Global-and-Local Collaborative Learning for Co-Salient Object Detection

Runmin Cong, Member, IEEE, Ning Yang, Chongyi Li, Huazhu Fu, Senior Member, IEEE, Yao Zhao, Senior Member, IEEE, Qingming Huang, Fellow, IEEE, and Sam Kwong, Fellow, IEEE

Abstract—The goal of co-salient object detection (CoSOD) is to discover salient objects that commonly appear in a query group containing two or more relevant images. Therefore, how to effectively extract interimage correspondence is crucial for the CoSOD task. In this article, we propose a global-and-local collaborative learning (GLNet) architecture, which includes a global correspondence modeling (GCM) and a local correspondence modeling (LCM) to capture the comprehensive interimage corresponding relationship among different images from the global and local perspectives. First, we treat different images as different time slices and use 3-D convolution to integrate all intrafeatures intuitively, which can more fully extract the global group semantics. Second, we design a pairwise correlation transformation (PCT) to explore similarity correspondence between pairwise images and combine the multiple local pairwise correspondences to generate the local interimage relationship. Third, the interimage relationships of the GCM and LCM are integrated through a global-and-local correspondence aggregation (GLA) module to explore more comprehensive interimage collaboration cues. Finally, the intra and inter features are adaptively integrated by an intra-and-inter weighting fusion (AEWF) module to learn co-saliency features and predict the co-saliency map. The proposed GLNet is evaluated on three prevailing CoSOD benchmark datasets, demonstrating that our model trained on a small dataset (about 3k images) still outperforms 11 state-of-the-art competitors trained on some large datasets (about 8k–200k images).

Index Terms—3-D convolution, co-salient object detection (CoSOD), global correspondence modeling (GCM), local correspondence modeling (LCM).

Manuscript received 27 January 2022; accepted 18 April 2022. Date of publication 22 July 2022; date of current version 15 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YZ0012100; in part by the Beijing Nova Program under Grant ZK201100006280016; in part by the National Natural Science Foundation of China under Grant 6202014, Grant U1936212, Grant 62120106009, Grant U21B2038, and Grant 61931008; in part by the Beijing Natural Science Foundation under Grant 4222013; in part by the Hong Kong Research Grants Council (RGC) General Research Funds under Grant 9042816 (CityU 11209819) and Grant 9042958 (CityU 11203820); in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA); in part by the Young Elite Scientist Sponsorship Program by the China Association for Science and Technology under Grant 2020QNRC001; in part by the Young Elite Scientist Sponsorship Program by the Beijing Association for Science and Technology; in part by the Hong Kong Scholars Program under Grant XJ2020040; in part by the CAAI-Huawei MindSpore Open Fund; and in part by the Miss Yang’s Project of the Fundamental Research Funds for the Central Universities under Grant 2021YJS046. This article was recommended by Associate Editor Y. Yuan. (Corresponding author: Chongyi Li.)

Please see the Acknowledgment section of this article for the author affiliations.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2022.3169431.

Digital Object Identifier 10.1109/TCYB.2022.3169431

I. INTRODUCTION

SALIENT object detection (SOD) simulates the human visual attention mechanism to locate the visually attractive or most prominent objects/regions from a scene [1], which has been applied to a large number of vision tasks, such as image classification [2], image compression [3], image retargeting [4], etc. With different data inputs, the SOD task can be further divided into some subtasks, such as RGB-D SOD [5]–[13], video SOD [14], remote sensing SOD [15]–[19], etc. In addition, with the explosive growth of data volume in recent years, people sometimes need to collaboratively perceive multiple relevant images, such as identifying teammates wearing the same uniform in different images. Therefore, simulating the human co-processing capabilities, co-salient object detection (CoSOD) aims to localize the common and salient objects in an image group containing multiple relevant images. As can be seen from the definition of CoSOD, it contains two key matters: 1) the saliency attribute and 2) the repetitiveness attribute. In other words, in addition to capturing the saliency attribute of each image, the interactive relationships among different images play an important role in determining whether the objects are shared in the entire image group. Moreover, the co-salient objects only belong to the same semantic category, but their appearance characteristics vary differently, which undoubtedly increases the difficulty of the task.

The existing CoSOD methods can be roughly divided into traditional methods [20]–[24] and deep-learning-based methods [25]–[41]. Traditional methods usually use handcrafted features to model the interimage relationships into feature clustering, similarity matching, rank constraint, etc. However, their performance is limited due to the lack of high-level semantic representation, especially for the complex scenes. Since entering the deep learning era, the CoSOD task has achieved a rapid development, and its performance has continued to set new records. For the process of generating group semantics, the modeling of the relationship among images can be classified into three categories: one is the global model, which combines all the intrafeatures into the designed model to learn the global semantics [25]–[30]. Another is the local model, which decouples the interimage relationship into the multiple pairwise correspondences [36]. The last one is the recurrent model, which utilizes recursive structure (such as RNN or GRU) to capture the group semantics [35], [37]. Among them, the recurrent model is sensitive to the input order of the images, the
global model is closer to a blind processing method requiring a delicately designed network to capture better group semantics, and the local model has a clearer physical meaning, but are more computationally expensive. On the whole, the global model and local model are naturally complementary to a certain degree, but there is no work to integrate them organically under a unified framework.

