Cascade Graph Neural Networks for RGB-D Salient Object Detection

Ao Luo[1⋆], Xin Li[2⋆], Fan Yang[2], Zhicheng Jiao[3], Hong Cheng[1✉], and Siwei Lyu[4]

[1] Center for Robotics, School of Automation Engineering, UESTC, Chengdu, China
[2] Group 42 (G42), Abu Dhabi, UAE
[3] University of Pennsylvania, Philadelphia, USA
[4] University at Albany, State University of New York, USA
{aoluo,xinli,fanyang}_uestc@hotmail.com; hcheng@uestc.edu.cn

Abstract. In this paper, we study the problem of salient object detection (SOD) for RGB-D images using both color and depth information. A major technical challenge in performing salient object detection from RGB-D images is how to fully leverage the two complementary data sources. Current works either simply distill prior knowledge from the corresponding depth map for handling the RGB-image or blindly fuse color and geometric information to generate the coarse depth-aware representations, hindering the performance of RGB-D saliency detectors. In this work, we introduce *Cascade Graph Neural Networks* (Cas-Gnn), a unified framework which is capable of comprehensively distilling and reasoning the mutual benefits between these two data sources through a set of cascade graphs, to learn powerful representations for RGB-D salient object detection. Cas-Gnn processes the two data sources individually and employs a novel *Cascade Graph Reasoning* (CGR) module to learn powerful dense feature embeddings, from which the saliency map can be easily inferred. Contrast to the previous approaches, the explicitly modeling and reasoning of high-level relations between complementary data sources allows us to better overcome challenges such as occlusions and ambiguities. Extensive experiments demonstrate that Cas-Gnn achieves significantly better performance than all existing RGB-D SOD approaches on several widely-used benchmarks. Code is available at https://github.com/LA30/Cas-Gnn.

Keywords: Salient object detection, RGB-D perception, graph neural networks

1 Introduction

Salient object detection is the crux to dozens of high-level AI tasks such as object detection or classification [52,80,69], weakly-supervised semantic segmentation [30,63], semantic correspondences [77] and others [35,72,71]. An ideal solution should identify salient objects of varying shape and appearance, show

⋆ Equal contribution
✉ Corresponding author

robustness towards heavy occlusion, various illumination and background. With the development of hardware (sensors and GPU), prediction accuracy of data-driven methods that use deep networks [87,42,74,56,79,68,67,84,22,10,37] have been improved significantly, compared to traditional methods based on hand-crafted features [41,12,81,82]. However, these approaches only take the appearance features from RGB data into consideration, making them unreliable when handling the challenging cases, such as poorly-lighted environments and low-contrast scenes, due to the lack of depth information.

The depth map captured by RGB-D camera preserves important geometry information of the given scene, allowing 2D algorithms to be extend into 3D space. Depth awareness has been proven to be crucial for many applications of scene understanding, *e.g.,* scene parsing [61,29], 6D object pose estimation [58,27] and object detection [24,49], leading to a significant performance enhancement. Recently, there have been a few attempts to take into account the 3D geometric information for salient object detection in the given scene, *e.g.,* by distilling prior knowledge from the depth [51] or incorporating depth information into a SOD framework [86,48,21]. These RGB-D models have achieved better performances than RGB-only models in salient object detection when dealing with challenging cases. However, as we demonstrate empirically, existing RGB-D salient object detection models fall short under heavy occlusions and depth image noise. One primary reason is that these models, which only focus on delivering or gathering information, ignore modeling and reasoning over high-level relations between two data sources. Therefore, it is hard for them to fully exploit the complementary nature of 2D color and 3D depth information for overcoming the ambiguities in complex scenes. These observations inspire us to think about: *How to explicitly reason on high-level relations over 2D appearance (color) and 3D geometry (depth) information for better inferring salient regions*?

Graph neural network (GNN) has been shown to be an optimal way of relation modeling and reasoning [54,11,88,73,62]. Generally, a GNN model propagates messages over a graph, such that the node's representation is not only obtained from its own information but also conditioned on its relations to the neighboring nodes. It has revolutionized deep representation learning and benefitted many computer vision tasks, such as 3D pose estimation [5], action recognition [87], zero-shot learning [70] and language grounding [1], by incorporating graph computation into deep learning frameworks. However, how to design a suitable GNN model for RGB-D based SOD is challenging and, to the best of our knowledge, is still unexplored.

In this paper, we present the first attempt to build a GNN-based model, namely *Cascade Graph Neural Networks* (Cas-Gnn), to explicitly reason about the 2D appearance and 3D geometry information for RGB-D salient object detection. Our proposed deep model including multiple graphs, where each graph is used to handle a specific level of *cross-modality* reasoning. In each graph, two basic types of nodes are contained, *i.e.,* `geometry nodes` storing depth features and `appearance nodes` storing RGB-related features, and they are linked to each other by edges. Through message passing, the useful mutual information

and high-level relations between two data sources can be gradually distilled for learning the powerful dense feature embeddings, from which the saliency map can be inferred. To further enhance the capability for reasoning over multiple levels of features, we make our Cas-Gnn to have these multi-level graphs sequentially chained by coarsening the preceding graph into two domain-specific `guidance nodes` for the following cascade graph. Consequently, each graph in our Cas-Gnn (except for the first cascade graph) has three types of nodes in total, and they distill useful information from each other to build powerful feature representations for RGB-D based salient object detection.

Our Cas-Gnn is easy to implement and end-to-end learnable. As opposed to prior works which simply fuse features of the two data sources, Cas-Gnn is capable of explicitly reasoning about the 2D appearance and 3D geometry information over chained graphs, which is essential to handle heavy occlusions and ambiguities. Extensive experiments show that our Cas-Gnn performs remarkably well on 7 widely-used datasets, outperforming state-of-the-art approaches by a large margin. In summary, our major contributions are described below:

1) We are the first to use the graph-based techniques to design network architectures for RGB-D salient object detection. This allows us to fully exploit the mutual benefits between the 2D appearance and 3D geometry information for better inferring salient object(s).
2) We propose a graph-based, end-to-end trainable model, called *Cascade Graph Neural Networks* (Cas-Gnn), for RGB-D based SOD, and carefully design *Graph-based Reasoning* (GR) module to distill useful knowledge from different modalities for building powerful feature embeddings.
3) Different from most GNN-based approaches, our Cas-Gnn ensembles a set of cascade graphs to reason about relations of the two data sources hierarchically. This cascade reasoning capability ensures the graph-based model to exploit rich, complementary information from multi-level features, which is useful in capturing object details and overcoming ambiguities.
4) We conduct extensive experiments on 7 widely-used datasets and show that our Cas-Gnn sets new records, outperforming state-of-the-art approaches.

2 Related Work

This work is related to RGB-D based salient object detection, graph neural network and network cascade. Here, we briefly review these three lines of works.

