信息科学研究所
Institute of Information Science

MePro 数字媒体信息处理研究中心
Center of Digital Media Information Processing

# CoADNet: Collaborative Aggregation-and-Distribution Networks for Co-Salient Object Detection

Runmin Cong (丛润民)

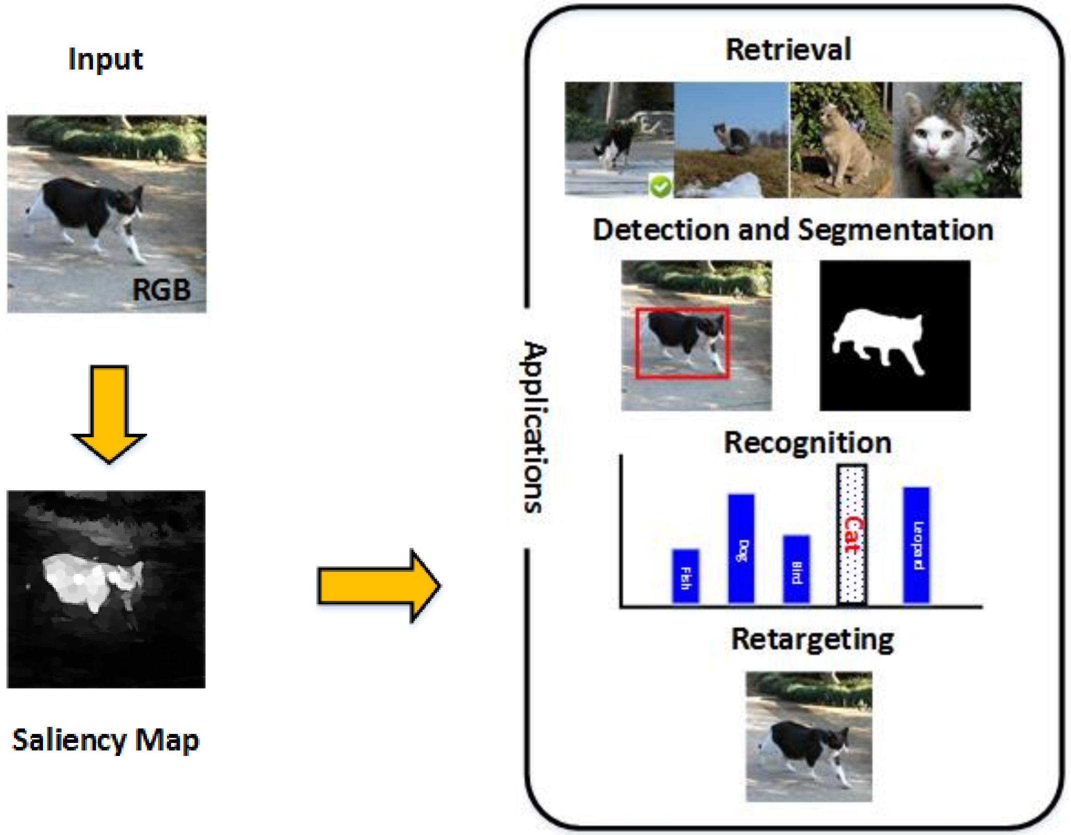
Beijing Jiaotong University

2020-12-06@NeurIPS’20 MeetUp

https://github.com/rmcong/CoADNet_NeurIPS20

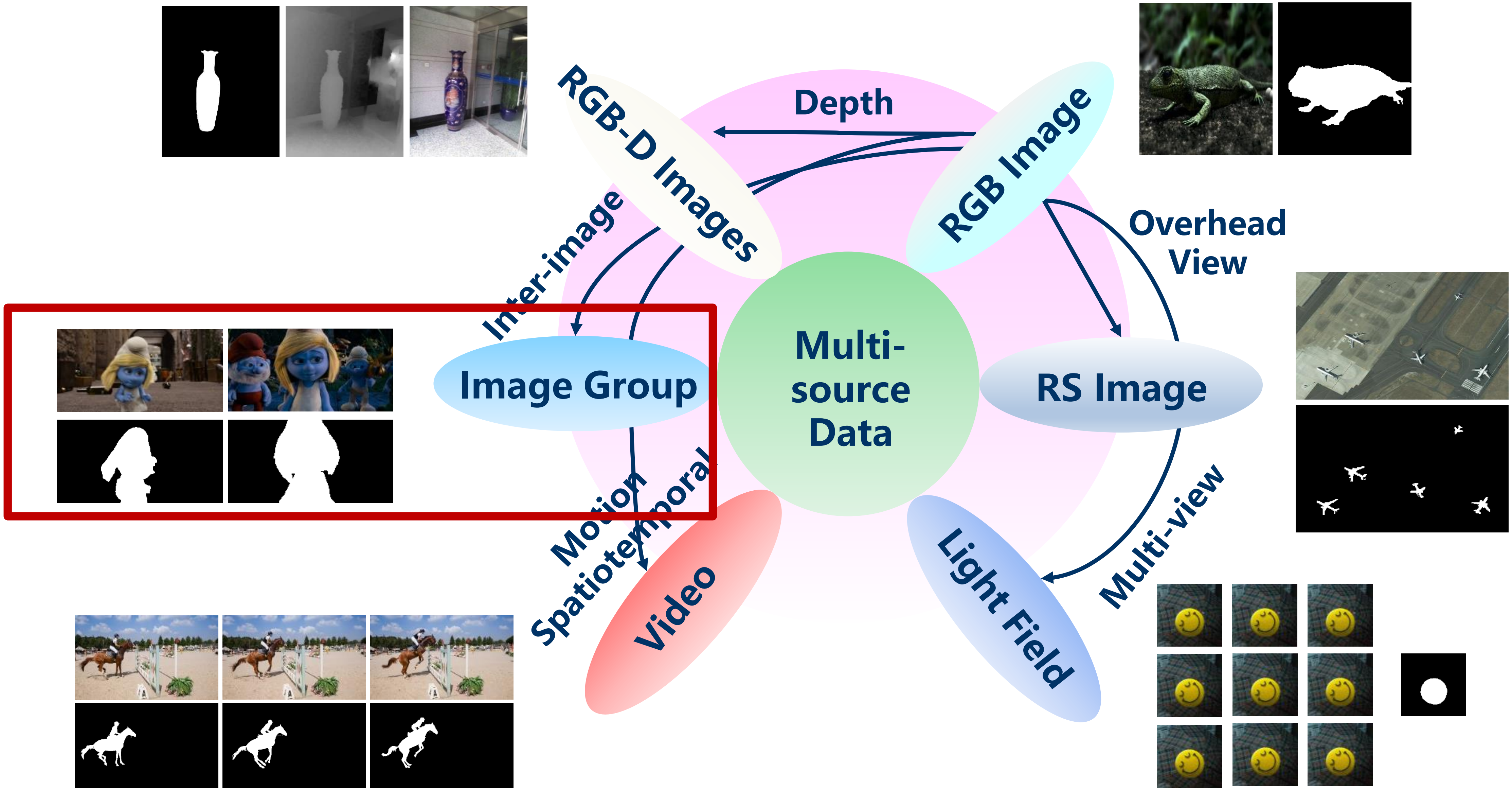
# Introduction

- **What is saliency detection?**

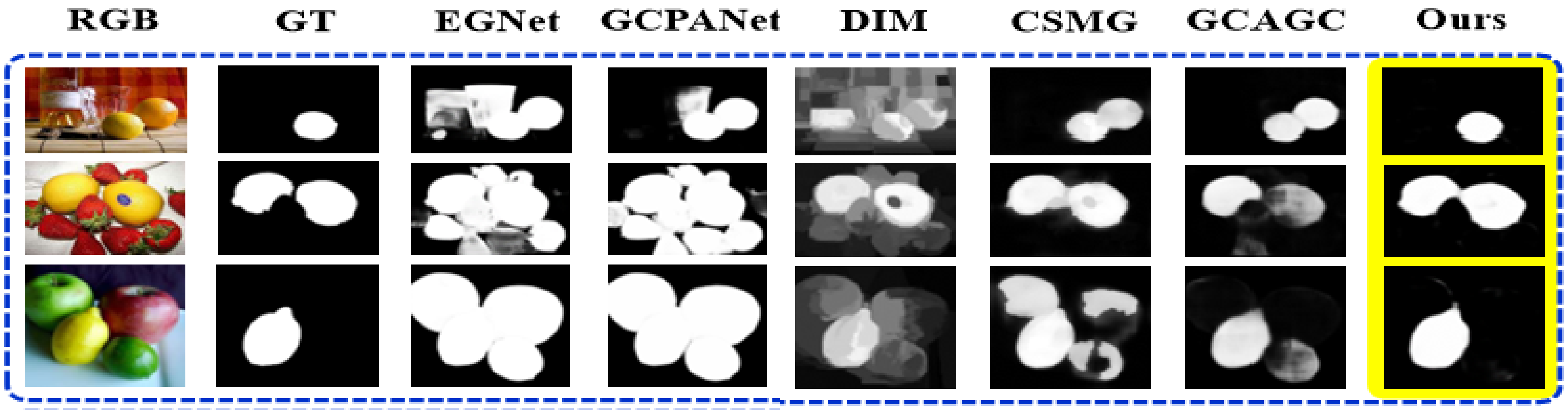


- Saliency detection aims to detecting the salient regions automatically, which has been applied in image/video segmentation, image/video retrieval, image retargeting, video coding, quality assessment, action recognition, and video summarization.
- The last decade has witnessed the remarkable progress of image saliency detection, and a plenty of methods have been proposed based on some priors or techniques, such as uniqueness prior, background prior, compactness prior, sparse coding, random walks, and deep learning.

# Introduction



# Introduction



- Co-Salient Object Detection (CoSOD) aims at discovering the salient objects that repeatedly appear in a query group containing two or more relevant images.
- One challenging issue is **how to effectively capture the co-saliency cues by modeling and exploiting the inter-image relationships**.

# Motivation

- **Insufficient group-wise relationship modeling.** The learned group representations in the previous studies vary with different order of the input group images, leading to unstable training and vulnerable inference.
- **Competition between intra-image saliency and inter-image correspondence.** The learned group semantics in the previous studies were directly duplicated and concatenated with individual features. In fact, this operation implies that different individuals receive identical group semantics, which may propagate redundant and distracting information from the interactions among other images.
- **Weakened group consistency during feature decoding.** In the feature decoding of the CoSOD task, existing up-sampling or deconvolution based methods ignore the maintenance of inter-image consistency, which may lead to the inconsistency of co-salient objects among different images and introduce additional artifacts.