To this end, we propose a novel network based on global-and-local collaborative learning (GLNet) to fully explore the semantic association among the group images from the global and local perspectives. The interimage relationship extraction includes a global correspondence modeling (GCM) that directly integrates all intrafeatures to learn the global semantics, and a local correspondence modeling (LCM) that combines corresponding relationships between pairwise images in the group. Considering that the global modeling method directly extracts relationships of multiple images at a time, it includes the dimensions of different images other than the traditional spatial and channel dimensions. Simple 2-D convolution may be difficult to extract the corresponding relationships in an image group. Thus, we introduce 3-D convolution into the CoSOD task for the first time to capture a more accurate and comprehensive global interimage relationship. For LCM, we design a pairwise correlation transformation (PCT) to explore similarity correspondence between pairwise images. The differences with the existing local model [36] are reflected in two aspects: 1) When measuring the pairwise correlation, the PCT is executed at the pixel level. Each spatial location is associated with all other locations through a max-pooling layer, which has the global description capability and 2) when fusing the multiple local correspondences, we use progressive 3-D convolution layers, which assist in reducing the redundancy and suppressing the interference. In addition, the global and local interinformation are integrated through a global-and-local correspondence aggregation (GLA) module, and an intra-and-inter weighting fusion (AEWF) strategy is designed to adaptively integrate the intra and interfeatures into co-saliency features to predict the corresponding co-saliency map. As can be seen in Fig. 1, our method can accurately and completely detect the co-salient object, even if the size of the object varies dramatically.

The major contributions of the proposed method are summarized as follows.

1) We propose an end-to-end network for CoSOD, the core of which is to capture a more comprehensive interimage relationship through the GLNet. The GCM module directly extracts the interactive information of multiple images intuitively, and the LCM module defines the interimage relationship through the form of multiple pairwise images.

2) For the GCM, different images are regarded as different time dimensions and, thus, 3-D convolution is used instead of the usual 2-D convolution to capture global group semantics.

3) For the LCM, we design a PCT to explore similarity correspondence between pairwise images. The two modules work together to learn more in-depth and comprehensive interimage relationships.

The remainder of this article is organized as follows. Section II reviews the related works. Section III introduces the proposed GLNet method in detail. The experimental results with quantitative evaluation are presented in Section IV. Finally, the conclusion is drawn in Section V.
proposed the first CNN-based CoSOD model, where the individual image saliency features were simply concatenated to learn the groupwise feature representation. Wang et al. [26] learned the groupwise semantic vector to represent the interimage correspondence. Zhang et al. [27] designed an effective aggregation-and-distribution strategy for interimage modeling and achieved highly competitive performance. Zhang et al. [29] proposed a gradient induction model that used image gradient information to draw more attention to the discriminative common saliency features. Zhang et al. [30] designed an adaptive graph convolutional network to capture the intra and interimage correspondence of an image group. Jin et al. [36] proposed that the weighted average each pairwise image was masked by the predicted saliency map as group semantics, and then used to compare the cosine similarity of each position of each image. Fan et al. [38] learned common information using co-attentional projection strategy, and at the same time established a CoSOD3k dataset for the CoSOD task. Tang et al. [39] regarded the CoSOD task as an end-to-end sequence prediction problem and proposed the CoSOD transformer (CoSformer) network. Qian et al. [40] used a two-stream encoder generative adversarial network (TSE-GAN) with progressive training. Zhang et al. [41] proposed a consensus-aware dynamic convolution model to perform the summarize and search process.

For the process of generating group semantics, these methods perceive the interimage relationships from a global perspective or a local perspective. The global perspective uses a direct way to extract the group semantic features of multiple images at once, while the local perspective allows the network to learn pairwise relationships in a step-by-step manner and finally fuse them into group semantics. In fact, we can capture group semantics in an all-round way from two perspectives. Therefore, we propose a novel GLNet, where the GCM and LCM modules are designed to fully explore the semantic association among the group image from different perspectives.

### III. PROPOSED METHOD

#### A. Overview

The framework of the proposed method is shown in Fig. 2. Given a query group, including $N$ relevant images $I = \{I^n\}_{n=1}^N$, the goal of the CoSOD task is to detect commonly salient objects and predict the corresponding co-saliency maps $M = \{M^n\}_{n=1}^N$. First, we use the backbone network (e.g., VGG [56], ResNet [57], Dilated ResNet [58]) to extract the multilevel features of each image in the group. Considering that the top-level convolutional features contain more semantic information and are more suitable for exploring the corresponding information among the group images, we use them as the final intrafeatures $F^n_{in} \in \mathbb{R}^{C \times H \times W}$, where $C$, $H$, and $W$ denote the channel, height, and width of the feature map, respectively. Then, in order to fully exploit the interimage cues among different images, a GLNet architecture is designed. In addition to the designed GCM that directly fuses all the intrafeatures through the 3-D convolution, we also design an LCM method from a local perspective by decoupling the multi-image relationship into multiple pairwise-image correspondences. Furthermore, the interactive information of these two complementary aspects is integrated through a 3-D convolution, we also design an LCM method from a local perspective by decoupling the multi-image relationship into multiple pairwise-image correspondences. Finally, the intra and inter features are adaptively integrated through an AEWF module and are fed into a decoder network to predict the corresponding co-saliency maps $M$. 

