910	0.866	0.062	0.824	0.940	0.898	0.035	0.911	0.966	0.929	0.026	0.771	0.885	0.850	0.055
POCINet$_{21}$ [18]	0.917	0.953	0.922	0.040	0.817	0.897	0.854	0.071	0.808	0.920	0.879	0.043	0.903	0.955	0.915	0.033	0.747	0.877	0.838	0.057
DCR$_{22}$ [49]	0.919	0.947	0.914	0.038	0.821	0.888	0.845	0.070	0.824	0.912	0.870	0.041	0.905	0.949	0.907	0.032	0.746	0.853	0.821	0.055
ICON$_{22}$ [23]	0.918	0.953	0.919	0.036	0.834	0.911	0.861	0.064	0.827	0.924	0.878	0.043	0.905	0.957	0.915	0.032	0.755	0.879	0.833	0.065
Ours	0.936	0.966	0.937	0.028	0.845	0.919	0.872	0.056	0.856	0.946	0.901	0.031	0.919	0.968	0.927	0.025	0.789	0.896	0.856	0.046

pixel for saliency prediction and ground truth, respectively. Ψ denotes all the parameters of the model and $\mathbf{Pr}(p_{ij}^s = l|\Psi)$ is the predicted probability.

IV. EXPERIMENT

In this section, we will evaluate and take a deep study on the proposed method with abundant experiments and analyses.

A. Setup Details

1) Datasets: Following previous works, we adopt five public datasets for evaluation, including ECSSD [55], DUTS [57], DUT-OMRON [58], PASCAL-S [56] and HKU-IS [24].

ECSSD [55] contains 1000 images with complicated structures, which are collected from the Internet.

PASCAL-S [56] contains 850 images, which can better demonstrate the semantic segmentation capability of the network.

DUTS [57] contains 10533 training images and 5019 test images, which are with different scenes and various sizes.

DUT-OMRON [58] has 5168 images with different sizes and complex structures.

HKU-IS [24] consists of 3000 training images and 1447 test images, which are with multiple disconnected objects.

We choose the training dataset of DUTS [57] to train the model.

2) Evaluation Criteria: We evaluate the performance of our model as well as other state-of-the-art methods from both visual and quantitative perspectives. The quantitative metrics include weighted F-measure (F_β) [63], Mean Absolute Error (MAE) [63], S-measure (S_m) [64], and E-measure (E_m) [65]. Given a continuous saliency map, a binary mask $\hat{B}$ is achieved by thresholding the saliency map B. Precision is defined as $Precision = \left|\hat{B} \cap G\right| \Big/ \left|\hat{B}\right|$, and recall is defined as $Recall = \left|\hat{B} \cap G\right| \Big/ |G|$.

F-measure is an overall performance indicator, which is computed by

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall}. \tag{11}$$

As suggested in [63], $\beta^2 = 0.3$.

MAE is defined as

$$MAE = \frac{1}{\hat{W} \times \hat{H}} \sum_i |B(i) - G(i)|, \tag{12}$$

where $\hat{W}$ and $\hat{H}$ are the width and height of the image, respectively.

S-measure (S_m) [64] is computed by

$$S_m = \alpha S_o + (1-\alpha) S_r, \tag{13}$$

where S_o and S_r represent the object-aware and region-aware structure similarities between the prediction and the ground truth, respectively. α is set to 0.5 [64].

E-measure (E_m) [65] combines local pixel values with the image-level mean value to jointly evaluate the similarity between the prediction and the ground truth.

3) Implementation Details: The proposed model is implemented with PyTorch and trained for 35 epochs with a batch size of 10 using an NVIDIA GeForce RTX 3090 GPU (24G memory). We adopt HRNet [51], pretrained on ImageNet [66], to initialize the parameters of our backbone. The input images have been resized to 352 × 352 resolution and enhanced with random horizontal rotation and color smoothing. We choose the SGD optimizer [67] with a momentum of 0.9 and weight decay of 0.0005. The learning rate is set to 0.001 and adjusted by a poly strategy with a power of 0.9. The training time of the model is 12.5 hours.

