Adaptive Group-wise Consistency Network for Co-saliency Detection

Zhen Bai, Zhi Liu, *Senior Member, IEEE*, Gongyang Li, and Yang Wang

***Abstract*—Co-saliency detection focuses on detecting common and salient objects among a group of images. With the application of deep learning in co-saliency detection, more accurate and more effective models are proposed in an end-to-end manner. However, two major drawbacks in these models hinder the further performance improvement of co-saliency detection: 1) the static manner-based inference, and 2) the constant quantity of input images. To address these limitations, we present a novel Adaptive Group-wise Consistency Network (AGCNet) with the ability of content-adaptive adjustment for a given image group with random quantity of images. In AGCNet, we first introduce intra-saliency priors generated from any off-the-shelf salient object detection model. Then, an Adaptive Group-wise Consistency (AGC) module is proposed to capture group consistency for each individual image, and is applied on three-scale features to capture the group consistency from different perspectives. This module is composed of two key components, where the content-adaptive group consistency block breaks the above limitations to adaptively capture the global group consistency with the assistance of intra-saliency priors and the ranking-based fusion block combines the consistency with individual attributes of each image feature to generate discriminative group consistency feature for each image. Following AGC modules, a specially designed Aggregated Decoder aggregates the three-scale group consistency features to adapt to co-salient objects with diverse scales for preliminary detection. Finally, we incorporate two normal decoders to progressively refine the preliminary detection and generate the final co-saliency maps. Extensive experiments on four benchmark datasets demonstrate that our AGCNet achieves competitive performance as compared with 19 state-of-the-art models, and the proposed modules experimentally show substantial practical merits.**

***Index Terms*—Co-saliency detection, group consistency, intra-saliency priors, content-adaptive layer, semantic information.**

I. Introduction

SALIENCY detection simulates the human visual attention mechanism during free-viewing within a single image to rapidly focus on the most attractive regions [1], [2]. As an extended branch of saliency detection, co-saliency detection explores the most repeatedly occurring salient objects with the same attributes across a group of relevant images.

This work was supported in part by the National Natural Science Foundation of China under Grant 62171269, and in part by the China Scholarship Council under Grant 202006890079. *(Corresponding author: Zhi Liu.)*

Zhen Bai, Zhi Liu, and Gongyang Li are with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China. Gongyang Li is also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: bz536476@163.com; liuzhisjtu@163.com; ligongyang@shu.edu.cn).

Yang Wang is with the Department of Compter Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada (email: ywang@cs.umanitoba.ca).

The most important information of the group of images can be represented by the extracted co-occurring patterns or the prime objects within contexts of the group [3], [4]. With this ability, co-saliency detection is widely used as an inherent part in many applications, such as image co-segmentation [5]–[7], image co-localization [8], and image retrieval [9].

Compared with saliency detection in a single image, mining the interaction of a group of images is a further essential step for co-saliency detection. Existing co-saliency detection models [10]–[25] generally focus on tackling two key issues to ensure that the detected objects are salient and similar with each other: 1) extracting representative features to characterize salient objects and 2) mining group consistency.

At the early stage, co-saliency detection models [10], [12], [13], [24], [26] are mainly based on handcrafted features, and assume that the co-salient objects in multiple related images should share certain shallow-level consistency. Researchers have designed some constraints to capture the group consistency of a given image group, *e.g.*, Kullback-Leibler divergence [26], cluster [10], and manifold ranking [12], [13], [24]. However, these models cannot sufficiently capture high-level object semantics, and the handcrafted features are unstable on complex scenes, *e.g.*, when there is a large appearance variance of co-salient objects across images, or when the co-salient objects are similar to the background. These factors often lead to poor performance.

Recently, the deep-learning based co-saliency detection models [4], [18]–[23], [27]–[29] demonstrate more powerful performance than handcrafted feature-based co-saliency detection models [30]. These models extract Convolutional Neural Network (CNN) features and model collaborative relationships of features from group-wise and single images, and obtain promising results. However, there are two major limitations in these models, which hinder the further performance improvement of co-saliency detection:

First, most end-to-end models [18], [20]–[23] capture group consistency in a static manner, where the model parameters are generally fixed once trained, which reduces their generalizability of handling objects of unseen categories. We take some results of IML [21] as examples in Fig. 1. Due to the training dataset including human category, the person who appears in the fifth image is wrongly identified by IML [21] as co-salient object. The real co-salient objects are cricket balls, which are not detected in the first image by IML model. Second, most end-to-end co-saliency detection models [20], [21], [23] are limited by a constant quantity of input images. Whether in the training or inference stage, the complete image groups have to be divided into some subgroups with

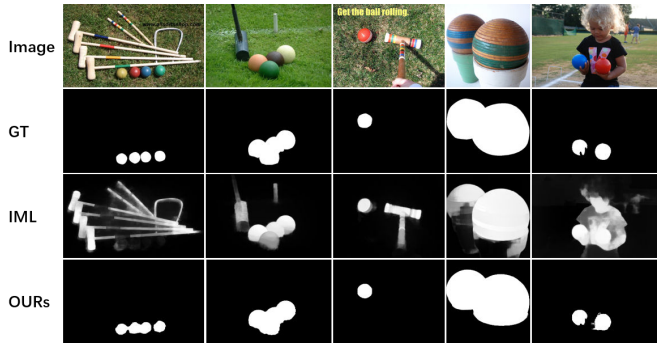

Fig. 1. Illustrations for the problems of existing end-to-end models to detect co-salient objects. Examples of cricket ball group in CoSOD3k [30]. GT represents ground truth.

a specified number of images. With different partial image groups as inputs, the global consistency capture within a complete image group is deficient for these models, resulting in the inconsistent inference for different combinations. When the number of images in a subgroup is fixed to a small value, some frequently occurring attributes will be enlarged. From the results of IML [21] in Fig. 1, when the cricket balls are highlighted, the mallets that often accompany them are mistakenly highlighted by IML. The root cause is that the ways (*e.g.*, feature concatenation) of modeling the inter-correlation among multiple images and combining the individual attributes with inter-correlation are improper.

To break the above limitations, we propose an end-to-end Adaptive Group-wise Consistency Network (AGCNet) to detect co-salient objects within an image group, where the number of images in the group can be variable. The key idea of AGCNet is to adaptively capture the global group consistency and integrate it with attributes of each individual image for more discriminative reflecting on each image of the group. Specifically, in light of that state-of-the-art image salient object detection (SOD) models can achieve comparable performance in co-saliency detection, we introduce the results of any off-the-shelf image SOD models into our model. In our view, the extracted salient regions from SOD models can be taken as the intra-saliency priors to conduct the semantic matching within an image group, modeling the inter-semantic correlation for capturing the global group consistency. Moreover, considering that co-salient objects in a group of images often belong to the same semantic category but vary in terms of some individual attributes [31], we further combine the global group consistency with individual attributes to keep the internal coherence of any co-salient objects. Besides, in light of that the co-salient objects of a group vary in scale even within the same category and existing attribute gap between different scales, we explore local and global complementation for different-scale group consistency in the decoder. Finally, on account of the excellent performance of the existing image SOD models in object edge detection, we combine group consistency with intra-saliency priors to promote the detected co-salient objects to retain fine edges.

In particular, our AGCNet is composed of three parts: an encoder for basic feature extraction, three Adaptive Group-wise Consistency (AGC) modules for modeling collaborative relationships among a group of image features, and a hybrid decoder for co-saliency reasoning. Specifically, our AGC module consists of a Content-adaptive Group Consistency (CGC) block and a Ranking-based Fusion (RF) block. The CGC block is capable of adaptively capturing more flexible global group consistency for current input group images. The intra-saliency priors are employed to mask the basic features to generate intra-semantic features as convolution kernels for a dynamic convolutional layer [32], which through filtering achieves point-to-point semantic matching for the given image group with random quantity. Subsequently, a weighted summation is applied on the matching information to generate the group consistency map, potentially scoring co-saliency for each pixel of each image. RF block targets for achieving the combination of global group consistency and attributes of each individual image. To ensure the internal consistency of each image, RF block further mines the self-correlation relationship within pixels of each image through an affinity matrix for the representation of the individual attributes. Finally, this block combines the group consistency map with the individual attributes to rearrange the channel group consistency features to drive the pixels of co-salient objects being highlighted more uniformly for subsequent reasoning. These two blocks mainly involve the operations of dynamic convolution, multiplication, addition and rearrange. Thus, this module can be applied to the input group with random quantity of images, and is content adaptive to the input image group, without bias to the category attributes of training data. The hybrid decoder includes an Aggregated Decoder and two normal decoders. Our Aggregated Decoder aggregates three-scale group consistency features through three Attentional Feature Fusion (AFF) blocks for preliminary detection of co-salient objects. Based on the attentional feature fusion mechanism, the AFF block globally and locally fuses group consistency features at different scales to remedy group information inconsistency. Since the group consistency is captured based on relatively high-level features, the proposed aggregated decoder that aggregates different-scale group consistency can only locate co-salient objects in the form of low-resolution co-saliency maps. Thus, we adopt two normal decoders to improve the integrity of co-salient objects by broadcasting the preliminary co-saliency to shallow-level features and further combine the intra-saliency priors for progressive refinement.

Our contributions can be summarized as follows:

- We propose an end-to-end Adaptively Group-wise Consistency Network, which introduces the intra-saliency priors generated from any off-the-shelf image SOD models, for co-saliency detection in an image group with random quantity. Our AGCNet can adaptively capture group consistency and aggregate multi-scale group consistency, highlighting co-salient objects with fine structures.
- We propose an Adaptively Group-wise Consistency module to model collaborative relationships among a group of image features in a dynamic manner. This module can adaptively achieve semantic matching with the assistance of intra-saliency priors and capture global group consistency. It can further integrate the self-correlation

of individual image with the global group consistency to generate discriminative group consistency features for each image.
- We propose an Aggregated Decoder to promote complementarity between different-scale group consistency features. This module globally and locally aggregates features of inconsistent semantics and scales to preliminary detect co-salient objects with various scales.