![Fig. 2. Flowchart of the proposed GLNet. Given a query image group, we first obtain the deep intra features with a shared backbone feature extractor. A GCM module is applied to capture the global interactive information among multiple images, while an LCM module further collects the local interactive information between pairwise images. The global and local features are generated into inter features through a GLA module. Then, the intra and inter features are further adaptively aggregated through an AEWF module to learn co-saliency features. In the end, a multilayer deconvolution is used to predict the full-resolution co-saliency maps $M$.](image-url)
B. Global Correspondence Modeling

Intuitively, the co-salient objects in an image group have the same semantic attributes in features. The extraction of the interimage correspondence among images is essential to determine which features are shared by multiple images. Therefore, the most intuitive way to achieve this is to directly integrate the different intrafeatures in the group, and then learn the common relationship attributes through some 2-D convolutional layers [25], [26]. This is a global modeling method that can directly extract multiple-image relationships at once. However, it is very difficult to adequately extract the relationship between different images using traditional 2-D convolutional layers because a new dimension of different images is added compared to a single image (we will verify this in the ablation study of Section IV). The superiority of 3-D convolution has been previously demonstrated in other vision tasks [59], [60]. As stated in [59], replacing 2-D convolutions with 3-D convolutions not only reduces the size of the model but also improves the performance of the model. Inspired by this, we treat the different images as different time slices and use 3-D convolution to replace 2-D convolution for GCM.

Specifically, the group features \( \mathbf{F} \in \mathbb{R}^{C \times N \times H \times W} \) are first obtained by concatenating all intrafeatures \( \mathbf{F}_{ia} \) of different images along channel dimension, which are the inputs of the 3-D convolutional layer. Then, three 3-D convolutional layers are used to capture the global correspondence

\[
g = \delta \circ f_2 \circ \delta \circ f_3 \circ \delta \circ f_2 \circ \delta \circ \mathbf{F}
\]

where \( f_2 \) and \( f_3 \) denote 3-D convolutional layers with filter sizes of \( 2 \times 3 \times 3 \) and \( 3 \times 3 \times 3 \), respectively, \( \delta \) is the ReLU activation function, and \( \circ \) denotes function composition. In order to introduce more compact global interimage presentation, we use the channel attention (CA) [61] and spatial attention (SA) [62] to highlight the important channels and spatial locations, thereby producing the final global correspondence \( \mathbf{G} \in \mathbb{R}^{C \times H \times W} \)

\[
\mathbf{G} = \text{SA}(\text{CA}(g))
\]

where \( \text{CA}(\cdot) \) and \( \text{SA}(\cdot) \) are the CA operation and SA operation, respectively.

C. Local Correspondence Modeling

The GCM captures the interimage relationship straightforwardly and globally from the image group at one time, which means that it needs to analyze the relationship between all images in the group at the same time. In fact, we can further decompose the relationship modeling among multiple images into more basic and smaller units, that is, the combination of multiple pairwise correspondences. Specifically, in order to capture the corresponding relationship between a certain image and all other \( N - 1 \) reference images in the group, we can model the pairwise relationship between it and others separately, and then integrate these relationships together to obtain its final interimage correspondence. This is the LCM we proposed, which is conducive to capturing more in-depth local interaction information among images. In order to illustrate simply and clearly how our LCM works in the GLNet, we take the \( k \)th image \( \mathbf{I}^k \) in a group including \( N \) images as an example, where \( k \in \{1, 2, \ldots, N\} \). The left side of Fig. 3 shows the pipeline of LCM for calculating the local interimage correspondence \( \mathbf{P}^k \) of image \( \mathbf{I}^k \).

Because the LCM only needs to process two images in the group, it is different from the global correspondence model which is similar to a black-box calculation. Therefore, we design a more physically meaningful pairwise correlation method to measure the semantic correlation. We, respectively, calculate the correlation between the current intrafeatures \( \mathbf{F}_{ia}^k \) and other reference intrafeatures \( \mathbf{F}_{ia}^j \) (\( j \neq k \)), thereby generating the multiple pairwise local correlation features \( \mathbf{W}^{kij} \in \mathbb{R}^{C \times H \times W} \). Then, these features \( \mathbf{W}^{kij} \) regarding image \( \mathbf{I}^k \) are integrated by progressive 3-D convolution layers to obtain its local interimage correspondence \( \mathbf{P}^k \in \mathbb{R}^{C \times H \times W} \).

The CoSOD task needs to determine the common attributes of objects by modeling the correspondence among images and further select the common and salient objects. For a pair of images, the extraction of the interimage relationship becomes simpler and clearer, which can be defined as the correlation between two image features. If the correlation of a location is high, it means that the location is more likely to have a common object. Specifically, we design a PCT to explore similarity correspondence between pairwise images, as shown in the right side of Fig. 3. Taking the current intrafeatures \( \mathbf{F}_{ia}^k \in \mathbb{R}^{C \times H \times W} \) and reference intrafeatures \( \mathbf{F}_{ia}^j \in \mathbb{R}^{C \times H \times W} \) as...
the inputs of the PCT, we first transform each of them into a new feature subspace through a 1 × 1 convolution layer, respectively. Next, a matrix multiplication operation is performed for computing the affinity matrix $A^{ij} ∈ ℝ^{HW × HW}$

$$A^{ij} = γ_i(f_{1} × 1(f^{ia}_{i} \cdot f^{ja}_{j}))^T ⊙ γ_j(f_{1} × 1(f^{ia}_{i} \cdot f^{ja}_{j}))$$  \( (3) \)

where $f_{1} × 1$ is the 2-D convolutional layer with the filter size of 1 × 1, ⊙ represents the matrix multiplication, superscript T denotes the transposition, and $γ_i(·)$ reshapes the features into the dimension of $ℝ^{C×HW}$. The element $a^{ij}_{pq}$ in affinity matrix $A^{ij}$ represents the correlation probability between the location $p$ in features $F^{i}_{ia}$ and location $q$ in features $F^{j}_{ja}$, and a large response corresponds to a strong correlation. Furthermore, we need to measure the global correlation of each spatial position in image $I^k$ relative to all positions in image $I$. To this end, we apply the max pooling on the affinity matrix $A^{ij}$ by row to obtain an affinity vector $\bar{A}^{ij} ∈ ℝ^{HW × 1}$, then normalize and reshape it into a location-correlation weight map $\bar{A}^{ij} ∈ ℝ^{1 × HW}$