B. Comparison With the States-of-the-Arts

To better evaluate the performance of our model, we compare the proposed architecture with 21 state-of-the-art SOD methods, including 14 CNNs based methods (AFNet [28], BASNet [59], PoolNet [60], ASNet [39], PAGE [40],

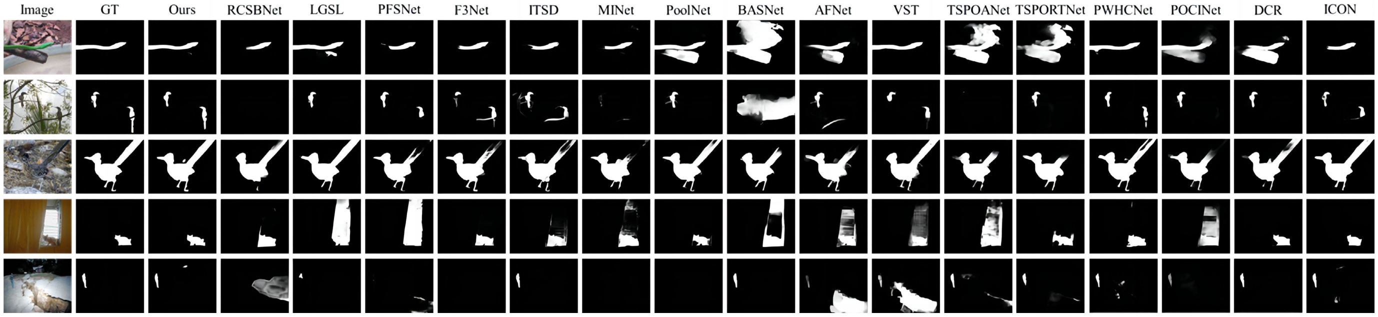


Fig. 8. Visual comparison of other state-of-the-art models. From top to bottom: Strong contrast, small objects, high similarity between foreground and background, strong light, and night scape.

TABLE II

ABLATION STUDY FOR THE PROPOSED METHOD. THE BEST METHOD IS MARKED BY RED

Model	ECSSD [55]				PASCAL-S [56]				DUTS [57]				HKU-IS [24]				DUT-OMRON [58]			
	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$	$F_\beta^w \uparrow$	$maxE_m \uparrow$	$S_m \uparrow$	$MAE\downarrow$
i) Sec. IV-C1: Effectiveness of components.																				
TBIE	0.924	0.949	0.915	0.039	0.841	0.905	0.857	0.062	0.858	0.935	0.885	0.035	0.911	0.953	0.905	0.033	0.779	0.885	0.842	0.047
TBIE + MTCC	0.933	0.958	0.926	0.033	0.848	0.910	0.861	0.060	0.870	0.942	0.897	0.032	0.920	0.961	0.917	0.028	0.785	0.900	0.855	0.047
TBIE + TIA	0.931	0.957	0.925	0.032	0.844	0.908	0.863	0.059	0.867	0.943	0.898	0.031	0.918	0.961	0.918	0.028	0.783	0.895	0.852	0.047
TBIE + MTCC + TIA	0.936	0.966	0.937	0.028	0.845	0.919	0.872	0.056	0.856	0.946	0.901	0.031	0.919	0.968	0.927	0.025	0.789	0.896	0.856	0.046
ii) Sec. IV-C2: Integration mechanisms for contrast and part-whole relations.																				
SCMC [19]	0.930	0.957	0.925	0.033	0.843	0.912	0.865	0.057	0.865	0.945	0.900	0.030	0.915	0.961	0.918	0.028	0.785	0.888	0.847	0.047
POGU [18]	0.927	0.959	0.926	0.033	0.845	0.912	0.868	0.058	0.846	0.937	0.893	0.034	0.911	0.960	0.916	0.029	0.782	0.898	0.853	0.047
MTCC	0.936	0.966	0.937	0.028	0.845	0.919	0.872	0.056	0.856	0.946	0.901	0.031	0.919	0.968	0.927	0.025	0.789	0.896	0.856	0.046
iii) Sec. IV-C3: TIA *vs.* ECA.																				
ECA [53]	0.929	0.959	0.927	0.033	0.841	0.912	0.863	0.060	0.859	0.942	0.893	0.033	0.919	0.962	0.917	0.029	0.788	0.891	0.849	0.047
TIA	0.936	0.966	0.937	0.028	0.845	0.919	0.872	0.056	0.856	0.946	0.901	0.031	0.919	0.968	0.927	0.025	0.789	0.896	0.856	0.046
iv) Sec. IV-C4: CNNs map (TIA) *vs.* CapsNets map for attention (TIA*).																				
TIA*	0.923	0.958	0.925	0.034	0.837	0.913	0.868	0.059	0.851	0.940	0.894	0.033	0.912	0.959	0.915	0.029	0.789	0.894	0.853	0.046
TIA	0.936	0.966	0.937	0.028	0.845	0.919	0.872	0.056	0.856	0.946	0.901	0.031	0.919	0.968	0.927	0.025	0.789	0.896	0.856	0.046
v) Sec. IV-C5: Coarse map *vs.* feature maps for MTCC.																				
MTCC-FM	0.922	0.959	0.928	0.035	0.829	0.911	0.866	0.065	0.841	0.935	0.889	0.037	0.918	0.966	0.923	0.028	0.776	0.895	0.850	0.054
MTCC	0.936	0.966	0.937	0.028	0.845	0.919	0.872	0.056	0.856	0.946	0.901	0.031	0.919	0.968	0.927	0.025	0.789	0.896	0.856	0.046
vi) Sec. IV-C6: MTCC *vs.* addition/mulltiplication.																				
Addition	0.925	0.959	0.926	0.034	0.838	0.914	0.867	0.060	0.849	0.941	0.895	0.033	0.911	0.961	0.916	0.029	0.782	0.894	0.850	0.047
Multiplication	0.921	0.958	0.925	0.035	0.834	0.910	0.863	0.061	0.848	0.943	0.896	0.033	0.908	0.960	0.916	0.030	0.785	0.896	0.852	0.046
MTCC	0.936	0.966	0.937	0.028	0.845	0.919	0.872	0.056	0.856	0.946	0.901	0.031	0.919	0.968	0.927	0.025	0.789	0.896	0.856	0.046