II. Related works

A. Image Salient Object Detection

Most early image SOD models [1], [33]–[40] adopted bottom-up strategy and generally based on handcrafted features with cognitive assumptions, *e.g.*, local contrast [36], global contrast [33], and background priors [37], *etc*. Classically, Cheng *et al.* [33] employed color histogram contrast to characterize global contrast to infer saliency. Liu *et al.* [38] extracted center-surround histograms, color spatial distributions and multi-scale contrast features, and adopted the conditional random field algorithm to fuse these for prediction. Besides, frequency domain analysis [40] and low-rank recovery [36] are the commonly used traditional algorithms for image SOD.

The deep learning-based image SOD has attracted lots of research attention and achieved remarkable progress. Early, Han *et al.* [41] followed background prior assumptions to measure the saliency of each region by reconstructing error between detected regions and background regions. Li *et al.* [42] extracted deep features to infer the saliency for each pixel and superpixel, respectively. Whereafter, a number of end-to-end SOD models [2], [43]–[45] were proposed, and they were designed with multi-scale or multi-stream network to learn more comprehensive CNN features. To improve the performance of image SOD, many researches [46]–[50] adopted the edge information as additional auxiliary information. By exploiting the correlation between saliency and contour, Zhou *et al.* [51] designed a two-stream framework to separately generate preliminary saliency map and edge map, and combined these two maps for final prediction. In addition to these image SOD models, Zhang *et al.* [52] proposed a well-performing detection model with weak supervision. Some researchers made attempts to employ top-tier image SOD models on co-saliency detection benchmarks. These image SOD models surprisingly achieved comparable results with deep learning-based co-saliency detection models, such as CPD [53], EGNet [49], and BASNet [50], as mentioned in [30]. This indicates that if the result of each image generated from image SOD models can be employed to support co-saliency detection in a proper way, a more powerful co-saliency detection model will be designed.

B. Co-saliency Detection

Similar to image SOD, traditional co-saliency detection models relied heavily on handcrafted features to characterize co-saliency with manually designed metrics. Jacobs *et al.* [3] defined the co-saliency detection task, and made the first attempt to detect common salient objects in image pair by exploring local variations. Li *et al.* [54] established the first public image pair benchmark. Whereafter, Li *et al.* [55] expanded the co-saliency detection task from image pair to multiple images.

Most traditional co-saliency detection models took the intra-saliency regions generated from existing image SOD models as proposals, then employed various matching techniques, *e.g.*, independent component analysis [26], cluster [10], cellular automata [56], and propagation [57], to capture the inter-correlation of these proposals. In order to promote the integrity of the co-salient objects, many models did not regard inter-correlation as the results, but combined intra-saliency and inter-correlation through diverse fusion techniques, *e.g.*, fixed weight fusion [54], adaptive weight fusion [24], [25], and region-wise adaptive fusion [13], for final inference.

The deep learning-based co-saliency detection models demonstrated more powerful performance than traditional models [30]. The early deep learning-based works only roughly combined CNN features with traditional co-saliency detection models. Based on CNN features, Zhang *et al.* [27] used the clustering algorithm and a principled Bayesian module to infer co-salient objects. Zhang *et al.* [58] applied high-level semantic CNN features and used a self-paced multiple-instance algorithm to capture the group consistency. Yao *et al.* [59] employed spectral rotation co-clustering algorithm twice to divide lots of images into a series of subgroups with similar foreground objects and to segment out the co-salient objects.

Afterwards, deep learning technology was applied in both feature extraction and co-saliency reasoning. Tsai *et al.* [4] proposed an unsupervised CNN based model to adaptively learn the deep features for co-saliency detection. Zhang *et al.* [29] novelly employed gradient information of consensus representation among a group of images to reflect the discriminative co-salient features for co-saliency detection. Hu *et al.* [60] employed Graph Convolutional Network (GNN) [61] to capture common information and regarded the co-saliency detection task as a classification task to conduct the binary classification for superpixels. In addition to these models, a large number of end-to-end co-saliency detection models [18]–[23], [62], [63] have been proposed with various strategies, such as multi-scale inter-correlation propagation [20], [23], RNN [22], [64], co-attention mechanism [63], GNN [19]. Notably, some models [18], [65] introduced extra labels for training. Wang *et al.* [65] extracted the co-category information from group-wise images as the group consistency with the supervision of category labels. However, a fine-tuning process must be used if applying the trained model to an unseen category. Zhang *et al.* [18] designed a collaborative aggregation-and-distribution network to capture both salient and repetitive visual patterns from five images with the supervision of image SOD label and co-saliency detection label.

While most end-to-end models made a great performance progress, some frameworks [18], [21]–[23], [63] are limited by the fixed quantity of input images with the influence of group consistency capture module, and cannot summarize the global shared attributes for the given group with random quantity. Moreover, these models are full of static convolutional layers that perform inference in a static manner with

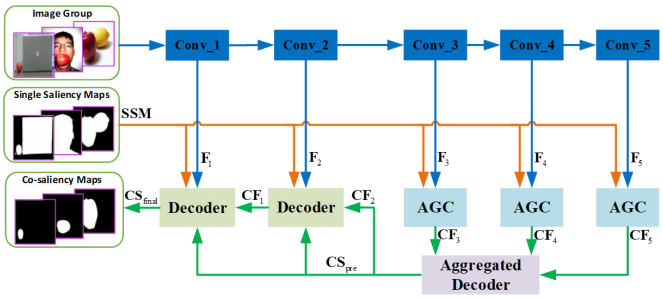

Fig. 2. Architecture overview of the proposed AGCNet. For a given image group with random quantity, we first utilize VGG to obtain the multi-scale CNN features of each image, and introduce corresponding single saliency map **SSM** generated from state-of-the-art image SOD models to AGCNet. Then, three relatively high-level features paired with **SSM** are fed into three Adaptive Group-wise Consistency (AGC) modules, respectively, generating a group of group-wise consistency features at three scales. Subsequently, three-scale group consistency features are aggregated by an Aggregated Decoder to generate a group of preliminary co-saliency maps $\mathbf{CS}_{\text{pre}}$ and composited group consistency features $\mathbf{CF}_2$. Finally, $\mathbf{CF}_2$ is combined with **SSM** to refine $\mathbf{CS}_{\text{pre}}$ through two normal decoders, generating the final co-saliency maps $\mathbf{CS}_{\text{final}}$ with the same sizes as the input images.

fixed parameters, hindering the performance improvement on unseen category of co-salient objects. Although the RNN-based co-saliency detection models [64], [66] made some improvements, the learned group representations vary in different order of the given image groups, resulting in unstable inference. The crucial temporal relations can be constructed by RNN for video sequences, while the relationships do not exist in image groups. Thus, modeling collaborative relationships of the input image groups by RNN for co-saliency detection is sub-optimal.

Compared with existing end-to-end co-saliency detection models, the proposed AGCNet breaks the hindrances of static manner-based inference and the constant image quantity. AGCNet not only fuses the results of existing SOD models as the former works [1], [4], [24], but also introduces the results to guide the dynamic convolutional layer of AGC module to capture global group consistency. This module endows our FCN-based AGCNet content-adaptive ability for any quantities of input images, and ensures the generalization ability of AGCNet on unseen categories without additional annotated category labels.

III. Proposed Approach

In this section, we first sketch the architecture of the proposed AGCNet in Sec. III-A. Subsequently, we describe the most important AGC module in Sec. III-B, and present the Aggregated Decoder in Sec. III-C. Finally, we state the implementation details in Sec. III-E.

A. AGCNet Overview

As reported in [30], some image SOD models [49], [50], [53] have achieved comparable performance with state-of-the-art co-saliency detection models. In our AGCNet, we use the output of SOD model as a form of prior information in co-saliency detection. In our AGCNet, we take saliency maps generated from CPD [53], which is a specialized model for single image SOD, as intra-saliency priors, *i.e.*, $\mathbf{SSM} \in \mathbb{R}^{N\times 1\times 224\times 224}$, to suppress the noise of non-salient regions. In Fig. 2, we illustrate the architecture of the proposed AGCNet, which mainly involves three parts: the feature extraction (*i.e.*, five convolution blocks Conv_1∼Conv_5), the group consistency capture (*i.e.*, three AGC modules), and the hybrid decoder (*i.e.*, one Aggregated Decoder and two normal decoders).

Given an image group with N images, denoted as $\mathbf{I} \in \mathbb{R}^{N\times 3\times 224\times 224}$, we take VGG-16 [67] pre-trained on ImageNet [68] as backbone for feature extraction, generating a set of basic features with five scales for each image of this image group, *i.e.*, $\mathbf{F}_i \in \mathbb{R}^{N\times C_i\times \frac{224}{2^i}\times \frac{224}{2^i}}$, $i \in \{1,2,3,4,5\}$. Notably, the last pooling and fully-connected layers of the original VGG-16 are discarded. Then, the basic features of three relatively high-level convolution blocks (*i.e.*, $\mathbf{F}_i$, $i \in \{3,4,5\}$) are respectively paired with **SSM** to flow into the AGC module to mine discriminative group-wise consistency features $\mathbf{CF}_i \in \mathbb{R}^{N\times 128\times \frac{224}{2^i}\times \frac{224}{2^i}}$, $i \in \{3,4,5\}$. In order to make accurate prediction for co-salient objects with various scales, the three-scale group consistency features then flow into an Aggregated Decoder to generate preliminary co-saliency maps, denoted as $\mathbf{CS}_{\text{pre}} \in \mathbb{R}^{N\times 1\times 56\times 56}$, and group consistency, denoted as $\mathbf{CF}_2 \in \mathbb{R}^{N\times 128\times 56\times 56}$. Finally, $\mathbf{CS}_{\text{pre}}$ is progressively broadcasted to shallow features through two normal decoders [69] D_i, $i \in \{1,2\}$, which combine $\mathbf{CS}_{\text{pre}}$, **SSM**, the shallow features $\mathbf{F}_i$, $i \in \{1,2\}$ and $\mathbf{CF}_i$, $i \in \{1,2\}$ to generate final co-saliency maps of the given group, $\mathbf{CS}_{\text{final}} \in \mathbb{R}^{N\times 1\times 224\times 224}$.