$$\bar{A}^{ij} = γ_2(softmax(maxpool(A^{ij})))$$  \( (4) \)

where maxpool(·) is the max pooling along the row, softmax(·) denotes the softmax layer for feature normalization, and $γ_2(·)$ reshapes the features into the dimension of $ℝ^{1 × HW}$. Finally, we combine the location-correlation weight map $\bar{A}^{ij}$ with the original features $F^{i}_{ia}$ in a residual connection manner, and use CA [61] and SA [62] for feature enhancement, thereby obtaining the pairwise correlation features $W^{ij}$ of the features $F^{i}_{ia}$ relative to the features $F^{j}_{ja}$

$$W^{ij} = SA(CA(\bar{A}^{ij} ⊙ F^{i}_{ia} ⊕ F^{j}_{ja}))$$  \( (5) \)

where ⊕ means elementwise multiplication broadcasted along feature planes, and CA(·) and SA(·) are the channel attention and spatial attention, respectively.

Repeating the above procedures, we can obtain $N − 1$ pairwise correlation features $W^{ij}$ ($i ∈ \{1, 2, …, N\}$ and $j ≠ k$) of image $I^k$. Then, they are fully integrated by progressive 3-D convolution layers with the filter size of 2 × 3 × 3 to obtain the local interfeatures $P^{i}_{k}$ of image $I^k$.

D. Global-and-Local Correspondence Aggregation

Both the GCM module and the LCM module can extract the interimage relationship among images, but their difference is that the global interfeatures $G$ describe the global correspondence among multiple images in the group, while the local interfeatures $P = \{P^{i}_{k}\}_{k=1}^{N}$ obtain the local relationship from multiple pairwise images. These two modules define the interimage relationship from different angles, and there is a certain degree of complementarity. Thus, we combine them to learn more comprehensive interfeatures $F^{ic}_{k} ∈ ℝ^{C×H×W}$ of image $I^k$ through a GLA module. Concretely, taking into account the advantages of 3-D convolution, we still use it for fusion learning. The GLA consists of a 3-D convolutional layer with the filter size of 2 × 3 × 3, the channel attention scheme, and a spatial attention scheme. The global interfeatures $G ∈ ℝ^{C×H×W}$ and local interfeatures $P^{i}_{k} ∈ ℝ^{C×H×W}$ are first concatenated along the channel dimension as the inputs of the GLA module. Then, we can obtain the final interfeatures $F^{ic}_{k}$ via a 3-D convolutional layer with ReLU activation as follows:

$$F^{ic}_{k} = SA(CA(δ(f_{2} × 3 × 3([G, P^{i}_{k}]))))$$  \( (6) \)

where $[·, ·]$ denotes the concatenation along the channel dimension.

E. Intra-and-Inter Weighting Fusion

As described previously, we construct the interfeatures $F^{ic}_{k}$ from global and local perspectives, aiming to capture the corresponding relationships among group images. Returning to the definition of CoSOD, we need to consider both the saliency attributes within an image and the correspondence among different images. Therefore, we should combine intrafeatures and interfeatures to obtain the co-saliency features. But for different scenes, intrafeatures and interfeatures play different roles, so we design a dynamic weighting strategy to adaptively fuse them instead of the clumsy addition or concatenation fusion strategies. Concretely, we first concatenate the features $F^{ic}_{k}$ and $P^{i}_{k}$, and apply a 1 × 1 convolutional layer for channel reduction to generate $F^{ic}_{c} ∈ ℝ^{C×H×W}$. Then, we learn a weighting map $α^{k} ∈ ℝ^{C×H×W}$ through a bottleneck convolutional layer, which determines the importance between intra and interfeatures. Therefore, the final co-saliency features $F^{co}_{k}$ of image $I^k$ can be obtained by

$$F^{co}_{k} = α^{k} ⊗ F^{ic}_{k} + (1 − α^{k}) ⊗ F^{i}_{ia}$$  \( (7) \)

$$α^{k} = σ(f_{p}(CA(F^{ic}_{c})))$$  \( (8) \)

where $f_{p}$ is a bottleneck convolutional layer, and $σ$ represents the sigmoid activation. For the final co-saliency map $M^{k}$ prediction, we need to upsample the co-saliency features $F^{co}_{k}$ to the same resolution with the input image, for which multilayer deconvolution operations are adopted.

F. Loss Function

During training, the proposed network is optimized by using the binary cross-entropy loss between each co-saliency map $M^{n}$ and the ground truth $T^{n}$ in an end-to-end manner. Specifically, given $N$ co-saliency maps and ground-truth masks (i.e., $\{M^{n}\}_{n=1}^{N}$ and $\{T^{n}\}_{n=1}^{N}$) in an image group, we define the loss function $ℓ$ by calculating the binary cross-entropy loss over all the samples

$$ℓ = \frac{1}{N} \sum_{n=1}^{N} ℓ^{bce}_{n}$$  \( (9) \)

where $ℓ^{bce}_{n} = −[T^{n}\log(M^{n}) + (1 − T^{n})\log(1 − M^{n})]$ is the binary cross-entropy loss of sample $n$, and $N$ denotes the image number in an image group.