TDBU [41], MINet [15], ITSD [14], F3Net [54], PFSNet [61], LGSL [62], RCSBNet [42]) TRACER [44] and MEMNet [43], 1 Transformer based method (VST [29]), and 6 CapsNet based methods (TSPOANet [68], PWHCNet [19], POCINet [18], DCR [49], ICON [23]).

1) Quantitative Comparisons: Table I lists F_β^w, S_m, $max E_m$ and MAE values of different methods. It is obvious that our method outperforms other methods on almost all the datasets regarding these four metrics. Especially, we perform best in terms of all metrics on ECSSD [55]. Besides, our method achieves three best metrics on complicated DUTS [57] and DUT-OMRON [58]. Compared with the best compared method, *i.e.*, MEMNet [43] which achieves 6 best metrics, our model achieves 13 best metrics, which indicates our model has the superiority on various scenes over MEMNet [43] and other methods.

2) Visual Comparisons: Visual comparisons between our model and other methods are shown in Fig. 8. To make the comparison more sufficient, we display various scenes, including strong contrast, small objects, high similarity between foreground and background, strong light, and night scape. It is obvious that most the state-of-the-art methods cannot handle all the listed scenes with introductions of noise or incomplete shapes. By contrast, our proposed method not only locates the salient objects accurately, but also ensures the integrity in every situation. This gets benefit from the primitive interaction mechanism for contrast from CNNs and part-whole relations from CapsNets in our model.

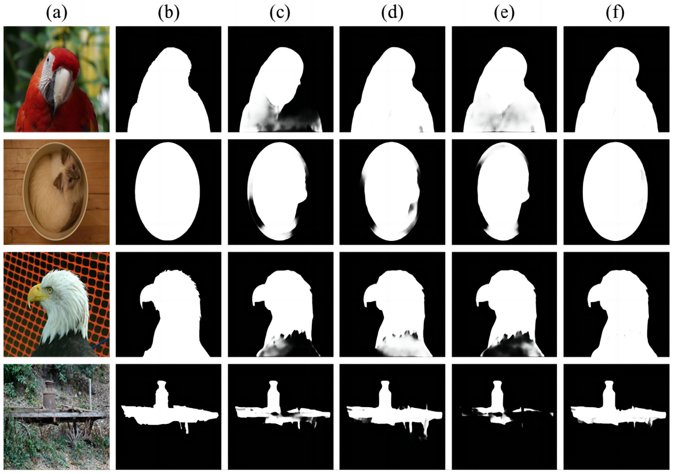


Fig. 9. Visual comparisons of different components. (a) Image, (b) GT, (c) TBIE, (d) TBIE + MTCC, (e) TBIE + TIA, (f) TBIE + MTCC + TIA.

C. Ablation Studies

We conduct the ablation experiments to verify the contributions of our main components. All these models described below are trained on the same DUTS training datasets under the same implementation details described in IV-A.3.