B. AGC Module

In Fig. 3, we take a group of three images as an example to illustrate how AGC module works. Clearly, our AGC module mainly consists of CGC block and RF block.

CGC Block: In real-life applications, co-salient objects in each image of a given group often vary in terms of texture, color, scale, and background. But they have the same semantic category attributes [31]. Thus, Wang *et al.* [65] mined the co-category vectors as group consistency. However, co-saliency detection needs to classify each pixel to segment out co-salient objects, not only to classify the image. Guaranteeing the detected co-salient objects are salient is one of the basic concepts of co-saliency detection. Based on these considerations, we introduce intra-saliency priors to extract intra-semantic information to match each pixel of input images for group consistency capture.

Firstly, we adopt the intra-saliency priors **SSM** generated from CPD model [53] to directly mask $\mathbf{F}_i \in \mathbb{R}^{N\times C_i\times H\times W}$ to filter the distractors of background and generate intra-saliency features. Once we obtain the masked features, we can directly apply global pooling operation to these features to generate intra-semantic vectors, as suggested in [29], [62]. However, since most images contain more than one salient object, many vectors represent hybrid semantic information of multiple salient objects. Thus, we try to generate relatively purer semantic vectors by adopting patch-wise average pooling with the patch size of 7×7 for the masked features and

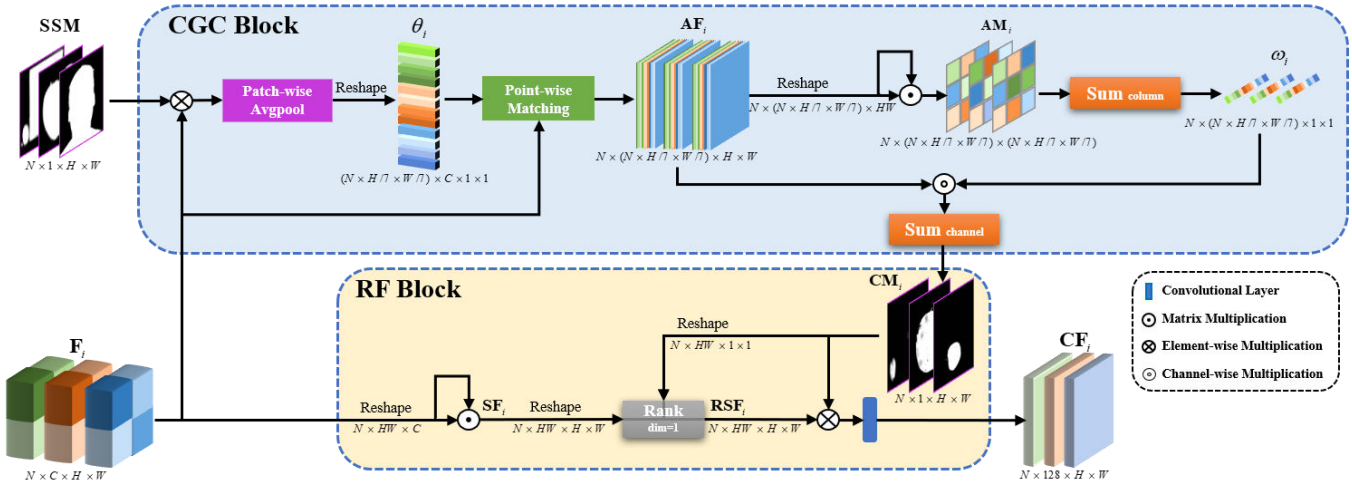

Fig. 3. **Illustration of the AGC module.**

reshaping the pooled features into a set of intra-semantic vectors, denoted by $\theta_i \in \mathbb{R}^{(N\times H/7\times W/7)\times C_i\times 1\times 1}$:

$$\theta_i = R(PAvg(\mathbf{F}_i \otimes \mathbf{SSM})), \tag{1}$$

where $PAvg(\cdot)$ and $R(\cdot)$ are the patch-wise average pooling and the reshape operation, respectively, and $\otimes$ is the element-wise multiplication. Notably, according to the size of input image, the patch size is fixed to 7×7 for any scale of masked features in AGCNet. In this way, intra-semantic vectors generated from multi-scale features with different receptive fields can cover salient objects with various scales.

Subsequently, we try to make semantic matching between the intra-semantic vectors and the pixels of detected image features $\mathbf{F}_i$. We achieve this matching through a dynamic point-wise convolutional layer [32] with θ_i as kernels. From the perspective of feature matching, the semantic vectors of every patch in the group θ_i take turns matching $\mathbf{F}_i$ at pixel level through the dynamic convolution, generating a set of pixel-level matching maps for each individual image, denoted as $\mathbf{af}_i^{(n)} \in \mathbb{R}^{(N\times H/7\times W/7)\times H\times W}$, and extending to a group of images, denoted as $\mathbf{AF}_i \in \mathbb{R}^{N\times(N\times H/7\times W/7)\times H\times W}$. Each matching map is associated with an intra-semantic vector, and the pixels with high correlation to the intra-semantic vector will be highlighted in the corresponding matching map. In this way, we formulate the matching maps for a group of images as follows:

$$\mathbf{AF}_i = CPconv(\mathbf{F}_i|\theta_i), \tag{2}$$

where $CPconv(\cdot)$ denotes the point-wise convolutional layer which is content-adaptive, and the channel size of $\mathbf{AF}_i$ is determined by the size of the group of features.

Then, we summarize these matching maps $\mathbf{af}_i^{(n)}$ in channel-wise to directly reflect the co-saliency score of each pixel. While considering the dependence of patch-wise intra-semantic vectors, we construct an affinity matrix to weight the matching maps before the summation. For each image, we resize the size of $\mathbf{af}_i^{(n)}$, *i.e.*, $\mathbb{R}^{(N\times H/7\times W/7)\times H\times W}$, into $\mathbb{R}^{(N\times H/7\times W/7)\times HW}$, and measure the relevance between each two matching maps via matrix multiplication to generate affinity matrix $\mathbf{am}_i^{(n)} \in \mathbb{R}^{(N\times H/7\times W/7)\times(N\times H/7\times W/7)}$.

$$\mathbf{am}_i^{(n)} = \mathbf{af}_i^{(n)} \odot \mathbf{af}_i^{(n)^T}, \tag{3}$$

where $\odot$ denotes matrix multiplication, and this operation is applied on $\mathbf{AF}_i$ to generate $\mathbf{AM}_i$ for the group. The importance of each matching map can be weighted by the summation of all elements in the column of $\mathbf{am}_i^{(n)}$. Concretely, the weight vector $\omega_i^{(n)} \in \mathbb{R}^{(N\times H/7\times W/7)\times 1\times 1}$ for $\mathbf{af}_i^{(n)}$ can be formulated as follows:

$$\omega_i^{(n)} = softmax\sum_{j=1}^{N\times H/7\times W/7}(\mathbf{am}_i^{(n)})_{\text{column}}. \tag{4}$$

With the generated weight vector, we summarize $\mathbf{af}_i$ into a group consistency map $\mathbf{cm}_i^{(n)} \in \mathbb{R}^{1\times H\times W}$ at channel-wise, formulated as follows:

$$\mathbf{cm}_i^{(n)} = \sum_{j=1}^{N\times H/7\times W/7}(\omega_i^{(n)} \circ \mathbf{af}_i^{(n)})_{\text{channel}}, \tag{5}$$

where $\circ$ is the channel-wise multiplication. In this way, a group of consistency maps $\mathbf{CM}_i$ can be generated.

Our CGC block benefits from the content-adaptive property of the dynamic convolution [32], [70] and the weighting summation, breaking the static-manner based inference and gaining content-adaptive ability for capturing group-consistency under various quantity and category of input groups.

RF Block: The consistency map reflects the potential co-saliency score for each pixel without the consideration of the dependence between pixels of each individual image. If we multiply $\mathbf{CM}_i$ with the $\mathbf{F}_i$ directly, the pixels belong to the same co-salient objects but with relatively individual attributes may fail to be distinguished due to its low matching, leading to sub-optimal prediction. Therefore, to detect more complete objects, we employ a RF block to combine the self-correlation of individual image with group consistency.

First, we mine the self-correlation relationships $\mathbf{sf}_i^{(n)}$ for each image feature $\mathbf{f}_i^{(n)} \in \mathbb{R}^{C_i\times H\times W}$ by computing inner pixel-wise correlations of individual image. Specifically, we reshape the size of $\mathbf{f}_i^{(n)}$ to $\mathbb{R}^{HW\times C_i}$, and use the inner product to construct an affinity matrix, which is formulated as follows:

$$\mathbf{sf}_i^{(n)} = \mathbf{f}_i^{(n)} \odot \mathbf{f}_i^{(n)^T}, \tag{6}$$

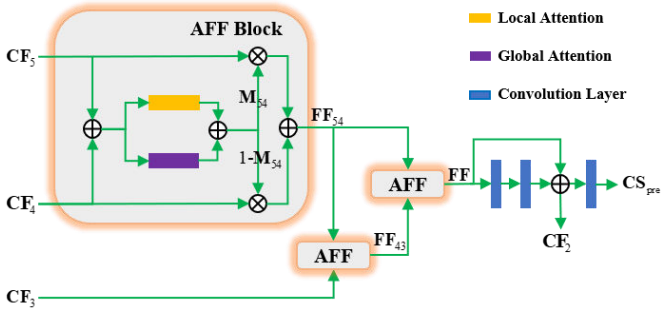

Fig. 4. **Illustration of the Aggregated Decoder.**

where the size of the affinity matrix $\mathbf{sf}_i^{(n)} \in \mathbb{R}^{HW\times HW}$, and the affinity matrixes of a group of image features $\mathbf{SF}_i \in \mathbb{R}^{N\times HW\times HW}$ can be constructed like this.