RGB-D Salient Object Detection. Unlike approaches for RGB-only salient object detection methods [75,23,39,67,84,22,43,17,22,37,41,12,81,82] which only focus on 2D appearance feature learning, RGB-D based SOD approaches [86,48,21] take two different data sources, *i.e.,* 2D appearance (color) and 3D geometry (depth) information, into consideration. Classical approaches extract handcrafted features from the input RGB-D data and perform cross-modality feature fusion by various strategies, such as random forest regressor [55] and minimum barrier distance [57]. However, with handcrafting of features, classic RGB-D

based approaches are limited in the expression ability. Recent works such as CPFP [86] integrates deep feature learning and cross-modality fusion within a unified, end-to-end framework. Piao *et al.* [48] futher enhance the cross-modality feature fusion through a recurrent attention mechanism. Fan *et al.* [21] introduce a depth-depurator to filter out noises in the depth map for better fusing cross-modality features. These approaches, despite the success, are not able to fully reason the high-order relations of cross-modality data, making them unreliable when handling challenges such as occlusions and ambiguities. In comparison, our Cas-Gnn considers a better way to distill the mutual benefit of the two data sources by modeling and reasoning their relations over a set of cascade graphs, and we show that such cross-modality reasoning boosts the performance significantly.

Graph Neural Networks. In recent years, a wide variety of graph neural network (GNN) based models [16,15,53,33] have been proposed for different applications [54,11,88,4,44]. Generally, a GNN can be viewed as a message passing algorithm, where representations for nodes are iteratively computed conditioned on their neighboring nodes through a differentiable aggregation function. Some typical applications in computer vision include semantic segmentation [50], action recognition [65], point cloud classification and segmentation [66], to name a few. In the context of RGB-D based salient object detection – the task that we study in this paper – a key challenge in applying GNNs comes from how the graph model learns high-level relations and low-level details simultaneously. To solve this problem, unlike existing graph models, we ensemble a set of sequentially chained graphs to form a unified, cascade graph reasoning model. Therefore, our Cas-Gnn is able to reason about relations across multiple feature levels to capture important hierarchical information for RGB-D based SOD, which is significantly different from all existing GNN based models.

Network Cascade. Network cascade is an effective scheme for a variety of high-level vision applications. Popular examples of cascaded models include DeCaFA for face alignment [14], BDCN for edge detection [26], Bidirectional FCN for object skeleton extraction [76], and Cascade R-CNN for object detection [6], to name a few. The core idea of network cascade is to ensemble a set of models to handle challenging tasks in a *coarse-to-fine* or *easy-to-hard* manner. For salient object detection in RGB-only images, only a few attempts employ the network cascade scheme. Li *et al.* [36] use a cascade network for gradually integrating saliency prior knowledge from coarse to fine. Wu *et al.* [67] design a cascaded partial decoder to enhance the learned features for salient object detection. Different from these approaches, our Cas-Gnn propagates the knowledge learned from a more global view to assist fine-grained reasoning by chaining multiple graphs, which aids a structured understanding of complex scenes.

3 Method

The key idea of Cas-Gnn is that it enables the fully harvesting of the 2D appearance and 3D geometric information by using a differentiable, cascade mod-

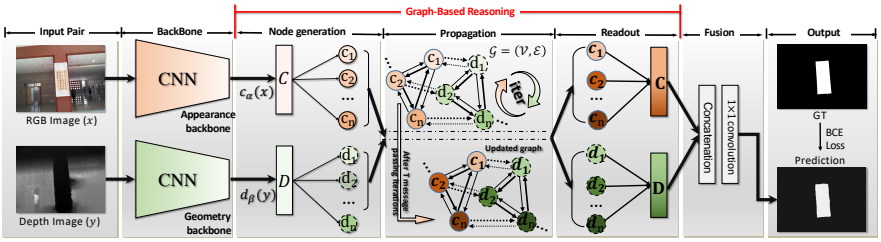


Fig. 1. Overall of our simple cross-modality reasoning model. Our model is built upon two VGG-16 based backbones, and uses a novel graph-based reasoning (GR) module to reason about the high-level relations between the generated 2D appearance and 3D geometry nodes for building more powerful representations. The updated node representations from two modalities are finally fused to infer the salient object regions.

ule to hierarchically reason about relations between the two data sources. In this section, we elaborate on how to design a graph reasoning module and how to further enhance the capability of graph-based reasoning using the network cascade technique.

3.1 Problem Formulation

The task of RGB-D based salient object detection is to predict a saliency map $z \in \mathcal{Z}$ given an input image $x \in \mathcal{X}$ and its corresponding depth image $y \in \mathcal{Y}$. The input space $\mathcal{X}$ and $\mathcal{Y}$ correspond to the space of images and depths respectively, and the target space $\mathcal{Z}$ consists of only one class. A regression problem is characterized by a continuous target space. In our approach, a graph-based model is defined as a function $f_\Theta : \{\mathcal{X}, \mathcal{Y}\} \mapsto \mathcal{Z}$, parameterized by Θ, which maps an input pair, *i.e.,* $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, to an output $f_\Theta(x, y) \in \mathcal{Z}$. The key challenging is to design a suitable model Θ that can fully exploit useful information from the two data sources (color and depth image) to learn powerful representations so that it can make the mapping more accurately.

3.2 Cross-modality Reasoning with Graph Neural Networks

We start out with a simple GNN model, which reasons over the cross-modality relations between 2D appearance (color) and 3D geometric (depth) information across multiple scales, for salient object detection, as shown in Fig. 1.

Overview. For RGB-D salient object detection, the key challenge is to fully mine useful information from the two complementary data sources, *i.e.,* the color image $x \in \mathcal{X}$ and the depth $y \in \mathcal{Y}$, and learn the mapping function $f_\Theta(x, y)$ which can infer the saliency regions $z \in \mathcal{Z}$. Aiming to achieve this goal, we represent the extracted multi-scale color features $C = \{c_1, \cdots, c_n\}$ and depth features $D = \{d_1, \cdots, d_n\}$ with a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ means a finite set of nodes and $\mathcal{E}$ stands for the edges among them. The nodes in the GNN model

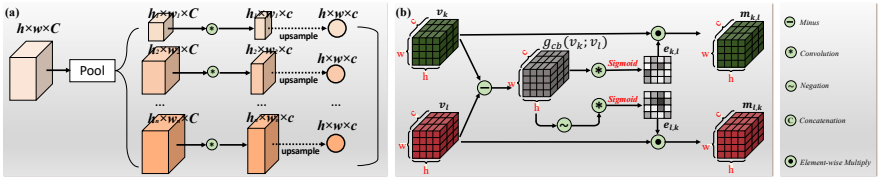


Fig. 2. Detailed illustration of our designs for (a) node embedding and (b) edge embedding. Zoom in for details.

are naturally grouped into two types: the `geometry nodes` $\mathcal{V}_1 = \{c_1, \cdots, c_n\}$ and the `appearance nodes` $\mathcal{V}_2 = \{d_1, \cdots, d_n\}$, where $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$. The edges $\mathcal{E}$ connect **i)** the nodes from the same modality ($\mathcal{V}_1$ or $\mathcal{V}_2$), and **ii)** the nodes of the same scale from different modalities, *i.e.,* $c_i \leftrightarrow d_i$ where $i \in \{1, \cdots, n\}$. For each node, c_i or d_i, we learn its updated representation, namely $\mathbf{c}_i^{(t)}$ or $\mathbf{d}_i^{(t)}$, by aggregating the representations of its neighbors. In the end, the updated features are fused to produce the final representations for salient object detection.