1) Effectiveness of Components: We verify the performance of each component by testing various simplified versions of our model. Table IV-C-i) lists performance of different

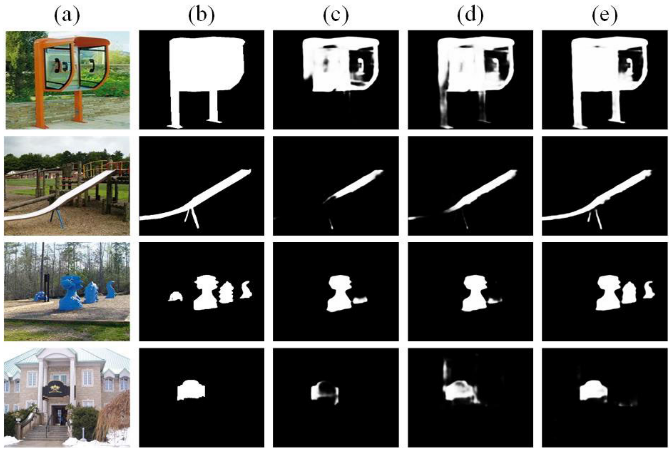


Fig. 10. Visual comparisons of different module versions. (a) Image; (b) GT; (c) SCMC [19]; (d) POGU [18]; (e) OURS (MTCC).

versions of the proposed model. As shown in Table IV-C-i), only using TBIE without other two modules can well tackle the task of saliency detection with good performance. The combinations of "TBIE + MTCC" and "TBIE + TIA" both surpass the baseline TBIE, which proves the effectiveness of the proposed MTCC and TIA. Ulteriorly, the whole model covering TBIE, MTCC, and TIA achieves a further performance gain. As shown in Fig. 9, on top of the baseline, *i.e.*, TBIE, MTCC enhances the object shapes capture (*e.g.*, rows 1 & 3) and object details (*e.g.*, row 4), and TIA strengthens the object wholeness (*e.g.*, row 1). The joint force of MTCC and TIA predicts the salient maps close to the ground truth. In summary, the proposed MTCC and TIA contribute significantly to the whole model for the task of saliency prediction.

2) Integration Mechanisms for Contrast and Part-Whole Relations: Our MTCC integrates contrast cues and part-whole relational cues for saliency prediction. To study its superiority, we compare our MTCC with two related integration mechanisms for contrast and part-whole relationships, including SCMC [19] and POGU [18]. For fair comparisons, we replace our MTCC with these two mechanisms in our framework for training. As Table IV-C-ii) shows, under the same setting, our MTCC beats SCMC [19] and POGU [18]. Visually in Fig. 10, our MTCC achieves better object wholeness (*e.g.*, top three rows) and inner details (*e.g.*, bottom row), compared with SCMC [19] and POGU [18]. This indicates our MTCC interact more efficiently contrast and part-whole relations than SCMC [19] and POGU [18].

3) TIA vs. ECA [53]: To take a deep study on our TIA, we compare it with a related work, *i.e.*, ECA [53]. Specially, we replace our TIA with ECA [53] within the proposed network architecture for fair comparisons. As represented in Table IV-C-iii), TIA beats ECA with a significant gap. As shown in Fig. 11, our TIA performs better for salient object detection in terms of objects wholeness and details. These observations prove that our TIA attends those primitive CNNs features and part-whole relations from CapsNets with a more intelligent attention mechanism compared with ECA [53] for salient object detection.

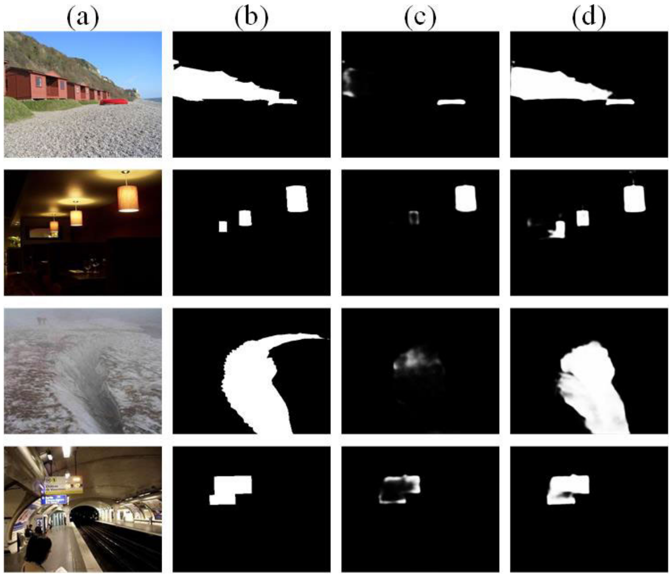


Fig. 11. Visual comparisons of different module versions. (a) Image; (b) GT; (c) ECA [53]; (d) OURS (TIA).