IV. EXPERIMENTS

In this section, we first introduce the datasets, the evaluation, and the implementation details. Then, we make some comparisons between the proposed GLNet and the state-of-the-art (SOTA) SOD and CoSOD methods to demonstrate our superiority. Finally, the ablation studies will be discussed to investigate the effectiveness of each module.


A. Datasets and Evaluation Metrics

1) Benchmark Datasets: We evaluate different methods on three widely used CoSOD benchmark datasets, including: 1) MSRC [68]; 2) iCoseg [74]; and 3) Cosal2015 [65]. Each image in these datasets has a corresponding pixelwise CoSOD ground truth. Among them, the MSRC dataset [68] is the first available dataset for CoSOD and consists of 210 images in seven groups after removing the nonsaliency grass group. The iCoseg dataset [74] contains 643 images distributed in 38 image groups, and the number of images in each group varies from 4 to 42. The Cosal2015 dataset [65] is a large and challenging dataset containing 2015 images in 50 groups, which includes various challenging factors, such as complex background and occlusion issues.

2) Evaluation Metrics: For quantitative evaluation, we introduce five widely used evaluation indicators in the SOD background and occlusion issues.

For F-measure ($F_{β}) [75]$: 3) MAE score [76]; 4) S-measure ($S_o)$ [77]; and 5) E-measure ($E_p)$ [78]. The P-R curve intuitively shows the variation trend between different precision and recall scores, and the closer the curve is to (1, 1), the better the method performance. F-measure ($F_{β}$) is a comprehensive measurement of precision and recall values. The specific formula is written as follows:

$$F_{β} = \frac{(1 + \beta^2)\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$$

where precision and recall are the precision score and recall score, respectively, and $\beta^2$ is fixed to 0.3 as suggested in [75].

MAE score is used to calculate the pixelwise difference between the predicted co-saliency map $M$ and ground truth $T$, which is formalized as

$$\text{MAE} = \frac{1}{W \times H} \sum_{w=1}^{W} \sum_{h=1}^{H} |M(w, h) - T(w, h)|$$

where $W$ and $H$ are the width and height of the image, respectively.

S-measure ($S_o$) describes the structural similarity between the co-saliency map and ground truth, which is defined as

$$S_o = \alpha \times S_o + (1 - \alpha) \times S_r$$

where $S_o$ represents the object-aware structural similarity, summing foreground and background comparison terms with weights. $S_r$ denotes the region-aware structural similarity, which divides the saliency map and ground truth into multiple regions and computes the weighted summation of the corresponding structural similarity measure. $\alpha$ is the balance parameter and set to 0.5 as suggested in [77].

E-measure ($E_p$) consists of a single term to account for both pixel and image-level properties, which is computed as

$$E_p = \frac{1}{W \times H} \sum_{w=1}^{W} \sum_{h=1}^{H} \theta(\phi)$$

where $\theta(\phi)$ indicates the enhanced alignment matrix.

Among these indicators, the larger the $F_{β}$, $E_p$, and $S_o$, the better the performance, while the MAE value is just the opposite (i.e., the smaller value corresponds to the better performance).

B. Implementation Details

1) Training Data: Most of the previous state-of-the-art CoSOD models are trained on the large-scale COCO-SEG dataset [26], which contains more than 200,000 image samples. As pointed out in [27], although there are a large amount of data in COCO-SEG, many objects in this dataset do not have the saliency attributes, so it is not completely suitable for the CoSOD task. Following the settings in [27], we take the CoSOD3k dataset [73] as our training dataset. The CoSOD3k dataset is specially designed for the CoSOD task, which contains 3316 images distributed in 160 image groups.

2) Implementation Settings: We implement the proposed model via PyTorch toolbox and train it on an RTX 2080Ti GPU in an end-to-end manner. We also implement our network by using the MindSpore Lite tool. In order to avoid overfitting, we use random flipping and rotating to augment the training samples. Following the input settings in [25] and [27], we set the input image number of the GLNet to 5, and select minibatch groups from all categories in the CoSOD3k dataset. Due to the limited computing resources, all input images are resized to $160 \times 160$, and rescaled to the original size for testing. We use Adam [79] to train our model with weight decay of $10^{-4}$ and the training process converges until 40,000 iterations. The learning rate decay strategy adopts cosine annealing decay, where the initial and minimum learning rate are set to $5e^{-6}$ and $5e^{-7}$, respectively. The code and results can be found at https://rmcong.github.io/proj_GLNet.html.

3) Backbone Setting: In our experiment, we choose the VGG16 network as our feature extractor for two reasons. First, most of the previous methods used the VGG as the backbone, such as GCAGC [30], ICNet [36], and so on. Therefore, for a fair comparison, we also choose the VGG network to extract the multilevel features of each image in the group, which is the dominant reason. Second, the CoSOD task deals with image group data, using other more powerful backbone networks (e.g., Dilated ResNet [58]) may significantly increase the model size. Taken together, after a tradeoff, we chose VGG as the backbone in the experiments.

C. Comparison With State-of-the-Art Methods

In order to demonstrate the effectiveness of GLNet, we compare it with 11 SOTA methods on the above three datasets, including two SOD methods for single image (e.g., GCPANet [64], and CPD [63]), and nine CoSOD methods (e.g., GICD [29], GCAGC [30], CODW [65], UMLF [67], CSMG [69], RCGS [26], CoADNet [27], ICNet [36], and GCoNet [28]). For a fair comparison, all results are directly provided by authors or are reproduced by the public source codes under the default parameters.