Feature Backbones. Before reasoning the cross-modality relations, we first extract the 2D appearance feature $\mathcal{C}$ and 3D geometry feature $\mathcal{D}$ through the appearance backbone network c_α and geometry backbone network d_β, respectively. Following most of the previous approaches [48,7,9,25,89], we take two VGG-16 networks as the backbones, and use the dilated network technique [78] to ensure that the last two groups of VGG-16 have the same resolution. For the input RGB image x and the corresponding depth image y, we can map them to semantically powerful 2D appearance representations $\mathcal{C} = c_\alpha(x) \in \mathbb{R}^{h \times w \times C}$ and 3D geometry representations $\mathcal{D} = d_\beta(y) \in \mathbb{R}^{h \times w \times C}$. Rather than directly fusing the extracted features $\mathcal{C}$ and $\mathcal{D}$ to form the final representations for RGB-D salient object detection, we introduce a *Graph-based Reasoning* (GR) module to reason about the cross-modality, high-order relations between them to build more powerful embeddings, from which the saliency map can be inferred more easily and accurately.

Graph-based Reasoning Module. The *Graph-based Reasoning* (GR) module g_χ takes the underlying 2D appearance features $\mathcal{C}$ and 3D geometry features $\mathcal{D}$ as inputs, and outputs powerful embeddings $\mathbf{C}$ and $\mathbf{D}$ after performing cross-modality reasoning: $\{\mathbf{C}, \mathbf{D}\} = g_\chi(\mathcal{C}, \mathcal{D})$. We formulate $g_\chi(\cdot, \cdot)$ in a graph-based, end-to-end differentiable way as follows:

1) Graph Construction: Given the 2D appearance features $\mathcal{C}$ and 3D geometry features $\mathcal{D}$, we build a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ which has two types of nodes: the `geometry nodes` $\mathcal{V}_1 = \{c_1, \cdots, c_n\}$ and the `appearance nodes` $\mathcal{V}_2 = \{d_1, \cdots, d_n\}$, where $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$. Each node c_i or d_i is a feature map for a predefined scale s_i and edges link **i)** the nodes from the same modality but different scales, *i.e.,* $c_i \leftrightarrow c_j$ or $d_i \leftrightarrow d_j$, and **ii)** the nodes of the same scale from different modalities, *i.e.,* $c_i \leftrightarrow d_i$. Next, we show how to parameterize the nodes $\mathcal{V}$, edges $\mathcal{E}$, and message passing functions $\mathcal{M}$ of the graph $\mathcal{G}$ with neural networks.

2) Multi-scale Node Embeddings $\mathcal{V}$: Given the 2D appearance features $\mathcal{C}$ and 3D geometry features $\mathcal{D}$, as shown in Fig. 2(a), we leverage the pyramid pooling module (PPM) [85] followed by a convolution layer and an interpolation layer to extract multi-scale features of the two modalities (n scales) as the initial node representations, resulting in $N = 2 \cdot n$ nodes in total. For the appearance node c_i and geometry node d_i, their initial node representations $\mathbf{c}_i^{(0)} \in \mathbb{R}^{h \times w \times c}$ and $\mathbf{d}_i^{(0)} \in \mathbb{R}^{h \times w \times c}$ can be computed as:

$$\mathbf{c}_i^{(0)} = \mathcal{R}_{h \times w}(Conv(\mathcal{P}(\mathcal{C}; s_i))); \quad \mathbf{d}_i^{(0)} = \mathcal{R}_{h \times w}(Conv(\mathcal{P}(\mathcal{D}; s_i))), \tag{1}$$

where $\mathcal{P}(\cdot\ ; s_i)$ means the pyramid pooling operation, which pools the given feature maps to the scale of s_i, and $\mathcal{R}(\cdot)$ is the interpolation operation which ensures multi-scale feature maps to have the same size $h \times w$.

3) Edge Embeddings $\mathcal{E}$: The nodes are linked by edges for information propagation. As mentioned above, in our constructed graph, edges link **i)** the nodes from the same modality but different scales, and **ii)** the nodes of the same scale from different modalities. For simplification, we use v_k and v_l, where v_k, $v_l \in \mathcal{V}$, to represent two nodes linked by the edge[1]. As shown in Fig. 2(b), the edge embedding $\mathbf{e}_{k,l}$ is used to represent the high-level relation on the two sides of the edge from v_k to v_l through a relation function $f_{rel}(\cdot\ ;\ \cdot)$:

$$\mathbf{e}_{k,l} = f_{rel}(\mathbf{v}_k; \mathbf{v}_l) = Conv(g_{cb}(\mathbf{v}_k; \mathbf{v}_l)) \in \mathbb{R}^{h \times w \times c}, \tag{2}$$

where $\mathbf{v}_k$ and $\mathbf{v}_l$ are node embeddings for nodes v_k and v_l respectively, $g_{cb}(\cdot\ ;\ \cdot)$ is a function that combines the node embeddings $\mathbf{v}_k$ and $\mathbf{v}_l$, and $Conv(\cdot)$ is the convolution operation which learns the relations in an end-to-end manner. For the combination function $g_{cb}(\cdot\ ;\ \cdot)$, we follows [66] and model it as: $g_{cb}(\mathbf{v}_k; \mathbf{v}_l) = \mathbf{v}_l - \mathbf{v}_k$. The resulting edge embedding $\mathbf{e}_{k,l}$ for node v_k to v_l is also a c-dimensional feature map with the size of $h \times w$, in which each feature reflects the pixel-wise relationship between linked nodes.

4) Message Passing $\mathcal{M}$: In our GNN model, each node aggregates feature messages from all its neighboring nodes. For the message $\mathbf{m}_{k,l}$ passed from all neighboring nodes v_k to v_l, we define the following message passing function $\mathcal{M}(\cdot\ ;\ \cdot)$:

$$\mathbf{m}_{k,l}^{(t)} = \sum_{k \in \mathcal{N}(l)} \mathcal{M}(\mathbf{v}_k^{(t-1)}, \mathbf{e}_{k,l}^{(t-1)}) = \sum_{k \in \mathcal{N}(l)} sigmoid(\mathbf{e}_{k,l}^{(t-1)}) \cdot \mathbf{v}_k^{(t-1)} \in \mathbb{R}^{h \times w \times c} \tag{3}$$

where $sigmoid(\cdot)$ is the sigmoid function which maps the edge embedding to link weight. Since our GNN model is designed for a pixel-wise task, the link weight between node is represented by a 2D map.

5)Node-state Updating $\mathcal{F}_{update}$: After the t_th message passing step, each node v_l in our GNN model aggregates information from its neighboring nodes to update its orginal feature representations. Here, we model the node-state updating

[1] In our formulation, the edges, message passing function and node-state updating function have no concern with the node types, therefore we simply ignore the node type for more clearly describing the ***3) edge embeddings***, ***4) message passing*** and ***5) node-state updating***.