4) CNNs Map vs. CapsNets Map for Attention: In our TIA, CNNs prediction map is activated to attend CapsNets prediction map. To take a deep insight into TIA, we compare our TIA with a modified version, named TIA*, in which CapsNets prediction map is activated to attend the CNNs prediction map. As listed in Table IV-C-iv, our TIA beats TIA* by a lot. Visually in Fig. 12, the saliency maps learned by our TIA obtain better object wholeness and inner details than those of TIA*. The improvements of TIA over TIA* demonstrates the superiority of the attention mechanism of our TIA for saliency prediction.

5) Coarse Map vs. Feature Maps for MTCC: To study the effectiveness of the coarse maps of CNNs and CapsNets in MTCC, we compare our framework with a modified version, named MTCC-FM, in which the feature maps instead of coarse maps of CNNs and CapsNets are employed for MTCC. As listed in Table IV-C-v), our MTCC performs better than MTCC-FM by a large margin. Besides, Fig. 13 describes some visual results of MTCC and MTCC-FM. It can be found that MTCC-FM produces much noise in the saliency map, while our MTCC predict the accurate salient objects. This is because that multiple channels of feature maps in MTCC-FM inevitably contain some noisy channels, which causes the degradation of performance. By contrast, the coarse maps learned from multiple feature maps in MTCC has cleared the noise to some extent, which improves the performance a lot.

6) MTCC vs. Addition/Multiplication: To explore the role of MTCC for the interaction of CNNs prediction map and CapsNets prediction map, we compare it with two modified versions, named MTCC-A and MTCC-M, which are implemented by directly integrating CNNs prediction map and CapsNets prediction map via addition and multiplication, respectively. As listed in Table IV-C-vi), our MTCC is superior over MTCC-A and MTCC-M, which indicates that

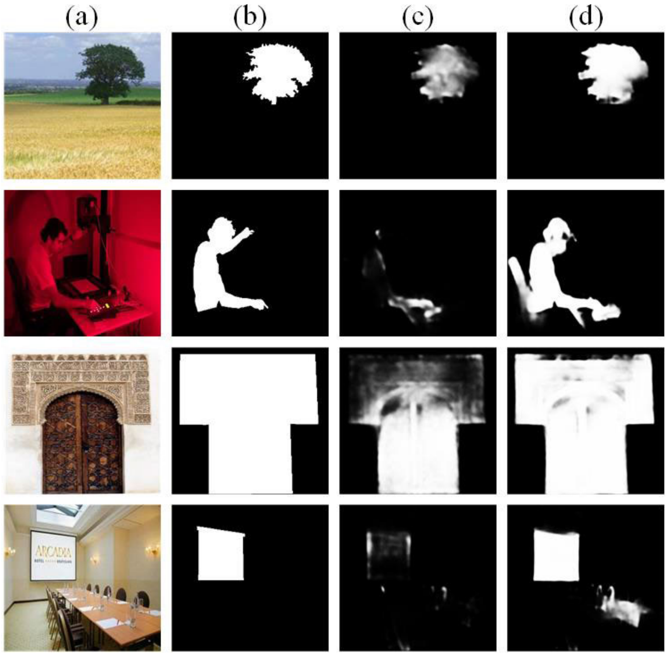


Fig. 12. Visual comparisons of different module versions. (a) Image; (b) GT; (c) TIA*; (d) OURS (TIA). TIA* means a modified TIA version, in which CapsNets prediction map is activated to attend the CNNs prediction map.