Then for fine-grained performance, we combine the group consistency with self-correlation. We reshape the size of $\mathbf{SF}_i$ into $\mathbb{R}^{N\times HW\times H\times W}$ as the correlation maps and rearrange the channel order of reshaped $\mathbf{SF}_i$ as suggested in [62], [69]. The number of pixels in $\mathbf{CM}_i$ is consistent with the channel size of $\mathbf{SF}_i$. For the pixel that has higher co-saliency scores in $\mathbf{CM}_i$, the channel of self-correlation map $\mathbf{SF}_i$ with the same index will be placed on the upper channel to generate the $\mathbf{RSF}_i \in \mathbb{R}^{N\times HW\times H\times W}$. After that, we multiply $\mathbf{RSF}_i$ and $\mathbf{CM}_i$, and take a regular convolutional layer to compress the channel size. The group consistency features $\mathbf{CF}_i \in \mathbb{R}^{N\times 128\times H\times W}$ is thus obtained. In this way, the pixel with a high co-saliency score will drive the pixels which have strong dependence on it to be highlighted.

In the RF block, we keep the internal coherence of each co-salient object and integrate the group consistency maps with individual attributes for more discriminative group consistency features. Notably, even the parameters of the adopted regular convolutional layer in the RF block are fixed after training, our AGC module still maintains the content-adaptive ability owning to the adopted rearrangement operation in RF block.

C. Aggregated Decoder

Based on the features of different scales, the captured common attributes are different. Thus, to handle diverse scenes, the combination of multi-scale group consistency is an indispensable process in reasoning. While feature fusion in the decoder of co-saliency models is usually implemented via the feature concatenation, which allocates the features with fixed weights regardless of the importance of different scales. From these considerations, we specifically design an Aggregated Decoder to effectively aggregate the three-scale group consistency features for reasoning. This decoder is mainly composed of three Attentional Feature Fusion (AFF) blocks and a residual connection, as shown in Fig. 4. Each AFF block corresponds to a feature fusion between two adjacent scales. We take the fusion of $\mathbf{CF}_4$ and $\mathbf{CF}_5$ as an example to describe the details of the AFF block.

In AFF block, we construct a soft selection between two features for fusion. Concretely, the soft selection mainly depends on the attention mechanism [76] for local channel context aggregation and global channel context aggregation, named as local attention (LA) and global attention (GA). The LA exploits point-wise channel interactions for each spatial position, while the GA aggregates channel context. For $\mathbf{CF}_4$ and $\mathbf{CF}_5$, we first take up-sampling operation for $\mathbf{CF}_5$ to keep the size of it to be consistent with $\mathbf{CF}_4$. Then, we formulate LA and GA as follows:

$$\mathbf{LA}_{54} = Pconv_{\beta,2}(\delta(Pconv_{\beta,1}(\mathbf{CF}_5 \oplus \mathbf{CF}_4))), \tag{7}$$

$$\mathbf{GA}_{54} = FC_{\beta,2}(\delta(FC_{\beta,1}(GAP(\mathbf{CF}_5 \oplus \mathbf{CF}_4)))), \tag{8}$$

where $\oplus$ is the element-wise summation, $Pconv_\beta$ is the point-wise convolutional layer with Batch Normalization (BN), δ denotes the Rectified Linear Unit (ReLU), FC_β is the fully connected layer with BN, GAP is the global average pooling, $\mathbf{LA}_{54}$ and $\mathbf{GA}_{54}$ respectively represent the local channel context aggregation and global channel context aggregation for $\mathbf{CF}_5$ and $\mathbf{CF}_4$.

We combine local channel context aggregation with global channel context aggregation to generate the soft selection matrix $\mathbf{M}_{54}$, formulated as follows:

$$\mathbf{M}_{54} = \zeta(\mathbf{LA}_{54} \oplus \mathbf{GA}_{54}), \tag{9}$$

where the size of $\mathbf{M}_{54}$ is consistent with the size of $\mathbf{CF}_4$, ζ denotes the Sigmoid, and elements in $\mathbf{M}_{54}$ belong to [0,1]. With $\mathbf{M}_{54}$, we fuse $\mathbf{CF}_4$ and $\mathbf{CF}_5$ to generate the fused features $\mathbf{FF}_{54}$ as follows:

$$\mathbf{FF}_{54} = \mathbf{CF}_5 \otimes \mathbf{M}_{54} \oplus \mathbf{CF}_4 \otimes (1 - \mathbf{M}_{54}). \tag{10}$$

As shown in Fig. 4, to keep more high-level semantic category attributes, we adopt an AFF block to fuse $\mathbf{FF}_{54}$ and $\mathbf{CF}_3$ and generate fused features $\mathbf{FF}_{43}$. And then fused features, *i.e.*, $\mathbf{FF}_{54}$ and $\mathbf{FF}_{43}$ flow into another AFF block to generate $\mathbf{FF}$. So thus, the three AFF blocks realize the aggregation of three-scale group consistency. Finally, we employ a residual connection operation on $\mathbf{FF}$ to generate $\mathbf{CS}_{\text{pre}}$ and $\mathbf{CF}_2$.

D. Normal Decoder

After the Aggregated Decoder locating co-salient objects for each input image in the form of low-resolution co-saliency map, we try to combine shallow-level features and the aggregated group consistency to improve the integrity of co-salient objects in images with full resolution. To this end, we adopt two normal decoders for further refinement. In light of the superiority of existing image SOD models in object edge detection, we also introduce the intra-saliency priors into two decoders to progressively generate sharper and more homogeneous co-saliency maps with full resolution.

In the normal decoder, $\mathbf{CS}_{\text{pre}}$ and $\mathbf{SSM}$ are utilized as masks for $\mathbf{F}_i$, $i \in \{1,2\}$, to generate two types of masked features. After up-sampling, the masked features and the semantic group consistency features $\mathbf{CF}_i$, $i \in \{1,2\}$ are concatenated for more accurate inference.

TABLE I

BENCHMARKING RESULTS OF 16 LEADING CO-SALIENCY DETECTION MODELS AND THREE IMAGE SOD MODELS ON FOUR DATASETS [27], [30], [71], [72]. '-' MEANS THAT THE AUTHORS DO NOT RELEASE RESULTS OR CODES, 'CO' AND 'SIN' ARE IN THE 'TYPE' COLUMN REPRESENT THE CORRESPONDING MODELS ARE CO-SALIENCY DETECTION MODELS AND SOD ONES, RESPECTIVELY. IML [21] ADOPTS COSAL2015 AS TRAINING DATA, THUS AUTHORS DO NOT TEST ON THIS DATASET. MSRC FOR GCAGC [19] IS IN THE SAME SITUATION. ↑ & ↓ DENOTE LARGER AND SMALLER IS BETTER, RESPECTIVELY. THE TOP THREE RESULTS ARE MARKED IN **RED**, **BLUE** AND **GREEN**, RESPECTIVELY.

Models	Type	# Param (M)	FLOPs (G)	Speed (FPS)	Cosal2015				iCoseg				MSRC				CoSOD3k			
					$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$
CBCS [10]	Co	-	-	3.3	.545	.568	.568	.234	.671	.763	.696	.166	.498	.671	.509	.297	.450	.496	.413	.204
CSHS [14]	Co	-	-	0.01	.595	.565	.586	.312	.747	.686	.739	.177	.671	.735	.656	.280	.568	.494	.566	.308
ESMG [12]	Co	-	-	0.8	.552	.511	.593	.248	.744	.680	.766	.149	.545	.610	.628	.291	.534	.451	.588	.239
CODR [11]	Co	-	-	0.03	.693	.608	.684	.203	.822	.744	.826	.107	.756	.771	.795	.192	.643	.526	.639	.222
DIM [73]	Co	-	-	0.04	.595	.525	.564	.312	.760	.679	.733	.174	.662	.681	.614	.302	.562	.456	.537	.327
CoDW [27]	Co	-	-	-	.650	.573	.608	.274	.751	.656	.709	.178	.714	.738	.675	.259	-	-	-	-
SPMIL [58]	Co	-	-	-	-	-	-	-	.782	.675	.745	.159	.769	.768	.742	.215	-	-	-	-
UCSG [74]	Co	-	-	-	.754	.690	.741	.159	.822	.779	.813	.118	**.795**	.819	.790	.175	-	-	-	-
CSMG [17]	Co	-	-	0.31	.776	.757	.783	.131	.812	.790	.820	.105	.728	.851	.749	.182	.712	.684	.707	.141
IML [21]	Co	963.8	1338.1	4.8	-	-	-	-	.833	.796	.843	.101	.786	.834	.799	.167	.736	.642	.742	.155
FEM [22]	Co	136.4	426.9	16.7	-	-	-	-	.844	.804	.846	.099	**.801**	.842	.817	.152	.680	.559	.681	.187
MGLCN [75]	Co	-	-	-	.805	.712	.800	.130	.861	**.868**	.868	.077	.788	.814	.784	.182	-	-	-	-
GCAGC [19]	Co	-	-	-	.810	.819	.836	.095	.859	.801	.874	.079	-	-	-	-	-	-	-	-
CoEG-Net [28]	Co	-	-	0.43	.838	**.857**	.872	.078	**.869**	.841	.898	.060	.712	.813	.756	.178	.778	**.794**	.820	.084
GICD [29]	Co	278.0	46.8	50.1	.839	.844	.879	.073	.821	.827	.881	.070	.665	.785	.733	.198	**.798**	.772	**.846**	.087
ICNet [62]	Co	**18.4**	**24.1**	**201.3**	**.857**	.854	**.897**	**.058**	.863	.855	**.917**	**.049**	.739	.816	.815	.155	.797	.787	.844	**.079**
CPD [53]	Sin	29.2	59.5	68.0	.825	.801	.845	.094	.857	.828	.891	**.057**	.733	.853	.780	.158	.779	.738	.805	.107
EGNet [49]	Sin	108.1	270.8	12.7	.824	.789	.840	.096	**.870**	.841	.897	.060	.730	.792	.776	.164	.784	.784	.808	.106
BASNet [50]	Sin	87.1	127.3	36.2	.820	.806	.842	.097	**.867**	.849	**.900**	**.057**	.774	**.857**	**.829**	**.131**	.773	.743	.805	.110
Ours-pre	Co	**16.8**	**19.2**	**172.0**	**.857**	**.865**	**.891**	**.063**	.851	**.864**	.898	.058	.786	**.861**	**.836**	**.123**	**.821**	**.810**	**.865**	**.073**
Ours	Co	**17.0**	**26.2**	**151.1**	**.868**	**.879**	**.903**	**.055**	.862	**.883**	**.912**	**.049**	**.791**	**.874**	**.850**	**.113**	**.829**	**.825**	**.879**	**.066**

