Supported by the Excellent Doctoral Degree Dissertation Fund of Tianjin University

Research on Visual Saliency Detection with Comprehensive Information

Discipline：Information and Communication Engineering

Major：Information and Communication Engineering

Author：Runmin Cong

Supervisor：Prof. Qingming Huang

Prof. Jianjun Lei

School of Electrical and Information Engineering

Tianjin University

May 2019

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得**天津大学**或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解**天津大学**有关保留、使用学位论文的规定。特授权**天津大学**可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

ABSTRACT

The Human visual system works as a filter to allocate more attention to the attractive and interesting objects for further processing. Visual saliency detection model simulates this system to perceive the scene, and has been widely used in many vision tasks, such as segmentation, retrieval, compression, coding, quality assessment, and so on. With the development of acquisition technology, more comprehensive information, such as depth cue, inter-image correspondence, or spatiotemporal relationship, is available to extend single image saliency detection to RGBD saliency detection, co-saliency detection, or video saliency detection. RGBD saliency detection model focuses on extracting the salient objects from RGBD images by combining the color and depth information. Co-saliency detection model introduces the inter-image correspondence constraint to discover the common salient objects in an image group. The goal of video saliency detection is to locate the motion-related salient object in video sequences, which considers the motion cue and spatiotemporal constraint jointly. In this thesis, comprehensive information is explicitly explore to address the challenges in different saliency detection tasks. Specially, the main works of this thesis are summarized as follows:

(1) For the stereoscopic images, considering the quality of the depth map and multiple cues fusion, a novel saliency detection method is proposed. First, according to the observation of depth distribution, a confidence measure for depth map is designed to reduce the negative influence of poor depth map on saliency detection. Moreover, a novel stereoscopic compactness saliency model is defined by integrating the color and depth information. In addition, a depth-refined foreground seeds selection mechanism is presented to assist in foreground saliency calculation by integrating color, depth, and texture cues. At last, the complementary compactness saliency and foreground saliency are fused to generate the final saliency map.

(2) In order to fully exploit the depth and inter-image correspondences, this thesis first attempts to address the co-saliency detection from an RGBD image group, in which the depth information is introduced as a novel cue in the designed model. In order to explore the inter-image relationship, the similarity matching methods on two levels are

proposed. The first one is the superpixel-level similarity matching scheme, which focuses on determining the matching superpixel set for the current superpixel based on three constraints from other images. The second is the image-level similarity measurement, which provides a global relationship on the whole image scale and works as a weighted coefficient for inter saliency calculation. Finally, the cross label propagation method is proposed to optimize the intra and inter saliency maps in a cross way, and generate the final co-saliency map.

(3) The existing co-saliency detection methods mainly rely on the designed cues or initialization, and lack the refinement-cycle. Thus, an effective co-saliency framework for RGBD images based on the refinement-cycle model is proposed, which integrates the addition scheme, deletion scheme, and iteration scheme. The addition scheme is used to enrich the saliency regions through the depth propagation and saliency propagation. Note that, a novel depth descriptor, named depth shape prior, is proposed in depth propagation to introduce the depth information and enhance the identification of co-salient objects. In the deletion scheme, the inter saliency is formalized as a common probability function to capture the inter-image correspondence. The iterative optimization scheme is designed to achieve more superior co-saliency result in a cycle way. The proposed method effectively exploits any existing 2D saliency model to work well in RGBD co-saliency scenarios.

(4) In order to balance the effectiveness and efficiency for inter-image correspondence capturing in co-saliency detection, a novel RGBD co-saliency model is proposed based on hierarchical sparsity reconstruction and energy function refinement. The multi-image correspondence is formulated as a hierarchical sparsity reconstruction framework, where the global sparsity reconstruction captures the global characteristic among the whole image group through a common foreground dictionary, and the pairwise sparsity reconstruction model utilizes a set of foreground dictionaries produced by other images to explore local inter-image information. Finally, in order to improve the intra-image smoothness and inter-image consistency, an energy function refinement model is proposed, which includes the unary data term, spatial smooth term, and holistic consistency term.

(5) Combining the spatial saliency in the single frame, the temporal cue in the inter frames, and the global constraints among the whole video, a novel method to detect the salient objects in video is proposed based on sparse reconstruction and propagation.

The single-frame saliency is calculated to represent the spatial saliency in each individual frame via the sparsity-based reconstruction, where the motion priors are defined as the motion compactness and uniqueness cues. Then, an efficient sparsity-based saliency propagation is presented to capture the correspondence in the temporal space and produce the inter-frame saliency map. Specifically, the salient object is sequentially reconstructed by the forward and backward dictionaries. Finally, in order to attain the spatiotemporal smoothness and global consistency of the salient object in the whole video, a global optimization model is formulated, which integrates unary data term, spatiotemporal smooth term, spatial incompatibility term, and global consistency term.