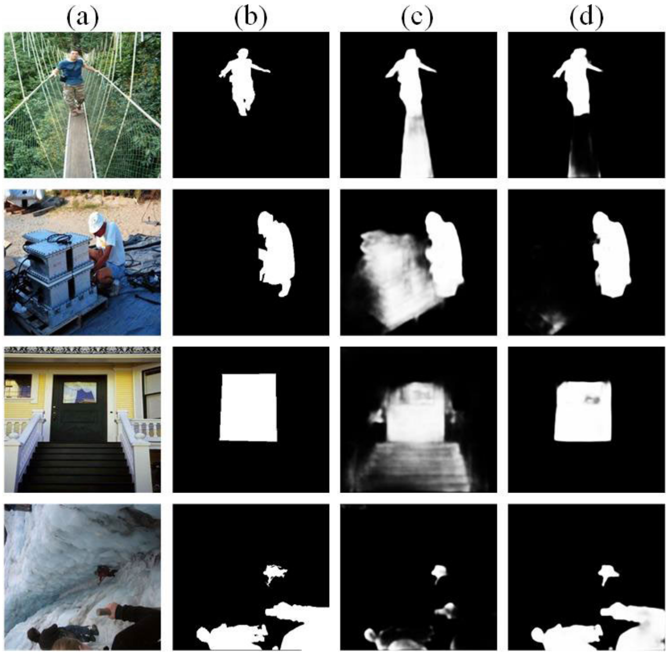


Fig. 13. Visual comparisons of different module versions. (a) Image; (b) GT; (c) MTCC-FM; (d) OURS (MTCC). MTCC-FM employs the feature maps instead of coarse maps of CNNs and CapsNets for MTCC.

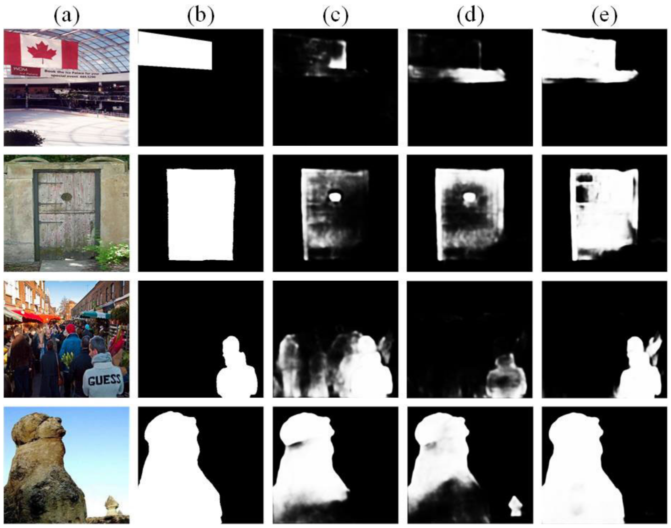


Fig. 14. Visual comparisons of different module versions. (a) Image; (b) GT; (c) MTCC-A; (d) MTCC-M; (e) OURS (MTCC). MTCC-A and MTCC-M, are implemented by directly integrating CNNs prediction map and CapsNets prediction map via addition and multiplication, respectively.

TABLE III

Complexity Comparison With CapsNets Saliency Methods. Top Two Methods Are Marked by red and green, Respectively

Method	Input size	FLOPs (G)	Time (s)
TSPOANet [20]	352 × 352	197.78	0.32
TSPORTNet [21]	352 × 352	267.50	0.35
POCINet [18]	352 × 352	332.30	0.1
DCR [49]	352 × 352	60.78	0.06
ICON [23]	352 × 352	64.90	0.013
PWHCNet [19]	256 × 256	137.64	0.167
Ours	352 × 352	50.55	0.05

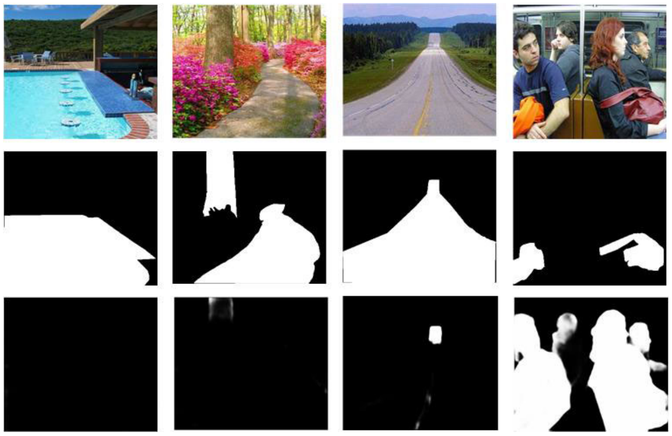


Fig. 15. Failure cases. From top to bottom: Image; GT; saliency maps.

the inter-type correlation guidance strategy of MTCC performs better than the simple addition and multiplication guidance strategies. Besides, as depicted in Fig. 14, our MTCC produces the saliency maps close to the ground truth, while MTCC-A and MTCC-M predict poor ones.