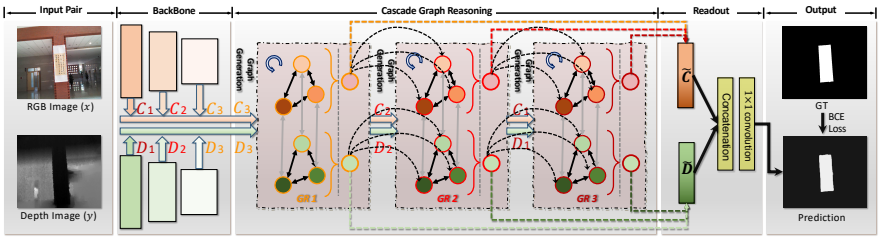


Fig. 3. The overall architecture of our CAS-GNN. Three graph-based reasoning (GR) modules are cascaded in a top-down manner to better distill multi-level information.

process with Gated Recurrent Unit [2],

$$\mathbf{v}_l^{(t)} = \sum_{k \in \mathcal{N}(l)} \mathcal{F}_{update}(\mathbf{v}_l^{(t-1)}, \mathbf{m}_{k,l}^{(t-1)}) = \sum_{k \in \mathcal{N}(l)} \mathcal{U}_{GRU}(\mathbf{v}_l^{(t-1)}, \mathbf{m}_{k,l}^{(t-1)}), \tag{4}$$

where $\mathcal{U}_{GRU}(\cdot\ ;\ \cdot)$ stands for the gated recurrent unit.

6)Saliency Readout $\mathcal{O}$: After T message passing iterations, we upsample all updated node embeddings of each modality to the same size through the interpolation layer $R(\cdot)$, and merge them, *i.e.*, $\mathbf{V}_1 = \{R(\mathbf{c}_i^{(T)})\}_{i=1}^n$ and $\mathbf{V}_2 = \{R(\mathbf{d}_i^{(T)})\}_{i=1}^n$, to form the embeddings:

$$\mathbf{C} = \mathcal{F}_{merge}(\mathbf{V}_1); \quad \mathbf{D} = \mathcal{F}_{merge}(\mathbf{V}_2), \tag{5}$$

where $\mathcal{F}_{merge}(\cdot)$ denotes the merge function which is implemented with a concatenation layer followed by a 3×3 convolution layer. The learned embeddings of each modality can be further fused to form the final representations for RGB-D salient object detection by the following operation:

$$\mathbf{S} = \mathcal{R}_{H \times W}(\mathcal{O}(\mathbf{C}, \mathbf{D})), \tag{6}$$

where $\mathcal{O}(\cdot)$ is the readout function that maps the learned representations to the saliency scores. Here, we implement it with a concatenation layer followed by two 1×1 convolution layers; $\mathcal{R}_{H \times W}(\cdot)$ is used to resize the generated results to the same size of input image $H \times W$ through the interpolation operation.

Overall, all components in our GNN model are formulated in a differentiable manner, and thus can be trained end-to-end. Next, we show how to further enhance the capability of GNN model through network cascade techniques.

3.3 Cascade Graph Neural Networks

In this part, we further enhance our GNN model for RGB-D salient object detection by using the network cascade technique. As observed by many existing works [28,40,59,20], the deep-layer and shallow-layer features are complementary to each other: the deep layer features encode high-level semantic knowledge while

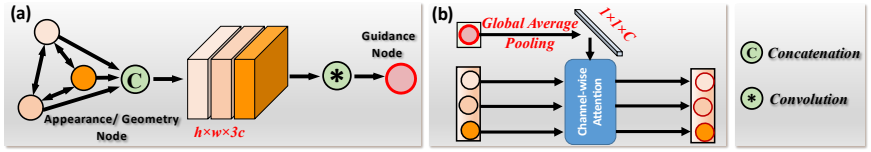


Fig. 4. Detailed illustration of our designs for (a) guidance node generation and (b) attention-based message propagation. Best viewed in color.

the shallow-layer features capture rich spatial information. Ideally, a powerful deep saliency model should be able to fully explore these multi-level features. Aiming to achieve this, we extend our GNN model to a hierarchical GNN model which is able to perform the reasoning across multiple levels for better inferring the salient object regions.

Hierarchical Reasoning via Multi-level Graphs. A straightforward scheme is to ensemble a set of graphs across multiple levels $\{\mathcal{G}_w\}_{w=1}^{W}$ to learn the embeddngs individually, and then fuse the learned representations to build the final representations. Formally, given the VGG-16 based appearance backbone c_α for RGB image $\mathcal{X}$ and geometry backbone d_β for depth image $\mathcal{Y}$, we follow [28] to map the inputs to W levels of side-output features, *i.e.*, the multi-level appearance features $\tilde{\mathcal{V}}_1 = \{\mathcal{C}_1, \cdots, \mathcal{C}_W\}$ and the multi-level geometry features $\tilde{\mathcal{V}}_2 = \{\mathcal{D}_1, \cdots, \mathcal{D}_W\}$. For the features of each level $w \in [1, W]$, we build a graph $\mathcal{G}_w$ and use our proposed *Graph-based Reasoning* (GR) module $g_\chi(\mathcal{C}_w, \mathcal{D}_w)$ to map them to the corresponding embeddings $\{\mathbf{C}_w, \mathbf{D}_w\}_{i=1}^{W}$. Then, these multi-level embeddings of each modality, $\tilde{V}_1 = \{\mathbf{C}_1, \cdots, \mathbf{C}_W\}$ and $\tilde{V}_2 = \{\mathbf{D}_1, \cdots, \mathbf{D}_W\}$, can be easily interpolated to have the same resolution through the interpolation layer $R(\cdot)$, *i.e.*, $\tilde{\mathbf{V}}_1 = \{R(\mathbf{C}_1), \cdots, R(\mathbf{C}_W)\}$ and $\tilde{\mathbf{V}}_2 = \{R(\mathbf{D}_1), \cdots, R(\mathbf{D}_W)\}$, and merged by the following function:

$$\tilde{\mathbf{C}} = \mathcal{M}_{cl}(\tilde{\mathbf{V}}_1); \quad \tilde{\mathbf{D}} = \mathcal{M}_{cl}(\tilde{\mathbf{V}}_2) \tag{7}$$

where $\mathcal{M}_{cl}(\cdot)$ is a merge function, which can be either element-wise addition or channel-wise concatenation. Then, the readout function $\mathcal{O}(\tilde{\mathbf{C}}, \tilde{\mathbf{D}})$ can be used to generate the final results.

Generally, this simply hierarchical approach enables the model to perform reasoning across multiple levels. However, as it treats the multi-level reasoning process independently, the mutual benefits are hard to be fully explored.

Cascade Graph Reasoning. To overcome the drawbacks of independent multi-level (graph-based) reasoning, we propose the *Cascade Graph Reasoning* (CGR) module by chaining these graphs $\{\mathcal{G}_w\}_{w=1}^{W}$ for joint reasoning. The resulting model is called *Cascade Graph Neural Networks* (Cas-Gnn), as shown in Fig. 3. Specifically, our Cas-Gnn includes multi-level graphs $\{\mathcal{G}_w\}_{w=1}^{W}$ which are linked in a top-down manner by coarsening the preceding graph into two domain-specific `guidance nodes` for the following cascade graph to perform the joint reasoning.