KEY WORDS: Visual saliency detection, co-saliency detection, video saliency detection, comprehensive information, RGBD images, depth cue, inter-image correspondence, spatiotemporal constraint.

中 文 摘 要

人类视觉系统可在大范围、复杂场景中定位出最吸引注意的感兴趣内容或区域，称之为视觉注意机制。该机制可以帮助人类快速捕获场景中的有效信息，以便快速、有效的分析场景内容。受此机制的启发，研究人员希望计算机可以模拟人类的视觉注意机制，具备自动定位场景中显著性内容的能力，进而为后续处理提供有效的辅助信息，实现计算资源的合理分配，这样视觉显著性检测任务应运而生。场景的显著性区域通常包含了人类感兴趣的重要目标或最能表达图像的内容，是能够在较短时间内吸引人的视觉注意力的区域，而视觉显著性检测就是找出这些感兴趣目标或区域的过程。作为一个跨计算机科学、神经学、生物学、心理学的交叉学科方向，视觉显著性检测已经被广泛应用于诸多研究领域，如检测、分割、裁剪、检索、压缩编码、质量评价、推荐系统等，具有十分广阔的应用前景。根据处理对象的不同，视觉显著性检测可以进一步划分为图像显著性检测、协同显著性检测和视频显著性检测等不同分支任务。本论文以不同的辅助信息为引导，探讨这三种显著性检测任务之间的区别与联系，并针对现有方法中存在的问题，提出相应的解决方案，以期促进相关领域的发展。具体研究内容如下：

(一)图像显著性检测。经过十余年的发展，面向彩色图像的单图显著性检测方法已经形成了较为完善的方法体系，新算法层出不穷，性能也不断被刷新，特别是深度学习方法的兴起使得算法性能发生了质的飞跃。具体来说，图像显著性检测方法可以粗略地分为两大类：一类是由数据驱动的自底向上的检测方法，这类方法主要利用底层线索（如颜色、纹理等）直接进行显著性模型构建；另一类是由任务驱动的自顶向下的检测方法，该类方法往往需要训练过程和特定的先验知识。实际上，人眼是通过双目立体视觉方式来感知客观世界的。换言之，人眼的双目视觉系统不仅可以获取场景的 2D 平面信息（如颜色、形状、结构等），还可以感知场景的深度信息，获得立体感。随着成像技术的进步与发展，成像设备不断更新换代，人们可以更加方便快捷地获取场景的深度信息，即以深度图的形式表示和存储场景的深度关系。作为颜色信息的补充，深度图可以提供许多有效的辅助信息，诸如形状、边缘、内部一致性等，进而进一步增强检测、识别等任务的效果。相比于 RGB 图像显著性检测任务，面向 RGBD 图像的显著性检测研究相对较少，研究人员在深度信息对人类感知系统的作用机理、深度信息的有效利用方法等方面还未达成共识，仍需进一步深入研究。

本论文提出了一种基于深度置信测度和多线索融合的 RGBD 图像显著性检

测算法。众所周知，高质量的深度图可以为显著性检测提供准确、有效的辅助信息，而低质的深度图如同噪声一样，引入到显著性模型后反而可能会降低检测性能。然而，现有的 RGBD 图像显著性检测算法并未对深度图质量进行有效评测和区分。基于此，通过观察深度图分布，利用深度图均值、变异系数和深度频率熵参量，构建了描述深度图可靠性的深度置信测度，以此控制模型引入深度信息的程度。测度数值越大，说明深度图质量越高，越可靠；反之，说明深度图质量较差，应该尽量降低深度信息引入的比例。此外，观察发现，显著性目标的深度在一定范围内（通常为靠近图像中心的区域）分布较为集中，而背景的深度分布比较分散，此特性类似于颜色域中的紧致性先验。因此，将颜色紧致性先验拓展至深度域，提出了一种融合颜色和深度信息的紧致性计算模型，得到紧致显著性图。为了获得更加鲁棒、稳定的检测性能，提出了一种基于深度修正的前景种子点选择机制，并综合考虑颜色、深度、纹理等特征，利用局部对比方法计算得到前景显著性图。最后，将紧致显著性图和前景显著性图进行加权融合，得到最终的 RGBD 显著性检测结果。在两个立体显著性检测数据库上的实验证明，所提出的 RGBD 显著性模型在定性和定量分析上都获得了较好的检测性能。该部分对应本论文中第 3 章内容。