7) Complexity: To study the computational complexity of our model with respect to the related CapsNets based salient detectors, Table III lists FLOPs and inference time of different CapsNets based methods. It shows that our model achieves the lowest FLOPs and second best inference time. That proves that our model shares a good efficiency within the scope of CapsNets based saliency detection.

D. Failure Case

Despite our method achieves promising performance, there are still many challenges. Fig. 15 depicts some failure cases. The salient objects in images of Fig. 15 are characterized with some scene semantics instead of simply high-contrast regions, which challenge our framework. In the future, we will take into account the salient semantics [38], [69] to improve our model on the real-world scene understanding.

V. Conclusion

In this paper, we have proposed a framework to extract the coherence between contrast cues and part-whole relations for salient object detection. Our key idea is integrating the correlations of these two cues from CNNs and CapsNets and let them interact. For more in-depth interaction, we have also developed an attention mechanism involving these two types of semantics to infer saliency. The evaluation of our model on five datasets has shown our excellent performance compared with other state-of-the-art methods. In the future, we will take into account salient semantics for high-level saliency understanding.

References

[1] L. Qin et al., "ID-YOLO: Real-time salient object detection based on the driver's fixation region," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15898–15908, Sep. 2022.

[2] T. Deng, K. Yang, Y. Li, and H. Yan, "Where does the driver look? Top-down-based saliency detection in a traffic driving environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2051–2062, Jul. 2016.

[3] M. Ju, C. He, J. Liu, B. Kang, J. Su, and D. Zhang, "IVF-net: An infrared and visible data fusion deep network for traffic object enhancement in intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 1220–1234, Jan. 2023.

[4] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, "Salient object detection for searched web images via global saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3194–3201.

[5] W. Wang, J. Shen, X. Lu, S. C. H. Hoi, and H. Ling, "Paying attention to video object pattern understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2413–2428, Jul. 2021.

[6] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.

[7] W. Wang, G. Sun, and L. Van Gool, "Looking beyond single images for weakly supervised semantic segmentation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, pp. 1–15, 2022, doi: 10.1109/TPAMI.2022.3168530.

[8] 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.

[9] B. Tian, Y. Li, B. Li, and D. Wen, "Rear-view vehicle detection and tracking by combining multiple parts for complex urban surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 597–606, Apr. 2014.

[10] W. Qian, Z. He, C. Chen, and S. Peng, "Navigating diverse salient features for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24578–24587, Dec. 2022.

[11] S. S. Thomas, S. Gupta, and V. K. Subramanian, "Event detection on roads using perceptual video summarization," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 9, pp. 2944–2954, Sep. 2018.

[12] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

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

[14] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9138–9147.

[15] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9410–9419.

[16] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 35–51.

[17] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8778–8787.

[18] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Integrating part-object relationship and contrast for camouflaged object detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5154–5166, 2021.

[19] Q. Zhang, M. Duanmu, Y. Luo, Y. Liu, and J. Han, "Engaging part-whole hierarchies and contrast cues for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3644–3658, Jun. 2022.

[20] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1232–1241.

[21] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3688–3704, Jul. 2022.

[22] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 3856–3866.

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

[24] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.

[25] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 660–668.

[26] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6593–6601.

[27] Y. Zeng, H. Lu, L. Zhang, M. Feng, and A. Borji, "Learning to promote saliency detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1644–1653.

[28] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1623–1632.

[29] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4702–4712.

[30] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1265–1274.

[31] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.

[32] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 825–841.

[33] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4039–4048.

[34] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[35] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3798–3807.

[36] Y. Tang and X. Wu, "Saliency detection via combining region-level and pixel-level predictions with CNNs," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 809–825.

[37] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "PoolNet+: Exploring the potential of pooling for salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 887–904, Jan. 2023.

[38] M.-M. Cheng, S.-H. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8006–8021, Nov. 2022.

[39] 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.

[40] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1448–1457.

[41] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, “An iterative and cooperative top-down and bottom-up inference network for salient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5961–5970.

[42] Y. Y. Ke and T. Tsubono, “Recursive contour-saliency blending network for accurate salient object detection,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2940–2950.

[43] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, “Pixels, regions, and objects: Multiple enhancement for salient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10031–10040.