1) Guidance Node: Unlike `geometry nodes` and `appearance nodes`, `guidance nodes` only deliver the guidance information, and will stay fixed during the message passing process. In our formulation, for reasoning the cross-modality relations of the w_th cascade stage, its preceding graph (from the deeper side-output level) is mapped into `guidance node` embeddings by the following functions:

$$\mathbf{g}_c^w = \mathcal{F}(\mathbf{V}_1^{(w-1)}); \quad \mathbf{g}_d^w = \mathcal{F}(\mathbf{V}_2^{(w-1)}), \tag{8}$$

where $\mathbf{g}_c^w$ and $\mathbf{g}_d^w$ are the guidance node embeddings of cascade stage w, and $\mathcal{F}(\cdot)$ is the graph merging operator, which coarsens the set of learned node embeddings ($\mathbf{V}_1^{(w-1)} = \{\mathbf{c}_i^{(w-1)(T)}\}_{i=1}^n$ or $\mathbf{V}_2^{(w-1)} = \{\mathbf{d}_i^{(w-1)(T)}\}_{i=1}^n$) of the preceding graph $\mathcal{G}_{(w-1)}$ by firstly concatenating them and then performing the fusion via a 3×3 convolution layer (See Fig. 4(a)).

2) Cascade Message Propagation: Each guidance node, $\mathbf{g}_c^w$ or $\mathbf{g}_d^w$, propagates the guidance information to other nodes of the same domain in the graph $\mathcal{G}_{(w)}$ through the attention mechanism:

$$\breve{\mathbf{v}}_c^{w(t)} = \mathbf{v}_c^{w(t)} \odot \mathcal{A}(\mathbf{g}_c^w); \quad \breve{\mathbf{v}}_d^{w(t)} = \mathbf{v}_d^{w(t)} \odot \mathcal{A}(\mathbf{g}_d^w) \tag{9}$$

where $\breve{\mathbf{v}}_c^{w(t)}$ and $\breve{\mathbf{v}}_d^{w(t)}$ denote the updated appearance node embeddings and geometry node embeddings for the cascade stage w after t_th message passing step respectively; $\odot$ means the channel-wise multiplication. $\mathcal{A}(\cdot)$ is the attention function, which can be formulated as:

$$A(\mathbf{g}_c^w) = sigmoid(\mathcal{P}(\mathbf{g}_c^w)); \quad A(\mathbf{g}_d^w) = sigmoid(\mathcal{P}(\mathbf{g}_d^w)); \tag{10}$$

where $\mathcal{P}(\cdot)$ is the global average pooling operation, and the *sigmoid* is used to map the guidance embeddings of each modality to the channel-wise attention vectors (See Fig. 4(b)). Therefore, the geometry and appearance node embeddings can incorporate important guidance information from previous graph $\mathcal{G}_{(w-1)}$ during performing the joint reasoning over $\mathcal{G}_w$ to create more powerful embeddings.

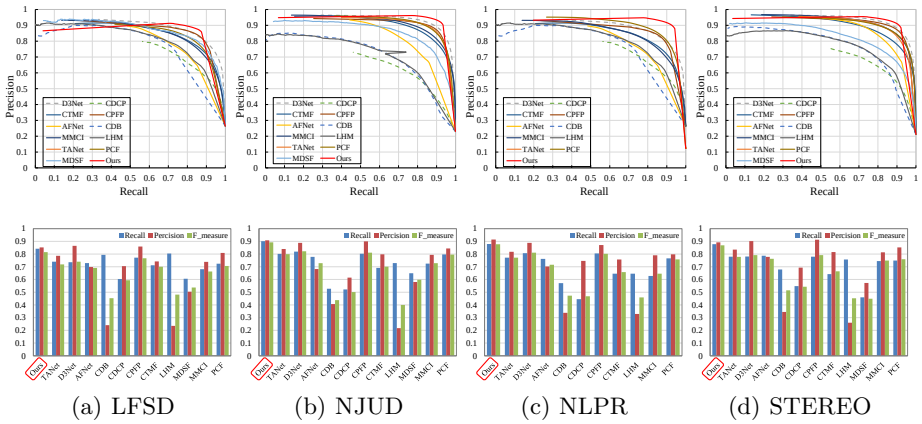
3) Multi-level Feature Fusion: Through the cascade message propagation, the *Cascade Graph Reasoning* (CGR) learns the embeddings of multi-level features under the guidance information provided by the guidance nodes. Here, we denote these learned multi-level embeddings as $\{\breve{\mathbf{C}}_1, \cdots, \breve{\mathbf{C}}_W\}$ and $\{\breve{\mathbf{D}}_1, \cdots, \breve{\mathbf{D}}_W\}$. To fuse them, we rewrite Eq. 7 to create the representations:

$$\breve{\mathbf{C}} = \mathcal{M}_{cl}(R(\breve{\mathbf{C}}_1), \cdots, R(\breve{\mathbf{C}}_W)); \quad \breve{\mathbf{D}} = \mathcal{M}_{cl}(R(\breve{\mathbf{D}}_1), \cdots, R(\breve{\mathbf{D}}_W)); \tag{11}$$

where $\breve{\mathbf{C}}$ and $\breve{\mathbf{D}}$ denote the merged representations for the appearance and geometry domain, respectively. Finally, the saliency readout operation (Eq. 6) is used to produce the final saliency map.

4 Experiments

In this section, we first provide the implementation details of our Cas-Gnn. Then, we perform ablation studies to evaluate the effectiveness of each core



(a) LFSD (b) NJUD (c) NLPR (d) STEREO

Fig. 5. Quantitative comparisons. The PR curves (Top) and weighted F-measures (Bottom) of the proposed method and state-of-the-art approaches on four datasets.

component of graph-based model. Finally, Cas-Gnn is compared with several state-of-the-art RGB-D SOD methods on six widely-used datasets.

Datasets: We conduct our experiments on 7 widely-used datasets: NJUD [31], STEREO [46], NLPR [47], LFSD [34], RGBD135 [13], and SSD [90]. For fair comparison, we follow most SOTAs [7,9,25] to randomly select 1,400 samples from the NJU2K dataset and 650 samples from the NLPR dataset for training, and use all remaining images for evaluation.

Evaluation Metrics: We adopt 5 most-widely used evaluation metrics to comprehensively evaluate the performance of our model, including the mean absolute error (`MAE`), the precision-recall curve (`PR Curve`), F-measure (F_β), S-measure (S_α) [18] and E-measure(E_ξ) [19]. Following previous SOTAs [7,9,25], we set β in F_β to 0.3 and α in S_α to 0.5 for fair comparison.

4.1 Implementation Details

Following [21,7,9,25], we utilize two VGG-16 networks as the backbones, where one is used for extracting the 2D appearance (RGB) features and the other for extracting 3D geometric (depth) features. We employ the dilated convolutions to ensure that the last two groups of backbones have the same resolution. In the *Graph-based Reasoning* (GR) module g_χ, three nodes are used in each modality for capturing information of multiple scales, resulting in a graph $\mathcal{G}$ with six nodes in total. $\mathcal{G}$ links all nodes of the same modality. For the nodes of different modalities, the edge only connects those nodes with the same scale. During the construction of the *Cascade Graph Reasoning* (CGR) module, the features from outputs of the second, third and fifth group of each backbone (different resolutions) are used as inputs for performing cascade graph reasoning. Similar to existing approaches [7,9,25], BCE loss is used to train our model.