# Contributions

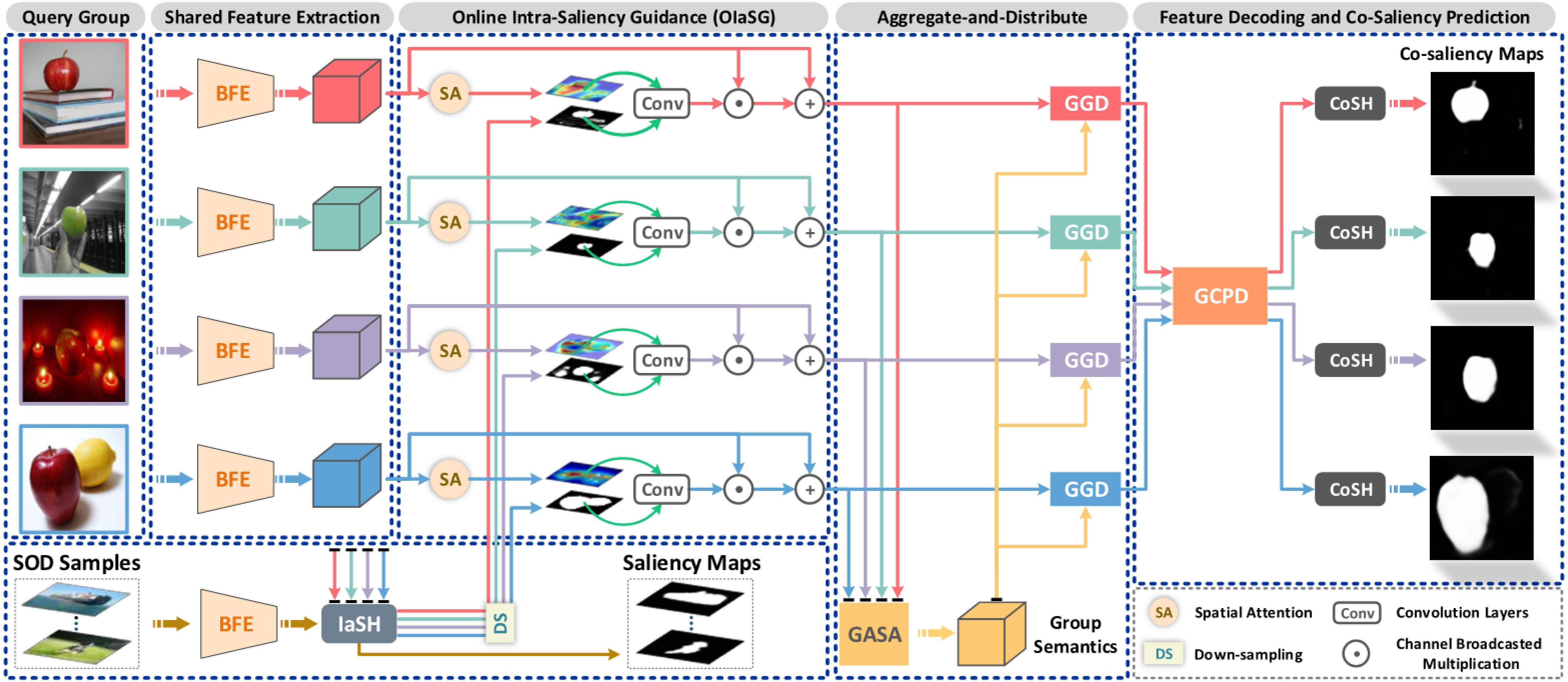
**The proposed CoADNet provides some insights and improvements in terms of modeling and exploiting inter-image relationships in the CoSOD workflow, and produces more accurate and consistent co-saliency results on four prevailing co-saliency benchmark datasets.**

We design an **online intra-saliency guidance (OIaSG)** module for supplying saliency prior knowledge, which is jointly optimized to generate trainable saliency guidance information.

We propose a **two-stage aggregate-and-distribute architecture** to learn group-wise correspondences and co-saliency features, including a group-attentional semantic aggregation (GASA) and a gated group distribution (GGD) module.

A **group consistency preserving decoder (GCPD)** is designed to exploit more sufficient inter-image constraints to generate full-resolution co-saliency maps while maintaining group-wise consistency.

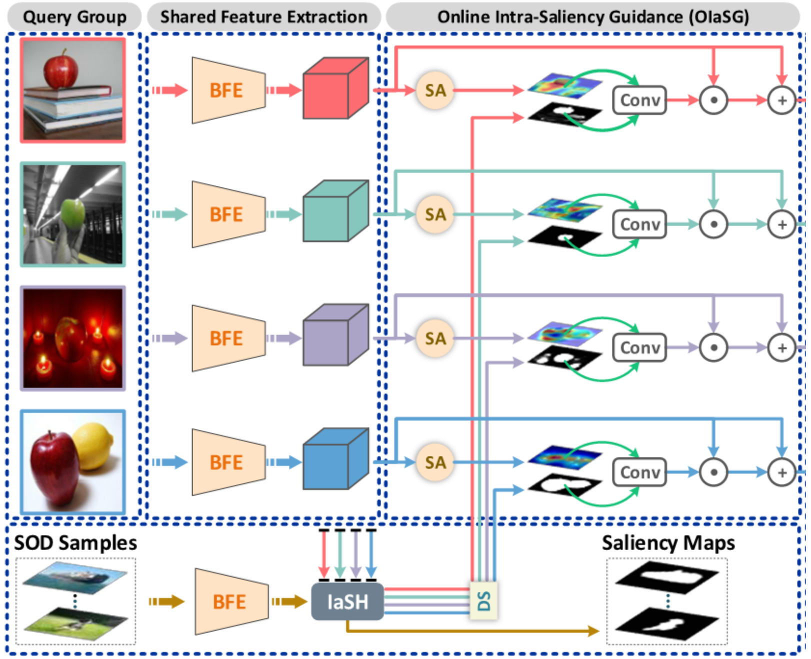
# Our Method


Query Group
Shared Feature Extraction
Online Intra-Saliency Guidance (OIaSG)
Aggregate-and-Distribute
Feature Decoding and Co-Saliency Prediction
BFE
SA
Conv
GGD
GCPD
CoSH
Co-saliency Maps
SOD Samples
IaSH
DS
Saliency Maps
GASA
Group Semantics
SA Spatial Attention
Conv Convolution Layers
DS Down-sampling
⊙ Channel Broadcasted Multiplication

# Online Intra-Saliency Guidance

The challenges of CoSOD are that 1) the salient objects within an individual image may not occur in all the other group images, and 2) the repetitive patterns are not necessarily visually attractive, making it difficult to learn a unified representation to combine these two factors. Thus, **we adopt a joint learning framework to provide trainable saliency priors as guidance information to suppress background redundancy**.

- intra-saliency head (IaSH) to infer online saliency maps;
- fuse online saliency priors with spatial feature in an attention way;
- In this way, we obtain a set of intra-saliency features (IaSFs) $\{U^{(n)}\}_{n=1}^{N}$ with suppressed background redundancy.