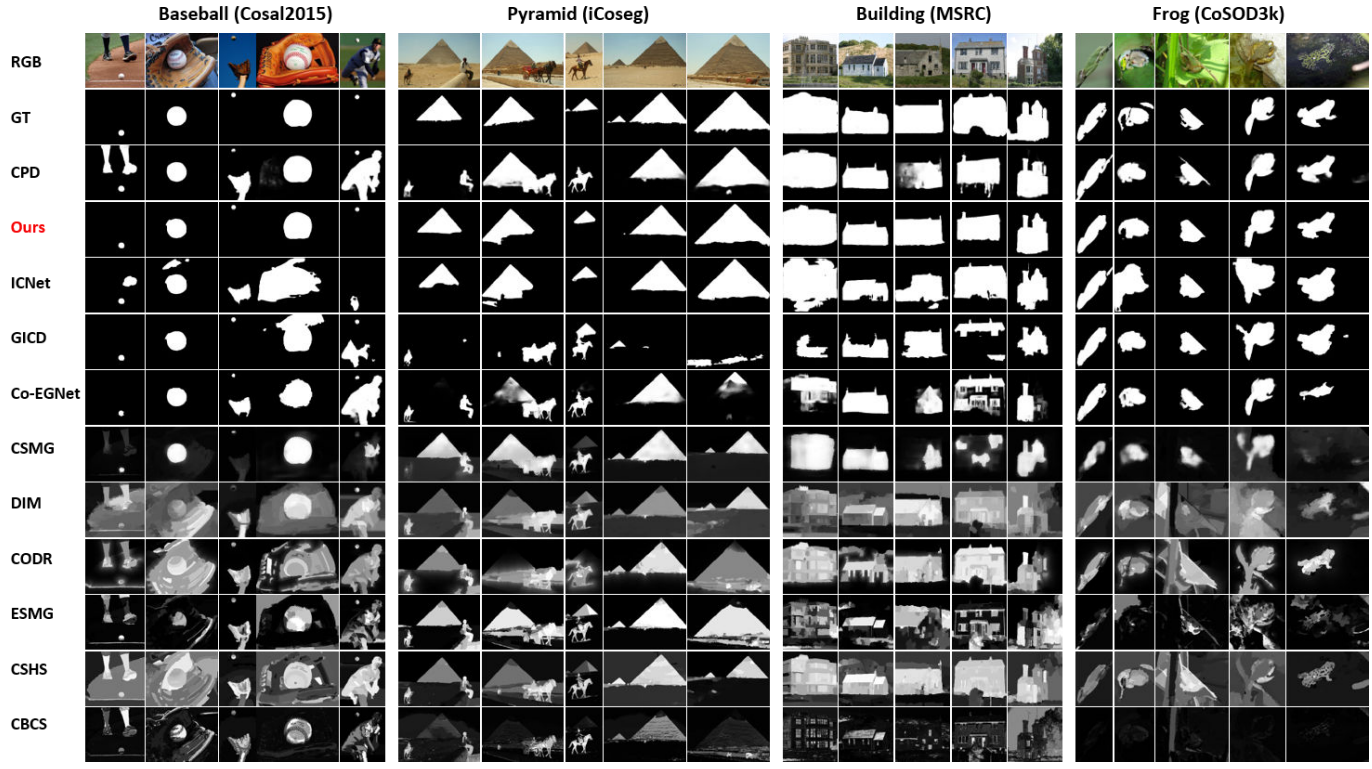

Fig. 5. Visual comparisons with five learning based co-saliency detection models (ICNet [62], GICD [29], Co-EGNet [28], CSMG [17], DIM [73]), and four classical non-learning based co-saliency detection models (CODR [11], ESMG [12], CSHS [14], CBCS [10]) on four datasets. Besides these, a image SOD model CPD [53] is also involved in the visual comparison, due to that the introduced SOD model of our posted results is CPD.

E. Implementation Details

1) Supervisions. We supervise N preliminary co-saliency maps and final co-saliency maps with corresponding ground-truths via the widely-used IoU loss [77], [78] as follows:

$$L = 1 - \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{(h,w)}(min(\mathbf{cs}^{(n)}, \mathbf{gt}^{(n)}))_{(h,w)}}{\sum_{(h,w)}(max(\mathbf{cs}^{(n)}, \mathbf{gt}^{(n)}))_{(h,w)}}, \tag{11}$$

where (h,w) denotes the position of pixel, $\mathbf{cs}^{(n)}$ is the concatenation of $\mathbf{cs}^{(n)}_{\text{pre}}$ and $\mathbf{cs}^{(n)}_{\text{final}}$ at channel, $max(\cdot,\cdot)$ and $min(\cdot,\cdot)$ are utilized to retain the element-wise maximum and minimum of two inputs, respectively.

2) Network Training Protocol. The experiments are implemented on Pytorch [79] by adapting a NVIDIA GTX 2080TI GPU (11G memory). Except for the parameters of the backbone, the additional parameters in the proposed AGCNet are initialized with the random normal distribution of which $\mu = 0, \sigma = 0:1$. We use Adam [80] as the optimizer to train our AGCNet with 60 epochs, and respectively set the learning rate and weight decay to 10^{-5} and 10^{-4}. The training data is a subset of the COCO dataset [81], including 65 groups of 9,213 images, as suggested by [28], [62]. All imported images are resized into 224×224. For each training iteration of training stage, we set the upper limited number of a batch to 11 due to the limited GPU memory, the images in each batch are all randomly selected from a same image group. In the testing stage, each image group with an arbitrary quantity of images constitutes a batch.

IV. Experiment

A. Dataset and Evaluation Metrics

1) Dataset. We conduct our experiments on four datasets (see Table I). The datasets, include iCoseg [71], MSRC [72], Cosal2015 [27] and CoSOD3k [30], are used for testing.

iCoseg [71] is originally proposed for co-segmentation task. After modification, it becomes the most widely used dataset in co-saliency detection task. The dataset contains 25 scenes, covering sports, animals, landmarks and so on. This dataset totally includes 643 images, which are divided into 38 groups. For each group, the co-salient objects and backgrounds of each image are roughly the same.

MSRC [72] is originally proposed for object classification task, which is used for co-saliency detection lately. This dataset consists of 7 groups of 233 images, and each group has 30-53 images. Most images have only one single category of salient objects and the synergy of co-salient objects of group tends to be semantic category consistency.

Cosal2015 [27] includes 50 groups of 2,015 images, and each group has 25-52 images. It is a relatively challenging dataset due to the diverse variances in appearances and complex backgrounds, and most images have more than one salient object.

CoSOD3k [30] is the largest co-saliency detection dataset, which is recently proposed with more realistic settings. Totally, it contains 13 super-classes, 160 groups and 3,316 images, where each super-class is carefully selected to cover diverse scenes. Thus, it is the most challenging dataset among the test datasets in this paper.

2) Evaluation Metrics. We use S-measure [82] ($S_\alpha, \alpha = 0.5$), maximum F-measure [83] ($F_\beta, \beta^2 = 0.3$), maximum E-measure [84] (E_ξ), and Mean Absolute Error [85] (MAE, M) to evaluate the performance of our proposed model and all compared models. The adopted evaluated tools are provided by Fan *et al.* [30].

S-measure is proposed for structure information evaluation, motivated by the studies of human behavioral vision. The S-measure combines region-aware and object-aware structural similarity as their final structure metric.

F-measure is essentially a region based similarity metric, which is adopted extensively in the field of saliency detection [50], [86], [87]. Following [62], we provide the maximum F-measure using varying fixed (0-255) thresholds.

E-measure is an enhanced alignment measure [84], which is specifically proposed for the evaluation of binary map. This measure is based on cognitive vision studies to combine local pixel values with the image-level mean value in one term, jointly capturing image-level statistics and local pixel matching information.

MAE is used to evaluate the pixel-level error between a predicted co-saliency map and the corresponding GT [85].

B. Comparisons with State-of-the-art models

To evaluate the effectiveness of the proposed model, 19 state-of-the-art models are adopted for comparison, including CBCS [10], CSHS [14], ESMG [12], CODR [11], DIM [73], CoDW [27], SPMIL [58], UCSG [74], CSMG [17], IML [21], FEM [22] MGLCN [75], GCAGC [19], CoEG-Net [28], GICD [29], ICNet [62], CPD [53], EGNet [49], and BASNet [50]. Among these, CBCS, CSHS, ESMG and CODR are four conventional co-saliency detection models which are based on handcrafted features, DIM, CoDW, SPMIL, CSMG and UCSG are five co-saliency detection models which are based on deep learning features, IML, FEM, MGLCN, GCAGC, GICD and ICNet are end-to-end deep learning-based models for co-saliency detection, CoEG-Net is a a two-stage model, and CPD, EGNet and BASNet are end-to-end image SOD models. Notably, in our comparison, the backbone of the compared image SOD models are VGG-16 [67]. For fair comparison, we use either the implementations with recommended parameter settings or co-saliency maps provided by the authors. These resources have been collected by Fan *et al.* [30][1]. Among the models with the released source code, Parameters, FLOPs and Speed of the end-to-end CNN-based models, which are IML, FEM, GICD, ICNet, CPD, EGNet and BASNet, are provided. For the traditional models based on handcrafted features and the models without CNN-based reasoning, we only compare the Speed.