(二)协同显著性检测。近年来，图像数据量急剧增长，人们处理的对象不再局限于单幅图像，而是需要联合多图信息同时处理一组具有近似目标或事件的图像集合数据。协同显著性检测作为一种新兴的、更具挑战性的任务逐渐引起了研究者的关注。与传统的单图显著性检测模型不同，协同显著性检测模型旨在从包含两个或多个相关图像的图像组中发现共同的显著性物体，而这些目标的类别、内在特征和位置往往都是未知的。由于其优越的可扩展性，协同显著性检测方法也已被广泛应用于协同分割、近似目标检测、目标协同识别、图像检索以及图像简报生成中。根据协同显著性检测的定义，协同显著性目标需同时满足两个特征属性：(i) 显著性，即每幅图像中的协同显著性目标应该是显著的；(ii) 共有性，即协同显著性目标应该是整个图像组中共有的目标，且具有近似的外观特征。因此，在协同显著性检测任务中，图像之间的对应关系用于判决单幅图像中的显著性目标是否是整个图像组中所共有的显著性目标。换句话说，有效捕获图像组中的图间对应关系是协同显著性检测任务中必不可少的环节和过程。现有的方法通常将图间关系建模为聚类问题、匹配问题、传播问题或学习问题等。建模为聚类问题的方法虽然具有较好的算法时效性，但其性能容易受噪声影响，算法精度有限。相比之下，基于相似匹配和传播的图间关系获取方法较为常用，其准确性较高，但需要以牺牲时间复杂度来换取。最近，基于学习的方法取得了更好的性能，

但其往往需要大量带标签的样本数据进行模型训练，数据准备阶段成本较高，未来还需进一步研究。同样地，深度信息也可以引入到协同显著性检测任务中，进一步提高检测性能。这样，面向 RGBD 图像的协同显著性检测任务应用而生。除了考虑图间对应关系外，如何有效挖掘深度信息也是需要着重解决的问题。围绕 RGBD 协同显著性检测任务，本论文开展了三方面的研究工作：

(1) 提出了一种基于多约束特征匹配和交叉标签传播的 RGBD 协同显著性检测模型，将深度信息看作颜色信息的补充特征，利用多约束特征匹配方法获取图间约束关系计算图间显著性图，并利用基于交叉标签传播的优化机制对图内和图间显著性图进行交叉优化。具体来说，首先利用现有的立体显著性检测模型计算图像组内每幅图像的图内显著性图。然后，分别在两个层级上进行特征匹配获取图间约束关系，生成图间显著性图。第一个层级是超像素级的相似性匹配，根据颜色、深度和显著性三种约束确定当前超像素在同组其他图像中的匹配超像素集合；基于图像组中相似性较高的两个图像之间存在共有显著性目标的可能性更大的观察，设计了第二个图间关系捕获层级，即图像级的图间相似性度量，用于提供整个图像尺度上的全局图间关系。根据捕获的超像素级和图像级的图间对应关系，超像素的图间显著性被定义为其他图像中对应超像素显著性值的加权和，其中加权系数由图像级的相似性度量计算得到。紧接着，为了获得更加均匀、一致的检测结果，将协同显著性检测的优化问题建模为一个“标签传播”问题，设计了基于交叉标签传播的优化机制对图内和图间显著性图进行交叉融合和优化，即首先利用图内显著性图对图间显著性图进行优化，然后利用优化后的图间显著性图对图内显著性图进行优化。最后，将原始显著性图与优化后的显著性图进行融合得到最终的协同显著性结果。在该工作中，还构建了一个具有真图标定、涉及 21 个图像组、包含 150 张图像的 RGBD 协同显著性检测数据集，且已对外开放下载。在两个 RGBD 协同显著性检测数据库上的实验证明了该方法的有效性。该部分对应本论文中第 4 章内容。

(2) 提出了一种迭代的 RGBD 协同显著性检测框架，其中补机制基于图内的深度和显著性传播来突出显著性区域，删机制通过提取的图间约束关系来抑制非共有的显著性区域，迭代机制以循环的方法获得更加均匀、一致的协同显著性结果。对于 RGBD 协同显著性检测任务，应该重点解决两个问题：一是如何充分挖掘深度信息，二是如何捕获图间关系对显著性目标进行筛选。鉴于单图 RGB 显著性检测的快速发展和优异性能，完全可以直接利用现有的单图 RGB 显著性检测方法生成图内显著性检测结果，而不需要设计一个包含各个模块的完整框架（如图内显著性检测模块、图间显著性检测模块和优化模块）来实现协同显著性