Query Group
Shared Feature Extraction
Online Intra-Saliency Guidance (OIaSG)
BFE
SA
Conv
⊙
+
SOD Samples
BFE
IaSH
DS
Saliency Maps

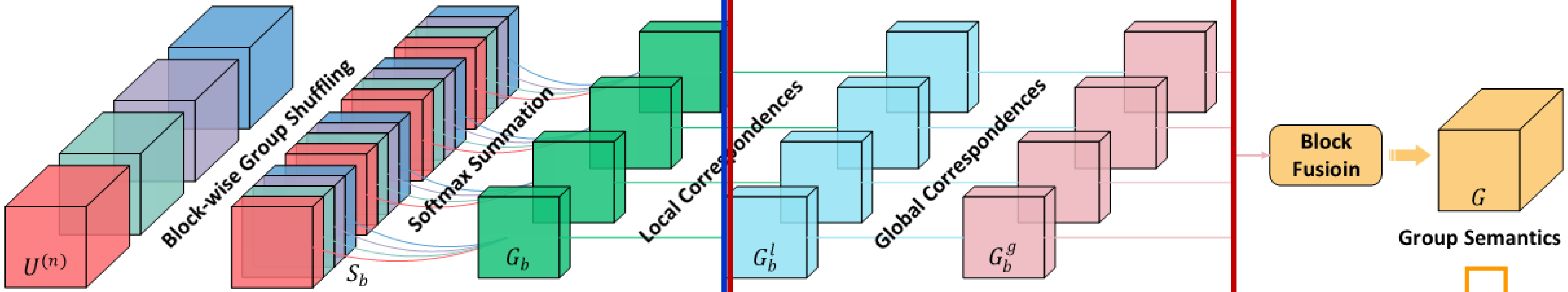
# Group-Attentional Semantic Aggregation

To efficiently capture discriminative and robust group-wise relationships, we investigate three key criteria:

1) **Insensitivity to input order** means that the learned group representations should be insensitive to the input order of group images;
2) **Robustness to spatial variation** considers the fact that co-salient objects may be located at different positions across images;
3) **Computational efficiency** takes the computation burden into account especially when processing large query groups or high-dimensional features.

⇨ we propose a computation-efficient and order-insensitive group-attentional semantic aggregation (GASA) module which builds local and global associations of co-salient objects in group-wise semantic context.

# Group-Attentional Semantic Aggregation



block-wise group shuffling that re-arranges the feature channels at the block level
channel-wise softmax + summation

local + global

model inter-image relationships with long-range semantic dependencies

high computational complexity

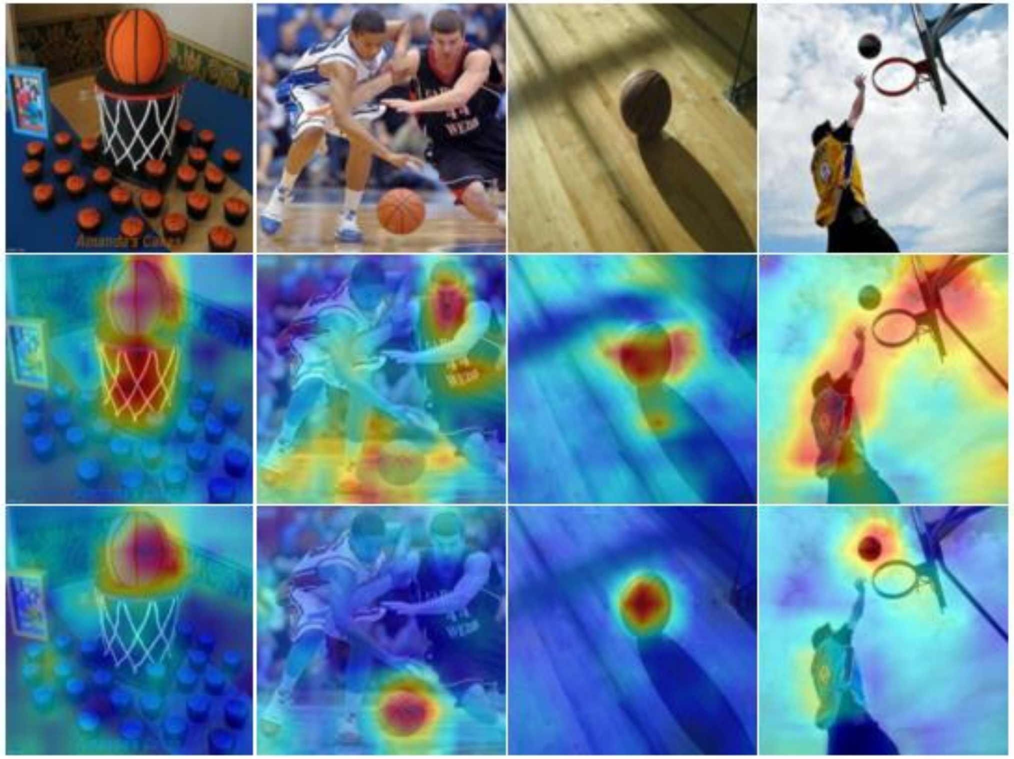
order-sensitivity

spatial variation of CoSal object

# Gated Group Distribution

The group-wise semantics encode the relationships of all images, which may include some distracting information redundancy for co-saliency prediction of different images.

**We propose a gated group distribution (GGD) module to adaptively distribute the most useful group-wise information to each individual.** To achieve this, we construct a group importance estimator that learns dynamic weights to combine group semantics with different IaSFs through a gating mechanism.