Fig. 4 shows the visual comparison results of different CoSOD methods. Intuitively, we can see that our method locates the co-salient objects more accurately and completely than other competing methods. For example, in the third group (i.e., goose), only our method can accurately locate the co-salient object in each image and show superiority in terms of

\[1\]https://www.mindspore.cn/
the internal consistency, even if the co-salient object is partially occluded. When there are obvious appearance changes, strong semantic interference, and complex backgrounds in the scene (such as the first group of axes in Fig. 4), our model can still accurately, completely, and clearly detect the co-salient object, benefiting from the comprehensive modeling of the interactive information among images.

For quantitative evaluation, Fig. 5 shows the P-R curves of all compared methods on three benchmark datasets. We can observe that our GLNet outperforms the other SOTA methods on all datasets. In particular, the P-R curves are much higher than the other methods on the challenging and largest Cosal2015 dataset. Meanwhile, Table I lists the quantitative results. In general, we can notice that our GLNet achieves better performance than most of comparison methods in terms of three evaluation indicators. For example, compared to the second-best method on the Cosal2015 dataset, the percentage gain of \( F \)-measure is 4.00%, and the MAE score reaches 2.85%. Our GLNet model achieves the best \( F \)-measure of 0.8945 and 0.8880 on the MSRC and iCoseg datasets, respectively, where the performance gains on the MSRC dataset and iCoseg dataset, respectively reach 3.69% and 1.09% compared with the second-best method. For the \( E \)-measure, the minimum percentage gain against the comparison methods reaches 3.43% on the Cosal2015 dataset, and 2.28% on the iCoseg dataset, respectively.

We also draw the FPS–\( F \)-measure map in Fig. 6 for a better comprehensive evaluation of performance and time on the Cosal2015 dataset. Generally speaking, more than 30 FPS can be considered real time, and our model reaches 33 FPS, which
meets the real-time requirement. As shown in Fig. 6, the GLNet achieves up to 5% gain in $F$-measure compared with the existing models with similar speed. In summary, our method strikes a tradeoff between performance and real-time capability.

D. Ablation Study

We conduct thorough ablation studies to analyze the effects of the key modules in our GLNet on three datasets, and the visualization and quantitative results are shown in Fig. 7 and Table II. The specific experimental settings are as follows.

1) *full model* (model 1) is the full model containing all the modules proposed in this article.
2) *w/o (GCM + LCM)* (model 2) removes the GCM and LCM modules from the full model and degenerates the full model into a single-image SOD model.
3) *w/o LCM* (model 3) represents that only the LCM module is removed and the GCM module is retained in the full model.
4) *w/o GCM* (model 4) means that only the GCM module is removed and the LCM module is retained in the full model.
5) *w/o GCM-3D* (model 5) refers to that only the 3-D convolution in the GCM module of the full model is replaced by the ordinary 2-D convolution while still retaining the LCM module.

First, we remove the GCM and LCM modules, so that the full model degenerates into a single-image SOD model. As shown in the fourth column of Fig. 7, although it can detect the salient objects, there are a large number of backgrounds and noncommon salient objects are wrongly preserved, such as the towel in the first image, the red sofa in the second image, etc.
and the sticker on the wall in the third image. Similarly, it performs poorly in various indicators. It can be seen from Table II that compared to the full model (model 1), the $S$-measure, $F$-measure, $E$-measure, and MAE score of w/o (GCM+LCM) (model 2) are decreased by 6.59%, 5.80%, 4.44%, and 42.14% on the Cosal2015 dataset, respectively.

Then, we investigate the effects of the two ways of inter-image relationship modeling, that is: 1) GCM and 2) LCM. From Fig. 7, we can observe that GCM and LCM have different performances in objects positioning and background suppression, such as the towel and sticker in the third image. Moreover, it can be observed from Fig. 7 that when LCM and GCM are organically combined, the noncommon salient region is effectively suppressed (e.g., towel, sofa, and sticker), thereby further improving the performance. By comparing various indicators, it can be found that the performance of using LCM or GCM alone is decreased compared to the full model (model 1). For example, the MAE score of the model w/o LCM (model 3) is decreased by 17.24% on the Cosal2015 dataset, and the MAE score of the model w/o GCM (model 4) is dropped by 10.23% on the MSRC dataset.

Finally, let us verify the effect of 3-D convolution in the GCM module. Compared to the full model (model 1), all indicators in w/o GCM-3-D (model 5) are obviously reduced on the three datasets, especially for the $S$-measure and $F$-measure on the Cosal2015 dataset. Specifically, the $S$-measure drops from 0.8460 to 0.8150, and the $F$-measure decreases from 0.8936 to 0.8657. It further shows that the 3-D convolution used in the GCM module can more efficiently extract multiple image relations at once, with clear advantages over traditional 2-D convolution.

In summary, the ablation studies have proved the effectiveness and advantages of proposed modules from both qualitative and quantitative aspects, including the LCM, GCM, and 3-D convolution in GCM.

E. Future Work

In future work, three aspects of CoSOD can be focused on. One is a large-scale and appropriate dataset might improve the accuracy. Currently, CoSOD3k is the largest one tailored for the CoSOD task, but also contains only 3316 images. Compared with other SOD tasks, there is much room for efforts in the dataset scale. This is also the focus of our future work, building a large-scale and specialized dataset to further promote the development of the CoSOD research community. Another is the design of the model. Our current network uses a fixed number of image group inputs, which makes the network less flexible. Therefore, at the model design level, we will further explore the recurrent architecture to solve this problem and improve the detection performance. Meanwhile, different CNN models also have a certain impact on performance, as explored in [80]. This is also an important aspect of the design of the model in the future. Last but not the least, although our model reaches a real-time level of 33 FPS, it still has room for optimization. In the future, more attention can be paid to lightweight model design to better adapt to practical applications.