Table 1. Ablation analysis for different graph-related settings.

Methods	Settings		NJUD		RGBD135	
	N	T	F_β	MAE	F_β	MAE
Cas-Gnn	2	3	0.887	0.039	0.890	0.033
Cas-Gnn	6	3	0.903	0.035	0.906	0.028
Cas-Gnn	10	3	0.905	0.035	0.909	0.028
Cas-Gnn	6	1	0.881	0.038	0.885	0.031
Cas-Gnn	6	3	0.903	0.035	0.906	0.028
Cas-Gnn	6	5	0.907	0.034	0.908	0.028

Table 2. Ablation analysis on three widely-used datasets.

Methods	Param.	FLOPs	NJUD [31]		STEREO [46]		RGBD135 [13]	
			F_β	MAE	F_β	MAE	F_β	MAE
Baseline	40.66M	65.64G	0.801	0.073	0.813	0.071	0.759	0.052
Baseline + IL	40.91M	66.21G	0.838	0.065	0.841	0.064	0.788	0.046
Baseline + NL	40.98M	66.86G	0.851	0.059	0.852	0.060	0.807	0.043
Baseline + GR (ours)	41.27M	68.91G	**0.874**	**0.051**	**0.864**	**0.048**	**0.854**	**0.031**
Baseline + CMFS	41.88M	72.63G	0.820	0.068	0.822	0.067	0.780	0.047
Baseline + HR (ours)	42.03M	73.19G	0.886	0.041	0.871	0.045	0.890	0.033
Baseline + CGR (ours)	42.28M	73.62G	**0.903**	**0.035**	**0.901**	**0.039**	**0.906**	**0.028**

We implement our Cas-Gnn using the Pytorch toolbox. The fully equipped model is trained on a PC with GTX 1080Ti GPU for 40 epochs with the mini-batch size of 8. The input RGB images and depth images are all resized to 256×256. To avoid overfitting, we perform the following data augmentation techniques: random horizontal flip, random rotate and random brightness. We adopt the Adam with a weight decay of 0.0001 to optimize the network parameters. The initial learning rate is set to 0.0001 and the 'poly' policy with the power of 0.9 is used as a mean of adjustment.

4.2 Ablation Analysis

In this section, we perform a series of ablations to evaluate each component in our proposed network.

Conventional Feature Fusion vs. Graph-based Reasoning. To show the effectiveness of graph-based reasoning, we implement a simple baseline model that directly fuses features from the same multi-modality backbones by first performing the concatenate operation and then learning to fuse the learned features for RGB-D based SOD by two 1×1 convolutions. Clearly, our graph-based reasoning approach (GR module) achieves much more reliable and accurate results.

In addition, we further provide two strong baselines to show the superiority of our proposed graph-based reasoning approach. The first one is designed by using the one-shot induced learner (IL) [3,45] to adapt the learned 3D geometric features to 2D appearance space, making the cross-modality features can be better fused for RGB-D based SOD. The second one uses non-local (NL) module [64] to enable 2D appearance feature map to selectively incorporate useful information from 3D geometric features for building powerful representations. As shown in Tab. 2, our GR module significantly outperforms these strong baselines. This is because our GR module is capable of explicitly distilling complementary information from 2D appearance (color) and 3D geometry (depth) features while the existing feature fusion approaches fail to reason out high-level relations between them.

The Effectiveness of Cascade Graph Reasoning. A key design of our Cas-Gnn is the novel *Cascade Graph Reasoning* module (CGR). To verify the effectiveness of CGR, we use the a common multi-level fusion strategy described in [48] (CMFS) for comparison. As shown in Tab. 2, our CGR consistently

outperforms CMFS across all datasets. Moreover, our CGR is also superior to the hierarchical reasoning (HR) approach without the `guidance nodes` which

Table 3. Quantitative comparisons with state-of-the-art methods by S-measure (S_α), F-measure (F_β), E-measure (E_ξ) and MAE (M) on 7 widely-used RGB-D datasets.

		2014-2017					2018-2020									
	Metric	LHM [47]	CDB [38]	CDCP [91]	MDSF [55]	CTMF [25]	AFNet [60]	MMCI [9]	PCF [7]	TANet [8]	CPFP [86]	D^{3}Net [21]	DMRA [48]	UCNet [83]	ASIF [32]	**Ours**
NJUD	$S_\alpha \uparrow$	0.514	0.624	0.669	0.748	0.849	0.772	0.858	0.877	0.878	0.879	0.895	0.886	0.897	0.888	**0.911**
	$F_\beta \uparrow$	0.632	0.648	0.621	0.775	0.845	0.775	0.852	0.872	0.874	0.877	0.889	0.872	0.889	0.900	**0.903**
	$E_\xi \uparrow$	0.724	0.742	0.741	0.838	0.913	0.853	0.915	0.924	0.925	0.926	0.932	0.908	0.903	-	**0.933**
	$M \downarrow$	0.205	0.203	0.180	0.157	0.085	0.100	0.079	0.059	0.060	0.053	0.051	0.051	0.043	0.047	**0.035**
STEREO	$S_\alpha \uparrow$	0.562	0.615	0.713	0.728	0.848	0.825	0.873	0.875	0.871	0.879	0.891	0.886	**0.903**	0.868	0.899
	$F_\beta \uparrow$	0.683	0.717	0.664	0.719	0.831	0.823	0.863	0.860	0.861	0.874	0.881	0.868	0.885	0.893	**0.901**
	$E_\xi \uparrow$	0.771	0.823	0.786	0.809	0.912	0.887	0.927	0.925	0.923	0.925	**0.930**	0.920	0.922	-	**0.930**
	$M \downarrow$	0.172	0.166	0.149	0.176	0.086	0.075	0.068	0.064	0.060	0.051	0.054	0.047	0.040	0.049	**0.039**
RGBD135	$S_\alpha \uparrow$	0.578	0.645	0.709	0.741	0.863	0.770	0.848	0.842	0.858	0.872	0.904	0.901	-	-	**0.905**
	$F_\beta \uparrow$	0.511	0.723	0.631	0.746	0.844	0.728	0.822	0.804	0.827	0.846	0.885	0.857	-	-	**0.906**
	$E_\xi \uparrow$	0.653	0.830	0.811	0.851	0.932	0.881	0.928	0.893	0.910	0.923	0.946	0.945	-	-	**0.947**
	$M \downarrow$	0.114	0.100	0.115	0.122	0.055	0.068	0.065	0.049	0.046	0.038	0.030	0.029	-	-	**0.028**
NLPR	$S_\alpha \uparrow$	0.630	0.629	0.727	0.805	0.860	0.799	0.856	0.874	0.886	0.888	0.906	0.899	0.918	0.884	**0.919**
	$F_\beta \uparrow$	0.622	0.618	0.645	0.793	0.825	0.771	0.815	0.841	0.863	0.867	0.885	0.855	0.890	0.900	**0.904**
	$E_\xi \uparrow$	0.766	0.791	0.820	0.885	0.929	0.879	0.913	0.925	0.941	0.932	0.946	0.942	0.951	-	**0.952**
	$M \downarrow$	0.108	0.114	0.112	0.095	0.056	0.058	0.059	0.044	0.041	0.036	0.034	0.031	**0.025**	0.030	**0.025**
SSD	$S_\alpha \uparrow$	0.566	0.562	0.603	0.673	0.776	0.714	0.813	0.841	0.839	0.807	0.866	0.857	-	-	**0.872**
	$F_\beta \uparrow$	0.568	0.592	0.535	0.703	0.729	0.687	0.781	0.807	0.810	0.766	0.847	0.821	-	-	**0.862**
	$E_\xi \uparrow$	0.717	0.698	0.700	0.779	0.865	0.807	0.882	0.894	0.897	0.852	0.910	0.892	-	-	**0.915**
	$M \downarrow$	0.195	0.196	0.214	0.192	0.099	0.118	0.082	0.062	0.063	0.082	0.058	0.058	-	-	**0.047**
LFSD	$S_\alpha \uparrow$	0.553	0.515	0.712	0.694	0.788	0.738	0.787	0.786	0.801	0.828	0.832	0.847	**0.860**	0.814	0.849
	$F_\beta \uparrow$	0.708	0.677	0.702	0.779	0.787	0.744	0.771	0.775	0.796	0.826	0.819	0.849	0.859	0.858	**0.864**
	$E_\xi \uparrow$	0.763	0.766	0.780	0.819	0.857	0.815	0.839	0.827	0.847	0.863	0.864	**0.899**	0.897	-	0.877
	$M \downarrow$	0.218	0.225	0.172	0.197	0.127	0.133	0.132	0.119	0.111	0.088	0.099	0.075	**0.069**	0.089	0.073
DUT-RGBD	$S_\alpha \uparrow$	0.568	-	0.687	-	0.834	-	0.791	0.801	-	-	-	0.888	-	-	**0.891**
	$F_\beta \uparrow$	0.659	-	0.633	-	0.792	-	0.753	0.760	-	-	-	0.883	-	-	**0.912**
	$E_\xi \uparrow$	0.767	-	0.794	-	0.884	-	0.855	0.858	-	-	-	0.927	-	-	**0.932**
	$M \downarrow$	0.174	-	0.159	-	0.097	-	0.113	0.100	-	-	-	0.048	-	-	**0.042**