[44] M. S. Lee, W. Shin, and S. W. Han, “TRACER: Extreme attention guided salient object tracing network (student abstract),” in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 11, 2022, pp. 12993–12994.

[45] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, “EDN: Salient object detection via extremely-downsampled network,” *IEEE Trans. Image Process.*, vol. 31, pp. 3125–3136, 2022.

[46] M. Ma, C. Xia, C. Xie, X. Chen, and J. Li, “Boosting broader receptive fields for salient object detection,” *IEEE Trans. Image Process.*, vol. 32, pp. 1026–1038, 2023.

[47] S. Jiao et al., “Collaborative content-dependent modeling: A return to the roots of salient object detection,” *IEEE Trans. Image Process.*, vol. 32, pp. 4237–4246, 2023.

[48] Y. Liu, X. Dong, D. Zhang, and S. Xu, “Deep unsupervised part-whole relational visual saliency,” *Neurocomputing*, vol. 563, Jan. 2024, Art. no. 126916.

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

[50] S. Zhao, Z. Wen, Q. Qi, K.-M. Lam, and J. Shen, “Learning fine-grained information with capsule-wise attention for salient object detection,” *IEEE Trans. Multimedia*, early access, pp. 1–14, 2023, doi: 10.1109/TMM.2023.3234436.

[51] J. Wang et al., “Deep high-resolution representation learning for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[52] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[53] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-Net: Efficient channel attention for deep convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[54] J. Wei, S. Wang, and Q. Huang, “F^{3}Net: Fusion, feedback and focus for salient object detection,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12321–12328.

[55] Q. Yan, L. Xu, J. Shi, and J. Jia, “Hierarchical saliency detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

[56] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, “The secrets of salient object segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[57] L. Wang et al., “Learning to detect salient objects with image-level supervision,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3796–3805.

[58] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, “Saliency detection via graph-based manifold ranking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[59] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, “BASNet: Boundary-aware salient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7471–7481.

[60] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, “A simple pooling-based design for real-time salient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3912–3921.

[61] M. Ma, C. Xia, and J. Li, “Pyramidal feature shrinking for salient object detection,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, 2021, pp. 2311–2318.

[62] B. Xu, H. Liang, R. Liang, and P. Chen, “Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–9.

[63] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[64] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

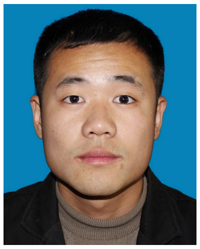
[65] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

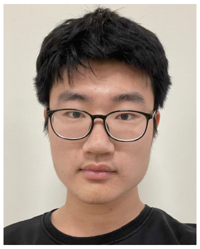
[67] H. Robbins and S. Monro, “A stochastic approximation method,” *Ann. Math. Statist.*, vol. 22, pp. 400–407, Sep. 1951.

[68] Y. Liu, J. Han, Q. Zhang, and C. Shan, “Deep salient object detection with contextual information guidance,” *IEEE Trans. Image Process.*, vol. 29, pp. 360–374, 2020.

[69] T. Do, K. Vuong, and H. S. Park, “Egocentric scene understanding via multimodal spatial rectifier,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2822–2831.



Yi Liu received the Ph.D. degree from Xidian University, China, in 2019. He is currently a Professor with Changzhou University, China. From 2018 to 2019, he was a Visiting Scholar with Lancaster University. His research interests include machine learning and computer vision, especially on saliency detection, capsule networks, 3D point cloud, and object detection.



Ling Zhou is currently pursuing the B.E. degree with the School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, Jiangsu, China. His research interests include image processing and deep learning.

Gengshen Wu (Member, IEEE) received the Ph.D. degree from Lancaster University, U.K. He is currently with the Faculty of Data Science, City University of Macau. His research interests include computer vision and pattern recognition.

Shoukun Xu received the Ph.D. degree from the China University of Mining and Technology, China, in 2001. He is currently a Professor with Changzhou University. He is also the Chair of China Computer Federation, Changzhou Branch, and a Distinguished Member of China Computer Federation. His research interests include digital twins, computer vision, and blockchain.

Jungong Han (Senior Member, IEEE) is currently a Chair Professor in computer vision with the Department of Computer Science, University of Sheffield, U.K. He also holds an Honorary Professorship with the University of Warwick, U.K. His research interests include computer vision, artificial intelligence, and machine learning. He is a fellow of the IAPR.