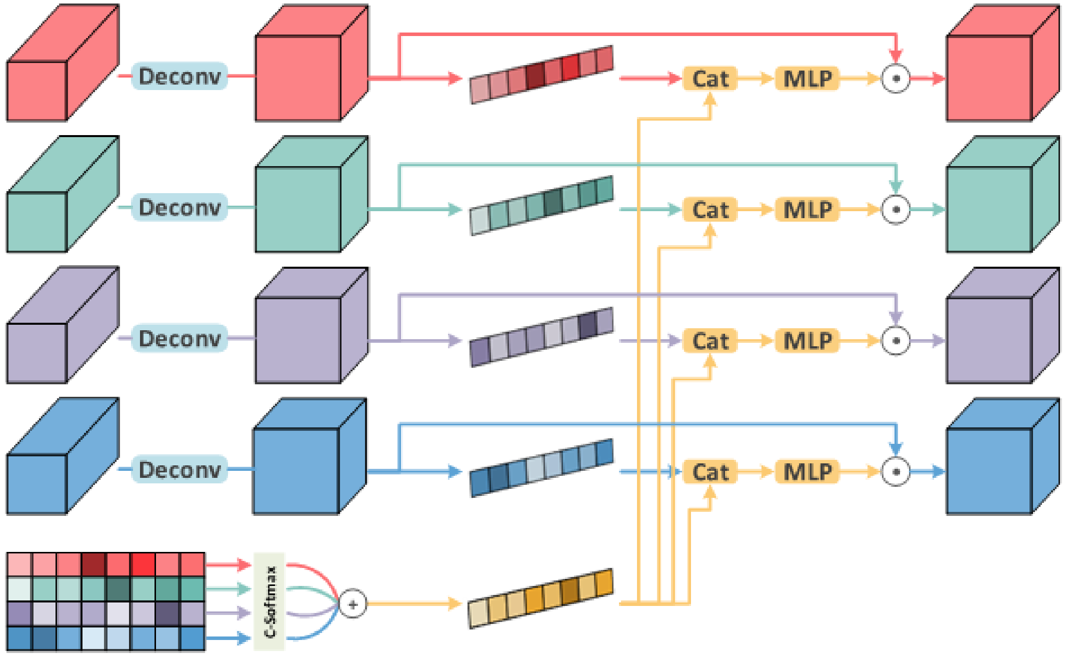
$$X^{(n)} = P \otimes G + (1 - P) \otimes U^{(n)}$$

$$P = \sigma(f^p(SE(U_g^{(n)})))$$

# Group Consistency Preserving Decoder

The most common up-sampling or deconvolution based feature decoders are not suitable for CoSOD tasks because they **ignore the inter-image constraints and may weaken the consistency between images during the prediction process**. Thus, **we propose a group consistency preserving decoder (GCPD) to consistently predict full-resolution co-saliency maps.**

- GCPD includes three cascaded feature decoding (FD) units；
- Learn a compact group feature vector y, and combine it with the vectorized deconvolution representations；
- the finest spatial resolution, which are further fed into a shared co-saliency head (CoSH) to generate full-resolution co-saliency maps



# Supervisions

We jointly optimize the co-saliency and single image saliency predictions in a multi-task learning framework.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_c + \beta \cdot \mathcal{L}_s$$

co-saliency loss:

$$\mathcal{L}_c = -(\sum_{n=1}^{N}(T_c^{(n)} \cdot log(M^{(n)}) + (1 - T_c^{(n)})] \cdot log(1 - M^{(n)})))/N$$

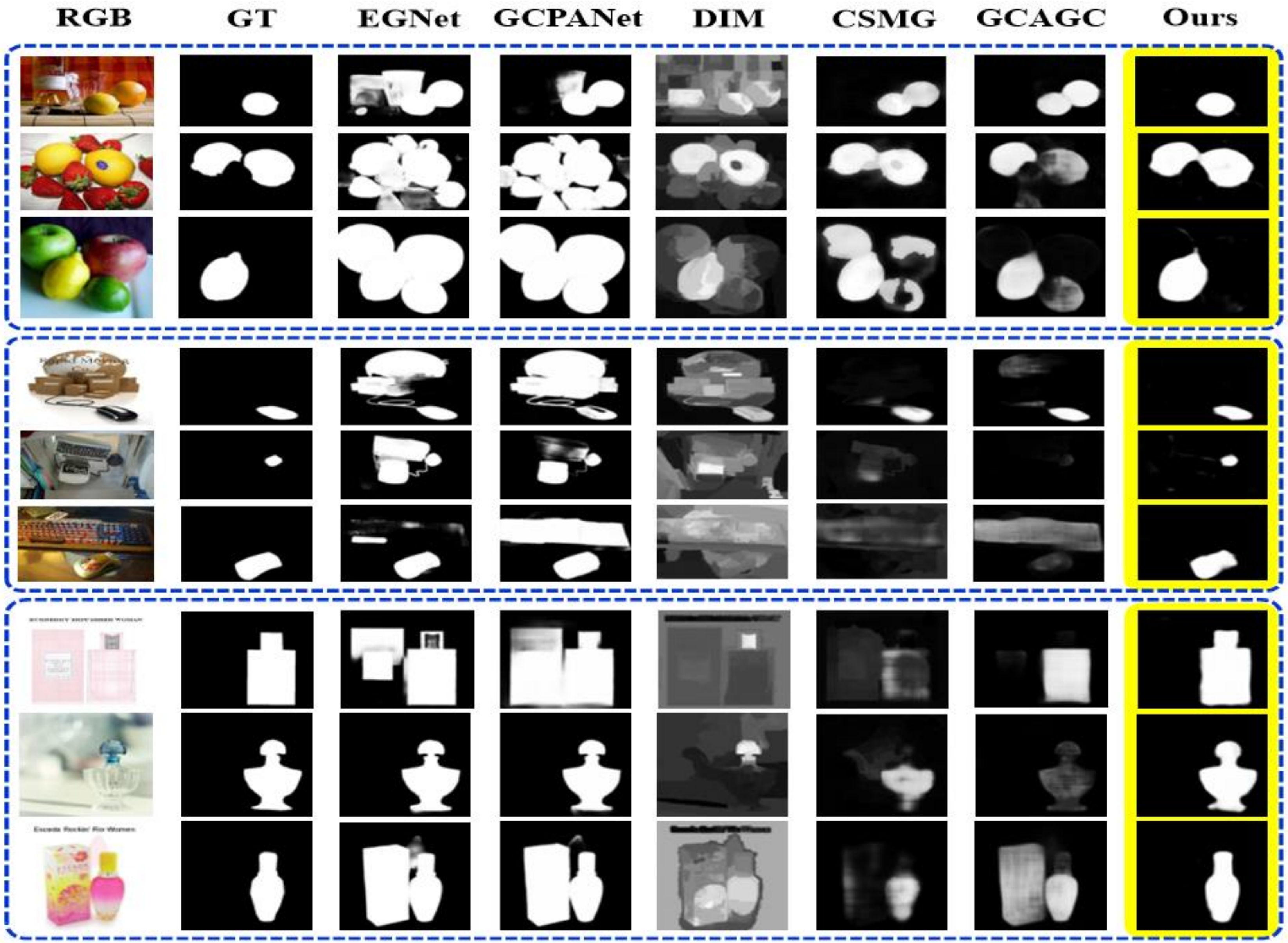
auxiliary saliency loss:

$$\mathcal{L}_s = -(\sum_{k=1}^{K}(T_s^{(k)} \cdot log(A^{(k)}) + (1 - T_s^{(k)}) \cdot log(1 - A^{(k)})))/K$$

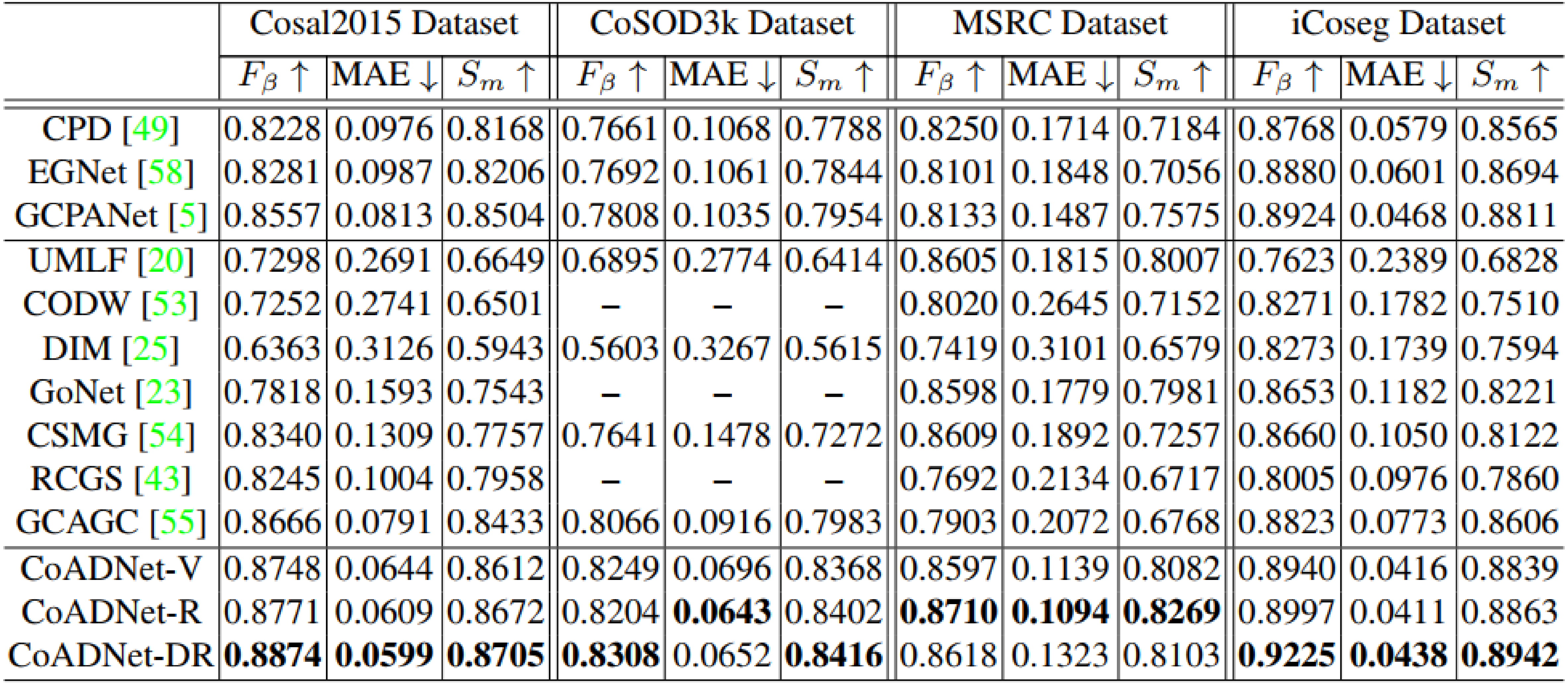
# Experiments

- Benchmark Datasets: CoSOD3k, Cosal2015, MSRC, and iCoseg.
- Evaluation Metrics: Precision-Recall (P-R) curve, F-measure, MAE score, and S-measure
- Implementation Details: a sub-group containing 5 images are randomly selected from a certain query group. All input images are resized to 224 ×224. In each training iteration, 24 sub-groups from COCO-SEG and 64 samples from DUTS are simultaneously fed into the network for optimizing the objective function. In our experiment, we provide the results under two backbones including ResNet-50 and Dilated ResNet-50, and the training process converges until 50,000 iterations. The average inference time for a single image is 0.07 seconds.