is described in Sec.3.3. This indicates that CGR (with the cascade techniques) can better distill and leverage multi-level information than existing strategies.

Node Numbers N. To investigate the impact of node numbers N in the GR module, we report the results of our GR module with different $N = 2 \cdot n$ in Tab. 1. We observe that when more nodes ($n = 1 \mapsto 3$) in each modality are used, the performance of our model improves accordingly. However, when more nodes are included in each modality ($n = 3 \mapsto 5$), the performance improvements are rather limited. This is caused by the redundant information from generated nodes. Therefore, we believe that setting 3 nodes in each modality ($N = 6$) should be a good balance of the speed and accuracy.

Message Passing Iterations T. We also evaluate the impact of message passing iterations T. As can be seen in Tab. 1, when more than three message passing iterations are used for graph reasoning, the model can achieve the best performance. Therefore, we set $T = 3$ in our GR module to guarantee a good speed and performance tradeoff.

4.3 Comparison with SOTAs

Quantitative Comparisons. We compare our Cas-Gnn with 14 SOTA models on 7 widely-used datasets in Tab. 3. In general, our Cas-Gnn consistently

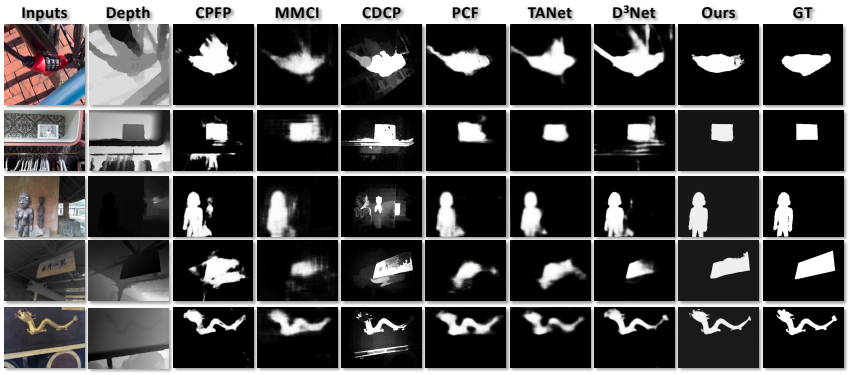


Fig. 6. Qualitative comparisons with state-of-the-art CNNs-based methods.

achieves the remarkable performance on all datasets with four evaluation metrics. Clearly, the results demonstrate that explicitly reason and distill mutual beneficial information can help to infer the salient object regions from the clutter images. In addition, we also show the results of widely-used PR curves and weighted F-measure in Fig. 5. As can be seen, our Cas-Gnn achieves the best performance on all datasets. All the comparisons with recent SOTAs indicate that mining the high-level relations of multi-modality data sources and perform joint reasoning across multiple feature levels are important, and will largely improve the reliability of deep model for handling cross-modality information.

Qualitative Comparisons. Fig. 6 shows some visual samples of results comparing the proposed Cas-Gnn with state-of-the-art methods. We observe that our Cas-Gnn is good at capturing both of the overall salient object regions and local object/region details. This is because our proposed cascade graph reasoning module is able to take both high-level semantics and low-level local details into consideration to build more powerful embeddings for inferring SOD regions.

5 Conclusion

In this paper, we introduce a novel deep model based on graph-based techniques for RGB-D salient object detection. Besides, we further propose to use cascade structure to enhance our GNN model to make it better take advantages of rich, complementary information from multi-level features. According to our experiments, the proposed Cas-Gnn successfully distills useful information from both the 2D (color) appearance and 3D geometry (depth) information, and sets new state-of-the-art records on multiple datasets. We believe the novel designs in this paper is important, and can be used to other cross-modality applications, such as RGB-D based object discover or cross-modality medical image analyse.

Acknowledgement: This research was funded in part by the National Key R&D Progrqam of China (2017YFB1302300) and the NSFC (U1613223).

References

1. Bajaj, M., Wang, L., Sigal, L.: G3raphground: Graph-based language grounding. In: ICCV (2019)
2. Ballas, N., Yao, L., Pal, C., Courville, A.: Delving deeper into convolutional networks for learning video representations (2016)
3. Bertinetto, L., Henriques, J.F., Valmadre, J., Torr, P., Vedaldi, A.: Learning feed-forward one-shot learners. In: NIPS (2016)
4. Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., Andreopoulos, Y.: Graph-based object classification for neuromorphic vision sensing. In: ICCV (2019)
5. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: ICCV (2019)
6. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR (2018)
7. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection. In: CVPR (2018)
8. Chen, H., Li, Y.: Three-stream attention-aware network for rgb-d salient object detection. TIP **28**(6), 2825–2835 (2019)
9. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. Pattern Recognition (2019)
10. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: ECCV (2018)
11. Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., Kalantidis, Y.: Graph-based global reasoning networks. In: CVPR (2019)
12. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. TPAMI (2014)
13. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: Proceedings of international conference on internet multimedia computing and service (2014)
14. Dapogny, A., Bailly, K., Cord, M.: Decafa: Deep convolutional cascade for face alignment in the wild. In: ICCV (2019)
15. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: NIPS (2016)
16. Duvenaud, D.K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., Adams, R.P.: Convolutional networks on graphs for learning molecular fingerprints. In: NIPS (2015)
17. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: ECCV (2018)
18. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: CVPR (2017)
19. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
20. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: CVPR (2020)
21. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks. TNNLS (2020)

22. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: CVPR (2019)
23. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: CVPR (2019)
24. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: ECCV (2014)
25. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE transactions on cybernetics (2017)
26. He, J., Zhang, S., Yang, M., Shan, Y., Huang, T.: Bi-directional cascade network for perceptual edge detection. In: CVPR (2019)
27. He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J.: Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. arXiv preprint arXiv:1911.04231 (2019)
28. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: CVPR (2017)
29. Jiao, J., Wei, Y., Jie, Z., Shi, H., Lau, R.W., Huang, T.S.: Geometry-aware distillation for indoor semantic segmentation. In: CVPR (2019)
30. Jin, B., Ortiz Segovia, M.V., Susstrunk, S.: Webly supervised semantic segmentation. In: CVPR (2017)
31. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: ICIP (2014)
32. Li, C., Cong, R., Kwong, S., Hou, J., Fu, H., Zhu, G., Zhang, D., Huang, Q.: Asif-net: Attention steered interweave fusion network for rgb-d salient object detection. TCYB (2020)
33. Li, G., Muller, M., Thabet, A., Ghanem, B.: Deepgcns: Can gcns go as deep as cnns? In: ICCV (October 2019)
34. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: CVPR (2014)
35. Li, X., Chen, L., Chen, J.: A visual saliency-based method for automatic lung regions extraction in chest radiographs. In: ICCWAMTIP (2017)
36. Li, X., Yang, F., Cheng, H., Chen, J., Guo, Y., Chen, L.: Multi-scale cascade network for salient object detection. In: ACM MM (2017)
37. Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: ECCV (2018)
38. Liang, F., Duan, L., Ma, W., Qiao, Y., Cai, Z., Qing, L.: Stereoscopic saliency model using contrast and depth-guided-background prior. Neurocomputing **275**, 2227–2238 (2018)
39. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR (2019)
40. Liu, N., Han, J.: Dhsnet: Deep hierarchical saliency network for salient object detection. In: CVPR (2016)
41. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. TPAMI (2010)
42. Liu, Y., Zhang, Q., Zhang, D., Han, J.: Employing deep part-object relationships for salient object detection. In: ICCV (2019)
43. Luo, A., Li, X., Yang, F., Jiao, Z., Cheng, H.: Webly-supervised learning for salient object detection. Pattern Recognition (2020)
44. Luo, A., Yang, F., Li, X., Nie, D., Jiao, Z., Zhou, S., Cheng, H.: Hybrid graph neural networks for crowd counting. In: AAAI (2020)
45. Nie, X., Feng, J., Zuo, Y., Yan, S.: Human pose estimation with parsing induced learner. In: CVPR (2018)

46. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: CVPR (2012)
47. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: a benchmark and algorithms. In: ECCV (2014)
48. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
49. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: CVPR (2018)
50. Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R.: 3d graph neural networks for rgbd semantic segmentation. In: ICCV (2017)
51. Ren, J., Gong, X., Yu, L., Zhou, W., Ying Yang, M.: Exploiting global priors for rgb-d saliency detection. In: CVPRW (2015)
52. Ren, Z., Gao, S., Chia, L.T., Tsang, I.W.H.: Region-based saliency detection and its application in object recognition. TCSVT (2013)
53. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. TNN (2008)
54. Shen, Y., Li, H., Yi, S., Chen, D., Wang, X.: Person re-identification with deep similarity-guided graph neural network. In: ECCV (2018)
55. Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., Ren, T.: Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. TIP (2017)
56. Su, J., Li, J., Zhang, Y., Xia, C., Tian, Y.: Selectivity or invariance: Boundary-aware salient object detection. In: ICCV (2019)
57. Wang, A., Wang, M.: Rgb-d salient object detection via minimum barrier distance transform and saliency fusion. SPL (2017)
58. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: CVPR (2019)
59. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: ECCV (2016)
60. Wang, N., Gong, X.: Adaptive fusion for rgb-d salient object detection. IEEE Access **7**, 55277–55284 (2019)
61. Wang, W., Neumann, U.: Depth-aware cnn for rgb-d segmentation. In: ECCV (2018)
62. Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L.: Zero-shot video object segmentation via attentive graph neural networks. In: ICCV (2019)
63. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: CVPR (2018)
64. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
65. Wang, X., Gupta, A.: Videos as space-time region graphs. In: ECCV (2018)
66. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. TOG (2019)
67. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: CVPR (2019)
68. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: ICCV (2019)
69. Xie, G.S., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., Shao, L.: Attentive region embedding network for zero-shot learning. In: CVPR (2019)
70. Xie, G.S., Liu, L., Zhu, F., Zhao, F., Zhang, Z., Qin, J., Yao, Y., Shao, L.: region graph embedding network for zero-shot learning. In: ECCV (2020)

71. Xie, G.S., Zhang, Z., Liu, L., Zhu, F., Zhang, X.Y., Shao, L., Li, X.: Srsc: Selective, robust, and supervised constrained feature representation for image classification. TNNLS (2019)
72. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
73. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? (2019)
74. Xu, Y., Xu, D., Hong, X., Ouyang, W., Ji, R., Xu, M., Zhao, G.: Structured modeling of joint deep feature and prediction refinement for salient object detection. In: ICCV (2019)
75. Yan, P., Li, G., Xie, Y., Li, Z., Wang, C., Chen, T., Lin, L.: Semi-supervised video salient object detection using pseudo-labels. In: ICCV (2019)
76. Yang, F., Li, X., Cheng, H., Guo, Y., Chen, L., Li, J.: Multi-scale bidirectional fcn for object skeleton extraction. In: AAAI (2018)
77. Yang, F., Li, X., Cheng, H., Li, J., Chen, L.: Object-aware dense semantic correspondence. In: CVPR (July 2017)
78. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2015)
79. Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. In: ICCV (2019)
80. Zhang, D., Meng, D., Zhao, L., Han, J.: Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. arXiv preprint arXiv:1703.01290 (2017)
81. Zhang, J., Sclaroff, S.: Saliency detection: A boolean map approach. In: ICCV (2013)
82. Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R.: Minimum barrier salient object detection at 80 fps. In: ICCV (2015)
83. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Sadat Saleh, F., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: CVPR (2020)
84. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: Capsal: Leveraging captioning to boost semantics for salient object detection. In: CVPR (2019)
85. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
86. Zhao, J.X., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L.: Contrast prior and fluid pyramid integration for rgbd salient object detection. In: CVPR (2019)
87. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: ICCV (2019)
88. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: CVPR (2019)
89. Zhu, C., Cai, X., Huang, K., Li, T.H., Li, G.: Pdnet: Prior-model guided depth-enhanced network for salient object detection. In: ICME (2019)
90. Zhu, C., Li, G.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: CVPRW (2017)
91. Zhu, C., Li, G., Wang, W., Wang, R.: An innovative salient object detection using center-dark channel prior. In: ICCVW. pp. 1509–1515 (2017)