Quantitative Comparisons. As shown in Table I, our model outperforms the compared models in terms of most metrics on four datasets. For example, F-measure and MAE scores of our model consistently outperform all compared models. For the most challenging dataset CoSOD3k, our model improves upon the second best model (except the preliminary results of our AGCNet, *i.e.*, Ours-pre) by about

[1] https://dpfan.net/CoSOD3K/

3.1%, 3.1%, 3.3% and 1.8% in terms of S-measure, F-measure, E-measure and MAE respectively. Since the images in the iCoseg and MSRC datasets typically contain one co-salient object, therefore the image SOD model CPD can easily handle these datasets. However, due to intrinsic limitation, CPD fails to handle the images with multiple objects in the Cosal2015 and CoSOD3k datasets. The co-salient objects in the training dataset tend to be the category consistent and the co-salient objects in the iCoseg dataset are further constrained by color. Due to the gap between the training datasets and the iCoseg dataset and the superiority of SOD model for simple dataset, AGCNet performs relatively weak on the iCoseg dataset compared with other models. From Fig. 3, the AGCNet embedded with three AGC modules seems to have high computational complexity, but that is not the case. Compared with the existing end-to-end co-saliency detection models, the proposed AGCNet requires the least parameters to be trained (even compared with the SOD which is a relatively easy task). The FLOPs and Speed of the proposed AGCNet slightly lag behind ICNet, but the performance of ICNet is lower than our model. And similar to our model, ICNet also needs additional assistance from existing image SOD models. Our model effectively captures global group consistency on multi-level features with the support of intra-saliency priors, and therefore exhibits competitive performance as compared with 16 co-saliency detection models and three image SOD models. In addition, Ours-pre is also competitive without two normal decoders. Although the performance is attenuated, the efficiency is reinforced. From the trade-off between performance and efficiency, the two normal decoders have a positive effect on co-saliency detection.

Visual Comparisons. We show co-saliency maps generated on various challenging scenes to demonstrate the superiority of AGCNet visually in Fig. 5. It can be observed that traditional models CBCS and CSHS can hardly locate common salient regions with handcrafted features. From the results of most cases, the image SOD model CPD can better find salient objects with sharp boundaries, but the non-common objects cannot be erased. For the co-salient objects with small size shown in the baseball group of Cosal2015 dataset, our model can successfully suppress the large non-common salient objects and perform significantly better than the compared models. The CGC block in our AGC module is particularly effective in handing extreme scale variation of co-salient objects. For the co-salient objects with low contrast like the pyramid group of iCoseg dataset, although the co-salient objects can not be detected by auxiliary CPD, our model can highlight the pixels of co-salient objects by connecting high correlation with other images of the same group, verifying the robust group consistency modeling capability. In contrast, GICD [29] relies on the group consensus and can not discriminatively put more weight to the co-salient objects. Even for co-salient objects with large size, *i.e.*, the building group of MSRC dataset, our model can propagate high-level semantic group consistency cues to shallow level to highlight salient objects more evenly without holes. In the case of the frog group of CoSOD3k dataset with background clutter and cross images variations, more complete object contours can be detected by our model with the help of the introduced **SSM** generated by CPD.

TABLE II
PERFORMANCE OF AGCNET WITH DIFFERENT QUANTITIES OF INPUTS.

number	Cosal2015			CoSOD3k		
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
3	0.858	0.872	0.064	0.820	0.814	0.073
5	0.863	0.875	0.059	0.826	0.819	0.069
10	0.866	0.878	0.057	0.828	0.823	0.067
Ours	**0.868**	**0.879**	**0.055**	**0.829**	**0.825**	**0.066**

TABLE III
PERFORMANCE OF AGCNET WITH INTRA-SALIENCY PRIORS OBTAINED FROM VARIOUS IMAGE SOD MODELS.

Models	Cosal2015			CoSOD3k		
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
CPD [53]	0.825	0.801	0.094	0.779	0.738	0.107
Our-CPD	0.868	**0.879**	0.055	**0.829**	0.825	**0.066**
EGNet [49]	0.824	0.789	0.096	0.784	0.784	0.106
Our-EGNet	0.870	0.877	0.055	0.827	**0.828**	0.068
BASNet [50]	0.820	0.806	0.097	0.773	0.743	0.110
Our-BASNet	**0.869**	0.876	**0.054**	0.826	0.820	0.071
w/o SSM	0.849	0.857	0.078	0.803	0.793	0.093

C. Ablation Studies

To gain insight of our key components, we do extensive ablation experiments to investigate the effectiveness of them, including the performance of AGCNet with different quantities of input images, the dependence of AGCNet on auxiliary intra-saliency priors, the design rationality of AGC module, the impacts of the AGC module number to AGCNet and the effectiveness of Aggregated Decoder. Compared with the MSRC and iCoseg datasets, the collected images of Cosal2015 and CoSOD3k dataset cover more diverse scenarios, which are more in line with realistic scenes. To this end, we perform the ablation studies on Cosal2015 and CoSOD3k datasets.

Performance of AGCNet with different quantities of input images. As presented in Table II, there is a performance gap between the numbers of input images, *i.e.*, 3 and 5. However, due to the size limitation of most models, existing end-to-end models are difficult to deal with a complete group with more than 5 images, which result in incorrectly reserving some attributes that are not shared in the image group, as shown in Fig. 1. AGCNet benefits from AGC module without such a limitation, as the number of input images increases to 10, the performance is close to that of the complete image group as inputs.

The Dependence of AGCNet on Auxiliary Intra-saliency priors. We exploit image SOD results as intra-saliency priors in our AGCNet. In order to find out the dependence of our model on image SOD model, we apply three end-to-end image SOD models, *i.e.*, CPD, EGNet and BASNet, on our model, and remove SSM from AGCNet to construct the variant w/o SSM. In Table III, we list the quantitative results of the original image SOD models and the corresponding applications for our model, *i.e.*, Our-CPD, Our-EGNet, Our-BASNet, on the

TABLE IV
PERFORMANCE OF DIFFERENT VARIANTS TO OUR AGC MODULE. SSM OF THE TABLE CORRESPONDS TO THE OPERATION THAT INTRODUCING SSM TO AGC MODULE, PA INDICATES THE PATCH-WISE AVERAGE POLLING IN THE PROCESS OF GENERATING INTRA-SEMANTIC VECTORS IN CGC BLOCK, CP IS THE CONTENT-ADAPTIVE CONVOLUTION LAYER FOR SEMANTIC MATCHING IN CGC BLOCK, WS DENOTES THE WEIGHTED SUMMATION IN CGC BLOCK, RF IS THE RF BLOCK.

Models	SSM	PA	CP	WS	RF	Cosal2015			CoSOD3k		
						$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
0	✓	✓	✓	✓	✓	**0.868**	**0.879**	**0.055**	**0.829**	**0.825**	**0.066**
1		✓	✓	✓	✓	0.854	0.855	0.063	0.811	0.793	0.076
2	✓		✓	✓	✓	0.862	0.877	0.057	0.819	0.802	0.070
3	✓	✓		✓	✓	0.847	0.832	0.079	0.801	0.799	0.085
4	✓	✓	✓		✓	0.863	0.869	0.057	0.821	0.813	0.069
5	✓	✓	✓	✓		0.853	0.854	0.063	0.820	0.808	0.071
6	✓					0.838	0.812	0.087	0.791	0.768	0.091

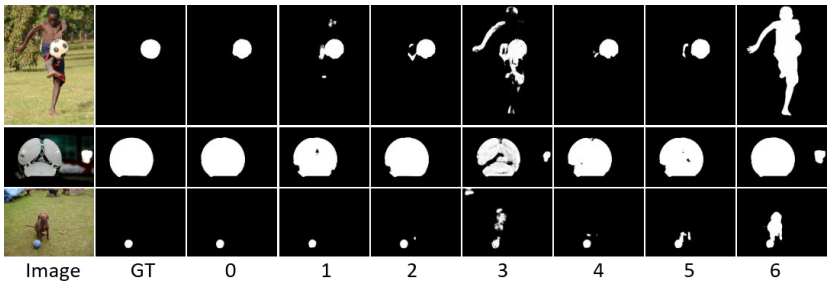

Fig. 6. Visual comparisons of AGCNet with variants about AGC module.

TABLE V
PERFORMANCE OF DIFFERENT NUMBER OF AGC MODULE.

AGC number	Cosal2015			CoSOD3k		
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
0	0.807	0.799	0.107	0.775	0.747	0.096
1	0.855	0.856	0.062	0.816	0.800	0.071
2	0.863	0.871	**0.055**	0.828	**0.830**	0.067
3	**0.868**	**0.879**	**0.055**	**0.829**	0.825	**0.066**
4	0.859	0.851	0.057	0.815	0.814	0.076
5	0.855	0.853	0.061	0.809	0.803	0.078

TABLE VI
PERFORMANCE OF DIFFERENT VARIANTS TO AGGREGATED DECODER.