# Experiments

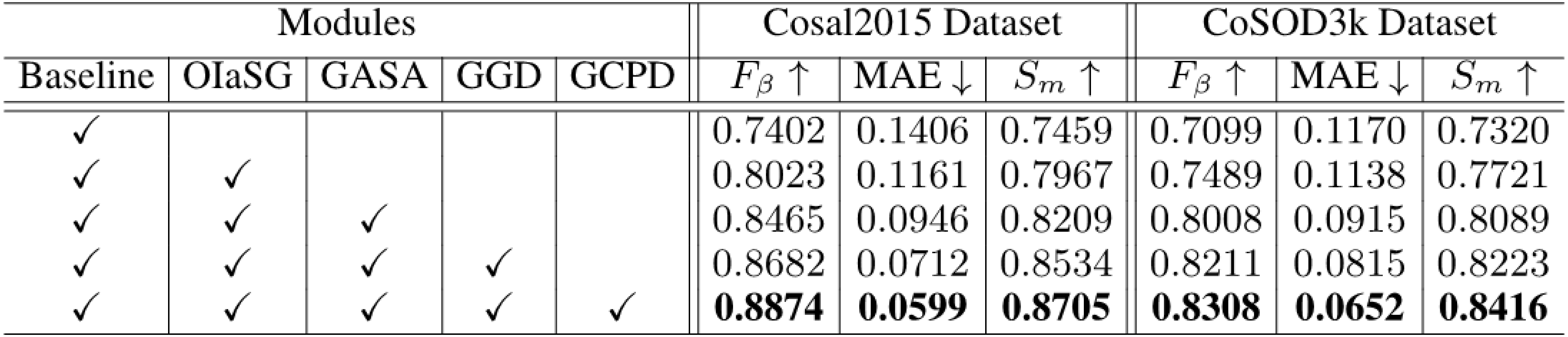


# Experiments



| | Cosal2015 Dataset | | | CoSOD3k Dataset | | | MSRC Dataset | | | iCoseg Dataset | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ |
| CPD [49] | 0.8228 | 0.0976 | 0.8168 | 0.7661 | 0.1068 | 0.7788 | 0.8250 | 0.1714 | 0.7184 | 0.8768 | 0.0579 | 0.8565 |
| EGNet [58] | 0.8281 | 0.0987 | 0.8206 | 0.7692 | 0.1061 | 0.7844 | 0.8101 | 0.1848 | 0.7056 | 0.8880 | 0.0601 | 0.8694 |
| GCPANet [5] | 0.8557 | 0.0813 | 0.8504 | 0.7808 | 0.1035 | 0.7954 | 0.8133 | 0.1487 | 0.7575 | 0.8924 | 0.0468 | 0.8811 |
| UMLF [20] | 0.7298 | 0.2691 | 0.6649 | 0.6895 | 0.2774 | 0.6414 | 0.8605 | 0.1815 | 0.8007 | 0.7623 | 0.2389 | 0.6828 |
| CODW [53] | 0.7252 | 0.2741 | 0.6501 | – | – | – | 0.8020 | 0.2645 | 0.7152 | 0.8271 | 0.1782 | 0.7510 |
| DIM [25] | 0.6363 | 0.3126 | 0.5943 | 0.5603 | 0.3267 | 0.5615 | 0.7419 | 0.3101 | 0.6579 | 0.8273 | 0.1739 | 0.7594 |
| GoNet [23] | 0.7818 | 0.1593 | 0.7543 | – | – | – | 0.8598 | 0.1779 | 0.7981 | 0.8653 | 0.1182 | 0.8221 |
| CSMG [54] | 0.8340 | 0.1309 | 0.7757 | 0.7641 | 0.1478 | 0.7272 | 0.8609 | 0.1892 | 0.7257 | 0.8660 | 0.1050 | 0.8122 |
| RCGS [43] | 0.8245 | 0.1004 | 0.7958 | – | – | – | 0.7692 | 0.2134 | 0.6717 | 0.8005 | 0.0976 | 0.7860 |
| GCAGC [55] | 0.8666 | 0.0791 | 0.8433 | 0.8066 | 0.0916 | 0.7983 | 0.7903 | 0.2072 | 0.6768 | 0.8823 | 0.0773 | 0.8606 |
| CoADNet-V | 0.8748 | 0.0644 | 0.8612 | 0.8249 | 0.0696 | 0.8368 | 0.8597 | 0.1139 | 0.8082 | 0.8940 | 0.0416 | 0.8839 |
| CoADNet-R | 0.8771 | 0.0609 | 0.8672 | 0.8204 | **0.0643** | 0.8402 | **0.8710** | **0.1094** | **0.8269** | 0.8997 | 0.0411 | 0.8863 |
| CoADNet-DR | **0.8874** | **0.0599** | **0.8705** | **0.8308** | 0.0652 | **0.8416** | 0.8618 | 0.1323 | 0.8103 | **0.9225** | **0.0438** | **0.8942** |

# Experiments



| Modules | | | | | Cosal2015 Dataset | | | CoSOD3k Dataset | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | OIaSG | GASA | GGD | GCPD | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ |
| ✓ | | | | | 0.7402 | 0.1406 | 0.7459 | 0.7099 | 0.1170 | 0.7320 |
| ✓ | ✓ | | | | 0.8023 | 0.1161 | 0.7967 | 0.7489 | 0.1138 | 0.7721 |
| ✓ | ✓ | ✓ | | | 0.8465 | 0.0946 | 0.8209 | 0.8008 | 0.0915 | 0.8089 |
| ✓ | ✓ | ✓ | ✓ | | 0.8682 | 0.0712 | 0.8534 | 0.8211 | 0.0815 | 0.8223 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **0.8874** | **0.0599** | **0.8705** | **0.8308** | **0.0652** | **0.8416** |