V. Conclusion

In this article, an end-to-end GLNet network was proposed to locate the co-salient objects in an image group by capturing the interimage information from both global and local perspectives. Our architecture contains two key modules for interimage correspondence extraction, that is: 1) global correspondence modeling (GCM) and 2) local correspondence modeling (LCM). The GCM directly captures the global group semantics by using the 3-D convolution, and the LCM explores the similarity correspondence between pairwise images through the designed PCT. The two complement each other and jointly improve the accuracy of the modeling interimage relationship. Experiments on three prevailing CoSOD datasets showed that our GLNet outperforms 11 SOTA competitors, even if our model is trained on a small dataset.

Acknowledgment

Runmin Cong is with the Institute of Information Science and the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China, and also with the Department of Computer Science, City University of Hong Kong, Hong Kong, SAR, China (e-mail: rmccong@bjtu.edu.cn).

Ning Yang and Yao Zhao are with the Institute of Information Science and the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: ningyang@bjtu.edu.cn; yzhao@bjtu.edu.cn).
REFERENCES

[1] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, “Review of visual saliency detection with comprehensive information,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[2] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, “Visual attention consistency under image transforms for multi-label image classification,” in Proc. CVPR, 2019, pp. 729–739.

[3] S. Han and N. Vasconcelos, “Image compression using object-based regions of interest,” in Proc. ICIP, 2006, pp. 3097–3100.

[4] R. Gallego, A. Ardizzone, and R. Pirrone, “Physical metaphor for streaming media retargeting,” IEEE Trans. Multimedia, vol. 16, no. 4, pp. 971–979, Jun. 2014.

[5] D. Jing, S. Zhang, R. Cong, and Y. Lin, “Occlusion-aware bi-directional guided network for light field salient object detection,” in Proc. ACM MM, 2021, pp. 1692–1701.

[6] H. Wen et al., “Dynamic selective network for RGB-D salient object detection,” IEEE Trans. Image Process., vol. 30, pp. 9179–9192, 2021.

[7] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, “Saliency detection for stereo images based on depth confidence analysis and multiple cues fusion,” IEEE Signal Process. Lett., vol. 23, no. 6, pp. 819–823, Jun. 2016.

[8] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, “Going from RGB to RGBD saliency: A depth-guided transformation model,” IEEE Trans. Cybern., vol. 50, no. 8, pp. 3627–3639, Aug. 2020.

[9] Z. Chen, R. Cong, Q. Xu, and Q. Huang, “DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection,” IEEE Trans. Image Process., vol. 30, pp. 7012–7024, 2021.

[10] G. Zhang et al., “Cross-modality discrepant interaction network for RGB-D salient object detection,” in Proc. ACM MM, 2021, pp. 2094–2102.

[11] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, “RGB-D salient object detection with cross-modality modulation and selection,” in Proc. ECCV, 2020, pp. 225–241.

[12] C. Li et al., “ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection,” IEEE Trans. Cybern., vol. 50, no. 1, pp. 88–100, Jan. 2021.

[13] Y. Mao, Q. Jiang, R. Cong, W. Gao, F. Shao, and S. Kwong, “Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3D images,” IEEE Trans. Multimedia, early access, May 19, 2021, doi: 10.1109/TMM.2021.3081260.

[14] J. Lei et al., H. Fu, F. Porikli, Q. Huang, and H. Hou, “Video saliency detection via sparsity-based reconstruction and propagation,” IEEE Trans. Image Process., vol. 28, no. 10, pp. 4819–4931, Oct. 2019.

[15] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, “Nested network with two-stream pyramid for salient object detection in optical remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 11, pp. 9156–9166, Nov. 2019.

[16] Q. Zhang et al., “ Dense attention fluid network for salient object detection in optical remote sensing images,” IEEE Trans. Image Process., vol. 30, pp. 1305–1317, 2021.

[17] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, “RRNet: Relational reasoning network with parallel multi-scale attention for salient object detection in optical remote sensing images,” IEEE Trans. Trans. Cybern., vol. 60, pp. 5584–5594, Feb. 2022.

[18] X. Zhou et al., “Edge-guided recurrent positioning network for salient object detection in optical remote sensing images,” IEEE Trans. Cybern., early access, Apr. 13, 2022, doi: 10.1109/TCYB.2022.3163152.
[46] G. Li and Y. Yu, “Deep contrast learning for salient object detection,” in Proc. CVPR, 2016, pp. 478–487.

[47] D. Zhang, H. Tian, and J. Han, “Few-cost salient object detection with adversarial-paced learning,” in Proc. NeurIPS, 2020, pp. 12236–12247.

[48] D. Zhang, J. Han, Y. Zhang, and D. Xu, “Synthesizing supervision for learning deep saliency network without human annotation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 7, pp. 1755–1769, Jul. 2020.

[49] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, “Deeply supervised salient object detection with short connections,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 4, pp. 815–828, Apr. 2019.

[50] J. Wei, S. Wang, and Q. Huang, “F³Net: Fusion, feedback and focus for salient object detection,” in Proc. AAAI, 2020, pp. 12321–12328.

[51] X. Hu, L. Zhu, J. Qin, C. W. Fu, and P. A. Heng, “Recurrently aggregating deep features for salient object detection,” in Proc. AAAI, 2018, pp. 6943–6950.

[52] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, “Salient object detection with recurrent fully convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1734–1746, Jul. 2019.

[53] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J.-F. Yang, and M.-M. Cheng, “EGNet: Edge guidance network for salient object detection,” in Proc. ICCV, 2019, pp. 8779–8788.