Variants	Cosal2015			CoSOD3k		
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
w/o GA	0.861	0.858	0.060	0.823	0.803	0.070
w/o LA	0.863	0.857	0.058	0.821	0.806	0.071
w/o AD	0.847	0.831	0.075	0.802	0.797	0.086
w/ 2AFF	0.865	0.866	0.058	0.825	0.810	0.068
Ours	**0.868**	**0.879**	**0.055**	**0.829**	**0.825**	**0.066**

most challenging two benchmark datasets in co-saliency field. Our model clearly improves the performance on co-saliency benchmarks, no matter which image SOD model is used for our model. The performance of w/o SSM is inferior to the performance of Our-CPD, Our-EGNet and Our-BASNet, which verifies the effectiveness of using intra-saliency priors to our model. Although our model depends on image SOD to a certain extent, while judging from the improvements made by the three auxiliary image SOD models, the performance of image SOD model does not have a great impact on our model, *e.g.*, even CPD performs worse than EGNet in terms of CoSOD3k dataset, Ours-CPD outperforms Ours-EGNet a bit.

The Rationality of AGC Module. AGC module in AGCNet plays the most important role in capturing consistency information within a group. In order to verify the design rationality of the AGC module, the five important operations as described in Section III-B in the module are removed or replaced by some conventional operations in turn. The quantitative performance is reported in Table IV. For the convenience of comparison, we use the original AGCNet as the baseline with the index '0'. For model '1', we discard the intra-saliency of SSM, and directly process the extracted original features with subsequent four operations. This variant differs from w/o SSM, due to the fact that SSM still works for the two normal decoders of AGCNet. The performance degradation of model '1' proves that SSM is indispensible to AGC module. The performance of model '2' is severely damaged by exchanging to global average pooling, the visual comparison in Fig. 6 also confirms that the patch-wise semantic matching is promoting for detection without less interference. We average all intra-semantic vectors of the detected group and take it as a common semantic vector, then expand the vector to the size of original feature to add each image feature. The above operation is used to replace the CP to construct model '3'. The performance of this replacement drops dramatically and the incomplete of detected soccer in the Fig. 6 all validate the generalization ability of CP. For the model '4', we directly summarize the matching maps without the process of affinity weighting, the lacking for the relation of intra-semantic vectors results of

the performance degradation. We construct the variant '5' by employing the group consistency map generated from the CGC block to mask the original feature. Without the individual attributes for each image, the variant can not well handle the misleading pixels of co-salient. For the model '6', except for the SSM, we discard the rest of operations of AGC module, just concatenate intra-saliency features of all group to generate common features, and concatenate intra-saliency feature with the common feature for each image to make inferring. Under this design, the performance of this variant reaches a low record, (*e.g.*, $S_\alpha : 0.868 \rightarrow 0.838$, $0.829 \rightarrow 0.791$; $F_\beta : 0.879 \rightarrow 0.812$, $0.825 \rightarrow 0.768$; $M : 0.055 \rightarrow 0.087$, $0.066 \rightarrow 0.091$), the non-common salient regions can not be effectively suppressed, just as shown in the results of CPD and ours in Fig. 6. The comparison of visual and quantitative results of these variant models all prove that each process of AGC module is indispensable.

The Impacts of the AGC Module Number to AGCNet. As shown in Table V, we construct model '0' by deleting AGC modules from the AGCNet, which indicates that the group consistency modeling process is omitted. By comparing the results of the proposed model with 0, 1, 2, 3 (ours), 4 and 5 AGC modules, we discover that the performances of '4' and '5' are lower than the original AGCNet with 3 AGC modules. Owning that most co-salient objects do not have clear common attributes on shallow-level features, with shallow-level features adopted for modeling group consistency, *i.e.*, model '4' and '5', the computation cost increases dramatically compared with model '3', but performs worse. With 2 AGC modules, the performance is comparable to the original AGCNet. The observations all illustrate that the high-level features with more semantic information are more effective for co-saliency task.

The Effectiveness of Aggregated Decoder. To evaluate the contribution of the proposed Aggregated Decoder to AGCNet on co-saliency task, we derive four variants: w/o GA, w/o LA, w/o AD and w/ 2AFF, the w/o GA and w/o LA of which respectively refer to that removing the global attention and local attention aggregating in turn. In light of that, there are three scales of group consistency that need to be aggregated. We delete one AFF which targets the fusion of $\mathbf{FF}_{54}$ and $\mathbf{FF}_{43}$, directly import $\mathbf{FF}_{43}$ to the next residual block to form w/ 2AFF. The variant w/o AD is constructed by replacing the Aggregated Decoder with the operation of feature concatenation. As presented in Table VI, the slight deterioration of performance of w/ 2AFF indicates that the task is more dependent on $\mathbf{CF}_5$ and $\mathbf{CF}_4$ with relatively more semantic category attributes than $\mathbf{CF}_3$. Embedding three AFF blocks can promote the reservation of $\mathbf{CF}_5$ and $\mathbf{CF}_4$, which is beneficial for co-saliency detection. The performance degradation of w/o GA, w/o LA and w/o AD confirm that the discriminate attention aggregations are reasonable as described in Section III-C and our aggregated decoder is necessary for our AGCNet.

With the support of existing image SOD models, the proposed AGC module, aggregated decoder and two normal decoders have different capabilities, are closely interdependent for a good tradeoff between effectiveness and efficiency, making AGCNet possible to be applied in practical applications. The AGC module is guided by intra-saliency priors to capture group consistency within any complete group with any quantities of images, but does not depend on the performance of intra-saliency priors. By constructing intra-semantic correlation of group and pixel-wise self-correlation of each single image, the AGC module can obtain more discriminative global group consistency to tackle the detection of easy confusing pixels compared with existing co-saliency detection models. And this module adopts less regular convolutional layers to retain the generalization for the group with unseen category. The aggregated decoder adaptively bridges the gap of adjacent-scale group consistency, to adaptively fuse three-scale group consistency for the preliminary localization of co-salient objects. And two normal decoders make further utilization of intra-saliency priors, benefiting from the advantage of edge detection of intra-saliency priors, further propagating group consistency to shallow features to improve integrity of the co-salient objects in full-resolution images.

V. Conclusion

In this paper, we propose an AGCNet for co-saliency detection. Promoted by the intra-saliency priors produced by existing image SOD models, our mainly proposed AGC module breaks the issues of the static manner-based inference and the constant quantity of input image, capturing the global group consistency within any unknown given group by semantic matching and weighted summarization. Moreover, this module integrates individual property with group consistency to extract discriminative group consistency features. The proposed Aggregated Decoder overcomes the semantic and scale inconsistency issue among multi-scale group consistency features for preliminary co-saliency detection. Experiments on four benchmark datasets demonstrated our AGCNet is competitive to 16 state-of-the-art co-saliency detection and 3 SOD models. And comprehensive ablation studies also validated the effectiveness and rationality of proposed modules.

References

[1] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient object segmentation via effective integration of saliency and objectness," *IEEE TMM*, vol. 19, no. 8, pp. 1742–1756, 2017.

[2] M. Huang, Z. Liu, L. Ye, X. Zhou, and Y. Wang, "Saliency detection via multi-level integration and multi-scale fusion neural networks," *Neurocomputing*, vol. 364, pp. 310–321, 2019.

[3] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," *ACM Symposium on User Interface Software and Technology*, pp. 219–228, 2010.

[4] C.-C. Tsai, K.-J. Hsu, L. Yen-Yu, X. Qian, and Y.-Y. Chuang, "Deep co-saliency detection via stacked autoencoder-enabled fusion and self-trained CNNs," *IEEE TMM*, vol. 22, no. 4, pp. 1016–1031, 2020.

[5] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE TMM*, vol. 18, no. 9, pp. 1896–1909, 2016.

[6] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "DeepCo3: Deep instance co-segmentation by co-peak search and co-saliency detection," in *IEEE CVPR*, 2019, pp. 8838–8847.

[7] C.-C. Tsai, W. Li, K.-J. Hsu, X. Qian, and Y.-Y. Lin, "Image co-saliency detection and co-segmentation via progressive joint optimization," *IEEE TIP*, vol. 28, no. 1, pp. 56–71, 2019.

[8] K. R. Jerripothula, J. Cai, and J. Yuan, "Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization," *IEEE TMM*, vol. 20, no. 9, pp. 2466–2477, 2018.

[9] J. Xue, L. Wang, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue and its applications to recognition and alpha matting," *Pattern Recognition*, vol. 46, no. 11, pp. 2874–2889, 2013.

[10] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE TIP*, vol. 22, no. 10, pp. 3766–3778, 2013.

[11] L. Ye, Z. Liu, J. Li, W. Zhao, and L. Shen, "Co-saliency detection via co-salient object discovery and recovery," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2073–2077, 2015.

[12] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 588–592, 2015.

[13] L. Li, Z. Liu, W. Zou, X. Zhang, and O. Le Meur, "Co-saliency detection based on region-level fusion and pixel-level refinement," in *IEEE ICME*, 2014, pp. 1–6.

[14] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 88–92, 2014.

[15] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE TMM*, vol. 14, no. 5, pp. 1520–9210, 2012.

[16] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling, "HSCS: Hierarchical sparsity based co-saliency detection for rgbd images," *IEEE TMM*, vol. 21, no. 7, pp. 1660–1671, 2018.

[17] K. Zhang, T. Li, B. Liu, and Q. Liu, "Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing," in *IEEE CVPR*, 2019, pp. 3090–3099.

[18] Q. Zhang, R. Cong, J. Hou, C. Li, and Y. Zhao, "CoADNet: Collaborative aggregation-and-distribution networks for co-salient object detection," in *NeurIPS*, 2020.

[19] K. Zhang, T. Li, S. Shen, B. Liu, J. Chen, and Q. Liu, "Adaptive graph convolutional network with attention graph clustering for co-saliency detection," in *IEEE CVPR*, 2020, pp. 9047–9056.

[20] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, F. Wu, and Y. Zhuang, "Deep group-wise fully convolutional network for co-saliency detection with graph propagation," *IEEE TIP*, vol. 28, no. 10, pp. 5052–5063, 2019.