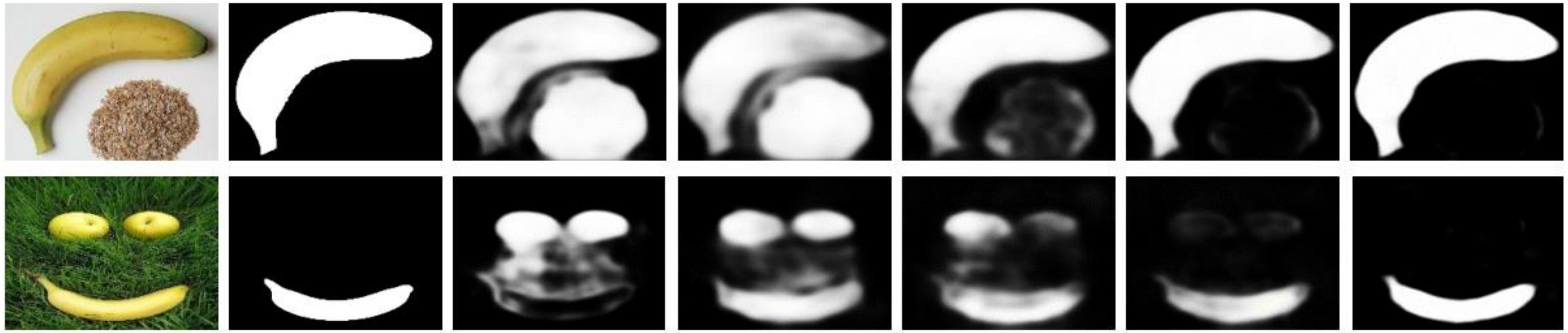


Figure 6: Visualization of different ablative results. From left to right: Input image group, Ground truth, Co-saliency maps produced by the Baseline, Baseline+OIaSG, Baseline+OIaSG+GASA, Baseline+OIaSG+GASA+GGD, and the full CoADNet.

# Experiments

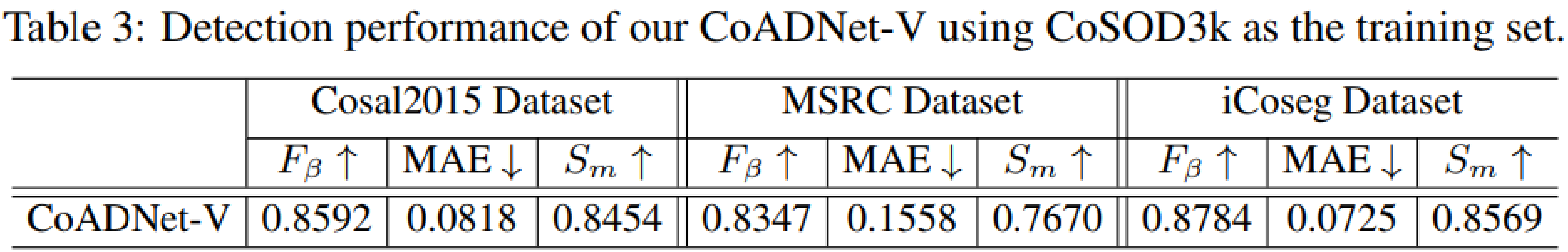


Table 3: Detection performance of our CoADNet-V using CoSOD3k as the training set.

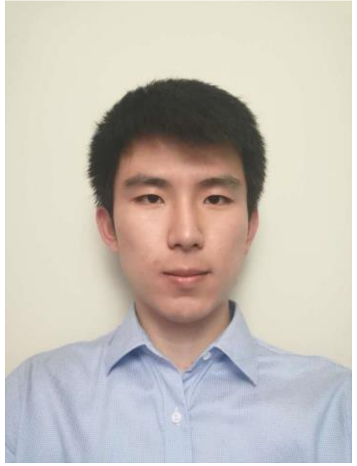
| | Cosal2015 Dataset | | | MSRC Dataset | | | iCoseg Dataset | | |
|---|---|---|---|---|---|---|---|---|---|
| | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ | $F_\beta \uparrow$ | MAE $\downarrow$ | $S_m \uparrow$ |
| CoADNet-V | 0.8592 | 0.0818 | 0.8454 | 0.8347 | 0.1558 | 0.7670 | 0.8784 | 0.0725 | 0.8569 |

**We need an appropriate dataset to train our CoSOD network!!**

# Conclusion

- We proposed an end-to-end CoSOD network by investigating how to model and utilize the inter-image correspondences.
- We first decoupled the single-image SOD from the CoSOD task and proposed an OIaSG module to provide learnable saliency prior guidance.
- Then, the GASA and GGD modules are integrated into a two-stage aggregate-and-distribute structure for effective extraction and adaptive distribution of group semantics.
- Finally, we designed a GCPD structure to strengthen inter-image constraints and predict full-resolution co-saliency maps.
- Experimental results and ablative studies demonstrated the superiority of the proposed CoADNet and the effectiveness of each component.

# Thanks to My Co-authors



张琦坚博士
@CityU

李重仪博士
@NTU

侯军辉助理教授
@CityU

赵耀教授
@BJTU

# 数字媒体信息处理研究中心

## Center of Digital Media Information Processing

北京交通大学数字媒体信息处理研究中心肇始于1998年，入选科技部“重点领域创新团队”、教育部“创新团队发展计划”。该中心现有教师12人，博、硕士研究生100余人。其中教授8人，副教授3人，包括教育部“长江学者” 特聘教授1人，国家杰出青年基金获得者人1人，国家“万人计划”科技创新领军人才1人，教育部新世纪优秀人才支持计划入选者2人，北京市杰出青年基金获得者1人，北京市科技新星3人，北京市科协青年人才托举工程入选者1人，香江学者1人。该中心的研究领域为数字媒体信息处理，研究方向主要包括图像\视频编码与传输、数字水印与数字取证、媒体内容分析与理解等。

赵　耀　教授

教育部长江学者特聘教授

国家杰出青年基金获得者

万人计划科技创新领军人才

教　授：赵　耀、朱振峰、倪蓉蓉、白慧慧
　　　　韦世奎、李晓龙、林春雨、张淳杰

副教授：常冬霞、刘美琴、丛润民

助　理：宋亚男

http://mepro.bjtu.edu.cn/index.html

https://rmcong.github.io/    rmcong@bjtu.edu.cn