[54] Y. Liu, M.-M. Cheng, X.-Y. Zhang, G.-Y. Nie, and M. Wang, “DNA: Deeply supervised nonlinear aggregation for salient object detection,” IEEE Trans. Cybern., early access, Feb. 2, 2021, doi: 10.1109/TCYB.2021.3051350.

[55] H. Li, G. Li, B. Yang, G. Chen, L. Lin, and Y. Yu, “Depthwise non-local module for fast salient object detection using a single thread.” IEEE Trans. Cybern., vol. 51, no. 12, pp. 6188–6199, Dec. 2021, doi: 10.1109/TCYB.2020.2969282.

[56] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015, pp. 1–14.

[57] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, 2016, pp. 770–778.

[58] F. Yu, V. Koltun, and T. Funkhouser, “Dilated residual networks,” in Proc. CVPR, 2017, pp. 636–644.

[59] Q. Li, Q. Wang, and X. Li, “Exploring the relationship between 2D/3D convolution for hyperspectral image super-resolution,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 10, pp. 8693–8703, Oct. 2021.

[60] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, “Scrubble-supervised video object segmentation,” IEEE/CAS J. Automatica Sinica, vol. 9, no. 2, pp. 339–353, Feb. 2022.

[61] H. Jie, S. Li, S. Gang, and S. Albanie, “Squeeze-and-excitation networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[62] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: convolutional block attention module,” in Proc. ECCV, 2018, pp. 3–19.

[63] Z. Wu, L. Su, and Q. Huang, “Cascaded partial decoder for fast and accurate salient object detection,” in Proc. CVPR, 2019, pp. 3907–3916.

[64] Z. Chen, Q. Xu, R. Cong, and Q. Huang, “Global context-aware progressive aggregation network for salient object detection,” in Proc. AAAI, 2020, pp. 10599–10606.

[65] D. Zhang, J. Han, C. Li, and J. Wang, “Co-saliency detection via looking deep and wide,” in Proc. CVPR, 2015, pp. 2994–3002.

[66] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. Li, “ImageNet: A large-scale hierarchical image database,” in Proc. CVPR, 2009, pp. 248–255.

[67] 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, Oct. 2018.

[68] J. Winn, A. Criminisi, and T. Minka, “Object categorization by learned universal visual dictionary,” in Proc. ICCV, 2005, pp. 1800–1807.

[69] K. Zhang, T. Li, B. Liu, and Q. Liu, “Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing,” in Proc. CVPR, 2019, pp. 3095–3104.

[70] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, “Learning to detect a salient object,” in Proc. CVPR, 2007, pp. 1–8.

[71] L. Wang et al., “Learning to detect salient objects with image-level supervision,” in Proc. CVPR, 2017, pp. 3796–3805.

[72] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. ECCV, 2014, pp. 740–755.

[73] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, “Taking a deeper look at the co-salient object detection,” in Proc. ICCV, 2020, pp. 291–2926.

[74] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, “iCoseg: Interactive co-segmentation with intelligent scribble guidance,” in Proc. CVPR, 2010, pp. 3169–3176.

[75] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in Proc. CVPR, 2009, pp. 1597–1604.

[76] D. Jeong, I. Hwang, and N. I. Cho, “Co-salient object detection based on deep saliency networks and seed propagation over an integrated graph,” IEEE Trans. Image Process., vol. 27, no. 12, pp. 5866–5879, Dec. 2018.

[77] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in Proc. ICCV, 2017, pp. 4548–4557.

[78] 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. IJCAI, 2018, pp. 698–704.

[79] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[80] F. Furlán, E. Rubio, H. Sossa, and V. Ponce, “CNN based detectors on planetary environments: A performance evaluation,” Front. Neurobot., vol. 14, Oct. 2020, Art. no. 590371.
Huazhu Fu (Senior Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2013.

He is a Senior Scientist with the Institute of High Performance Computing, A*STAR, Singapore. Previously, he was a Research Fellow with Nanyang Technological University, Singapore, from 2013 to 2015, a Research Scientist with I2R, A*STAR from 2015 to 2018, and a Senior Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE, from 2018 to 2021. His research interests include computer vision, machine learning, and AI in healthcare.

Dr. Fu received the Best Paper Award from ICME 2021. He has served as the Associate Editor for IEEE TRANSACTIONS ON MEDICAL IMAGING and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, the AC/Senior-PC for MICCAI, IJCAI, and AAAI, and the General Chair for MICCAI Workshop on OMIA. He is a member of the IEEE BISP TC.

Yao Zhao (Senior Member, IEEE) received the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996.

He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, video analysis and understanding, and artificial intelligence.

Dr. Zhao serves on the editorial boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, a Senior Associate Editor of the IEEE SIGNAL PROCESSING LETTERS, and an Area Editor of the Signal Processing: Image Communication. He was named as a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of the Ministry of Education of China in 2013.

Prof. Huang is the Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and Acta Automatica Sinica, and the Reviewer of various international journals, including the IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON IMAGE PROCESSING. He has served as a General Chair, a Program Chair, a Track Chair, and a TPC Member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PCM, and PSIVT.

Sam Kwong (Fellow, IEEE) received the Ph.D. degree from the University of Hagen, Hagen, Germany, in 1996.

From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada, Ottawa, ON, Canada. He joined Bell Northern Research Canada, Ottawa, as a member of Scientific Staff. In 1990, he became a Lecturer with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, where he is currently a Professor with the Department of Computer Science. His research interests are video and image coding and evolutionary algorithms.