[21] J. Ren, Z. Liu, X. Zhou, C. Bai, and G. Sun, "Co-saliency detection via integration of multi-layer convolutional features and inter-image propagation," *Neurocomputing*, vol. 371, pp. 137–146, 2020.

[22] J. Ren, Z. Liu, G. Li, X. Zhou, C. Bai, and G. Sun, "Co-saliency detection using collaborative feature extraction and high-to-low feature integration," in *IEEE ICME*, 2020, pp. 1–6.

[23] M. Li, S. Dong, K. Zhang, Z. Gao, X. Wu, H. Zhang, G. Yang, and S. Li, "Deep learning intra-image and inter-images features for co-saliency detection," in *BMVC*, 2018, pp. 1–13.

[24] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE TIP*, vol. 23, no. 9, pp. 4175–4186, 2014.

[25] C.-C. Tsai, X. Qian, and Y.-Y. Lin, "Image co-saliency detection via locally adaptive saliency map fusion," in *IEEE ICASSP*, 2017, pp. 1897–1901.

[26] H.-T. Chen, "Preattentive co-saliency detection," in *IEEE ICIP*, 2010, pp. 1117–1120.

[27] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *IEEE CVPR*, 2015, pp. 2994–3002.

[28] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE TPAMI*, 2021.

[29] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng, "Gradient-induced co-saliency detection," in *ECCV*, 2020, pp. 455–472.

[30] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a deeper look at co-salient object detection," in *IEEE CVPR*, 2020, pp. 2916–2926.

[31] Z.-J. Zha, C. Wang, D. Liu, H. Xie, and Y. Zhang, "Robust deep co-saliency detection with group semantic and pyramid attention," *IEEE TNNLS*, vol. 31, no. 7, pp. 2398–2408, 2020.

[32] B. Yang, G. Bender, Q. V Le, and J. Ngiam, "CondConv: Conditionally parameterized convolutions for efficient inference," in *NeurIPS*, 2019.

[33] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *IEEE CVPR*, 2011, pp. 409–416.

[34] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE TIP*, vol. 23, no. 5, pp. 1937–1952, 2014.

[35] X. Zhou, Z. Liu, G. Sun, L. Ye, and X. Wang, "Improving saliency detection via multiple kernel boosting and adaptive fusion," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 517–521, 2016.

[36] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *IEEE ICCV*, 2013, pp. 2976–2983.

[37] J. Li, F. Meng, and Y. Zhang, "Saliency detection using a background probability model," in *IEEE ICIP*, 2015, pp. 2189–2193.

[38] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE TPAMI*, vol. 33, no. 2, pp. 353–367, 2011.

[39] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. D. Feng, "Robust saliency detection via regularized random walks ranking," in *IEEE CVPR*, 2015, pp. 2710–2717.

[40] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE CVPR*, 2009, pp. 1597–1604.

[41] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE TCSVT*, vol. 25, no. 8, pp. 1309–1321, 2015.

[42] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE TNNLS*, vol. 29, no. 12, pp. 6038–6051, 2018.

[43] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *IEEE ICCV*, 2017, pp. 202–211.

[44] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep cnn features," *IEEE TIP*, vol. 25, no. 11, pp. 5012–5024, 2016.

[45] Y. Wu, Z. Liu, and X. Zhou, "Saliency detection using adversarial learning networks," *Journal of Visual Communication and Image Representation*, vol. 67, p. 102761, 2020.

[46] W. Wang, S. Zhao, J. Shen, S. C. H. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *IEEE CVPR*, 2019, pp. 1448–1457.

[47] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *IEEE ICCV*, 2019, pp. 7263–7272.

[48] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *IEEE CVPR*, 2019, pp. 1623–1632.

[49] 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 *IEEE ICCV*, 2019, pp. 8778–8787.

[50] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *IEEE CVPR*, 2019, pp. 7471–7481.

[51] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *IEEE CVPR*, 2020, pp. 9138–9147.

[52] H. Zhang, Y. Zeng, H. Lu, L. Zhang, J. Li, and J. Qi, "Learning to detect salient object with multi-source weak supervision," *IEEE TPAMI*, 2021.

[53] Z. Wu, L. Su, and H. Qingming, "Cascaded partial decoder for fast and accurate salient object detection," in *IEEE CVPR*, 2019, pp. 3902–3911.

[54] H. Li and N. K. Ngan, "A co-saliency model of image pairs," *IEEE TIP*, vol. 20, no. 12, pp. 3365–3375, 2011.

[55] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE TMM*, vol. 15, no. 8, pp. 1896–1909, 2013.

[56] Z. Zhang, Z. Wu, Q. Jiang, L. Du, and L. Hu, "Co-saliency detection based on superpixel matching and cellular automata," *KSII Transactions on Internet and Information Systems*, vol. 11, no. 5, pp. 2576–2589, 2017.

[57] C. Ge, K. Fu, Y. Li, J. Yang, P. Shi, and L. Bai, "Co-saliency detection via similarity-based saliency propagation," in *IEEE ICIP*, 2015, pp. 1845–1849.

[58] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE TPAMI*, vol. 39, no. 5, pp. 865–878, 2017.

[59] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE TIP*, vol. 26, no. 7, pp. 3196–3209, 2017.

[60] R. Hu, Z. Deng, and X. Zhu, "Multi-scale graph fusion for co-saliency detection," in *AAAI*, 2021.

[61] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *ACM Multimedia*, 2019, pp. 1437–1445.

[62] W.-D. Jin, J. Xu, M.-M. Cheng, Y. Zhang, and W. Guo, "ICNet: Intra-saliency correlation network for co-saliency detection," in *NeurIPS*, vol. 33, 2020, pp. 18 749–18 759.

[63] G. Gao, W. Zhao, Q. Liu, and Y. Wang, "Co-saliency detection with co-attention fully convolutional network," *IEEE TCSVT*, vol. 31, no. 3, pp. 877–889, 2021.

[64] B. Li, Z. Sun, L. Tang, and J. Shi, "Detecting robust co-saliency with recurrent coattention neural network," in *IJCAI*, 2019, pp. 818–825.

[65] C. Wang, Z.-J. Zha, D. Liu, and H. Xie, "Robust deep co-saliency detection with group semantic," in *AAAI*, 2019, pp. 8917–8924.

[66] K. Aditya and V. K. Raghavendra, "Weakly supervised multi-scale recurrent convolutional neural network for co-saliency detection and co-

segmentation," in *Neural Computing and Applications*. Springer, 2020, pp. 16 571–16 588.
[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[68] D. Jia, D. Wei, S. Richard, L. Li-Jia, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.
[69] Z. Wang, J. Xu, L. Liu, F. Zhu, and L. Shao, "RANet: Ranking attention network for fast video object segmentation," in *IEEE ICCV*, 2019, pp. 3977–3986.
[70] Y. Chen, X. Dai, M. Liu, D. Chen, Y. Lu, and Z. Liu, "Dynamic convolution: Attention over convolution kernels." in *IEEE CVPR*, 2020, pp. 11 027–11 036.
[71] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *IEEE CVPR*, 2010, pp. 3169–3176.
[72] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *IEEE ICCV*, vol. 2, 2005, pp. 1800–1807.
[73] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE TNNLS*, vol. 27, no. 6, pp. 1163–1176, 2016.
[74] K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, X. Qian, and Y.-Y. Chuang, "Unsupervised CNN-based co-saliency detection with graphical optimization," in *ECCV*, 2018, pp. 485–501.
[75] B. Jiang, X. Jiang, J. Tang, B. Luo, and S. Huang, "Multiple graph convolutional networks for co-saliency detection," in *IEEE ICME*, 2019, pp. 332–337.
[76] H. Jie, S. Li, and S. Gang, "Squeeze-and-excitation networks," in *IEEE CVPR*, 2018, pp. 7132–7141.
[77] H. Lin, X. Qi, and J. Jia, "AGSS-VOS: Attention guided single-shot video object segmentation," in *IEEE ICCV*, 2019, pp. 3948–3956.
[78] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE TIP*, vol. 30, pp. 3528–3542, 2021.
[79] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.
[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
[81] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, and D. Ramanan, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.
[82] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *IEEE ICCV*, 2017, pp. 4548–4557.
[83] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE CVPR*, 2009, pp. 1597–1604.
[84] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment Measure for Binary Foreground Map Evaluation," in *IJCAI*, 2018, pp. 698–704.
[85] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE CVPR*, 2012, pp. 733–740.
[86] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *IEEE ICCV*, vol. 2, 2001, pp. 416–423.
[87] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *IEEE CVPR*, 2016, pp. 678–686.

Zhen Bai received the B.E. degree from Wuhan Huaxia University of Technology, Wuhan, China, in 2016, the M.S. degree from the Zhengzhou University of Light Industry, China, Zhengzhou, in 2019, and is currently pursuing the Ph.D. degree with School of Communication and Information Engineering in Shanghai University, Shanghai, China. Her research interests include machine learning and saliency detection.

Zhi Liu (M'07-SM'15) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 1999, 2002 and 2005, respectively. He is currently a Professor at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He was a TPC member/session chair in ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, WIAMIS 2013, *etc*. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an Area Editor of *Signal Processing: Image Communication* and served as a Guest Editor for the special issue on *Recent Advances in Saliency Models, Applications and Evaluations* in *Signal Processing: Image Communication*.

Gongyang Li received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include image/video object segmentation and saliency detection.

Yang Wang received the B.Sc. degree from the Harbin Institute of Technology, Harbin, China, the M.Sc. degree from the University of Alberta, Edmonton, AB, Canada, and the Ph.D. degree from Simon Fraser University, Burnaby, BC, Canada, all in computer science. He was previously a NSERC Postdoc Fellow with the University of Illinois at Urbana-Champaign, Champaign, IL, USA. He is currently an Associate Professor of computer science with the University of Manitoba, Winnipeg, MB, Canada. His research interests include computer vision and machine learning.