检测任务，进而将重点放在深度信息和图间约束关系的获取上。基于此动机，该工作以现有的 2D 单图显著性结果作为随机初始化，利用一个循环修正模型实现了 RGBD 图像的协同显著性检测，即实现了从 RGB 显著性检测任务到 RGBD 协同显著性检测任务的转换。该框架由三个机制组成：补机制通过深度传播将深度信息引入模型中，并利用显著性传播进一步改善图内显著性结果。观察深度图发现，显著性目标内部的深度值较大且趋于平滑、一致，而且高质量的深度图还可以提供锐利且清晰的目标边界和形状信息。基于上述观察，提出了一种新的深度描述子——深度形状先验，用以捕获深度图的形状信息，进而增强显著性检测性能。删机制通过提取的图间约束关系来抑制非共有的显著性区域和背景区域。具体来说，在该机制中，构造了超像素级的相似性度量来表示两个超像素之间的相似性关系，并设计了一个共有概率函数来度量每个超像素属于共有区域的可能性，生成图间显著性结果。迭代机制以循环的方法将整个模型串联起来，以获得更加均匀、一致的协同显著性结果。迭代过程通过判断最大迭代次数和两次迭代结果的变化程度来决定是否终止。通过在两个 RGBD 协同显著性检测数据库上的定性和定量实验证明了该方法的有效性。该部分对应本论文中第 5 章内容。

(3) 提出了一种基于分层稀疏重建的 RGBD 协同显著性检测方法，其中全局重建通过一个共有前景字典来捕获整个图像组的全局特性，成对重建通过多组成对字典挖掘图像对之间的对应关系，能量函数优化用于改善图内平滑性和图间一致性。基于相似性匹配算法可以获得较为准确的图间对应关系，但其运算量较大，计算复杂度较高。而基于聚类的图间关系建模方法对噪声比较敏感，以准确性换取了时效性，使得算法性能大打折扣。因此，现有算法很难同时兼顾有效性和时效性。针对这一问题，稀疏表示技术提供了很好的解决方案，已被广泛应用于包括显著性检测在内的多项任务中。传统的基于稀疏表示的显著性检测方法通常利用背景或前景字典来重建每个处理单元（如超像素），并且通过计算重建误差来度量显著性值。实际上，除了用于描述单幅图像的显著性之外，稀疏表示还可用于捕获图像间对应关系，实现图间显著性检测。通过考虑全局和局部图像间信息，提出了一种分层稀疏重建模型以捕获更加全面的图间关系，主要包括两个互补机制：一方面，整个图像组中的共有显著对象应属于同一类别并具有相似的外观特征。因此，利用图像组内的所有图像信息构建一个全局前景字典，并以此对图像组内的每幅图像进行稀疏重建，捕获全局的图间对应关系，这被称为全局稀疏重建。另一方面，多个图像之间的关系可以分解为多个图像对之间对应关系的组合。因此，每幅图像可以被图像组中其他图像构造的前景字典进行重建，进而从局部视角获得多对图间显著性图，融合后得到成对的图间显著性结果，该过程称为成

对稀疏重建。此外，同一图像组中的共有显著性目标应该在外观上具有较高的相似性和一致性。因此，协同显著性检测模型应该保证检测结果具有较好的图内平滑性和图间一致性。为此，提出了一种能量函数修正模型以获得更加一致、准确的协同显著性结果，包括一元数据项、空间平滑项和全局一致项。数据项用于限制修正算法的更新变化程度，平滑项用于约束具有相似外观的空间相邻区域具有较为一致的显著性得分，全局一致项则专为协同显著性检测任务设计，使得共有显著性目标在整个图像组中保持一致。在两个 RGBD 协同显著性检测数据库上的定性和定量分析实验表明，该方法优于目前最先进的算法。该部分对应本论文中第 6 章内容。

(三)视频显著性检测。大数据时代的来临，使得数据形式发生了翻天覆地的变化，传统的图像数据已不足以满足人们日益增长的感官需求，视频数据量呈现出井喷式的增长，如何准确、连续地提取视频数据中的显著性目标成为亟待解决的新课题。视频显著性检测旨在通过联合空间和时间信息，实现视频序列中与运动相关的显著性目标的连续提取，已被广泛应用于视频目标检测、视频摘要、基于内容的视频检索等领域。不同于图像显著性检测，视频显著性检测需要同时结合时间信息和空间信息，连续地定位视频序列中与运动相关的显著性目标。与协同显著性检测相比，视频显著性检测还需考虑运动信息和时序特性，而且具有“相邻视频帧之间相关性较大”的先验。由于视频数据量大、场景变化明显、目标大小不一致、存在遮挡等问题，使得视频显著性检测研究难度较大，算法性能有待进一步提高。根据是否需要进行训练学习，视频显著性方法可以分为基于底层线索的方法和基于学习的方法两大类。其中，基于底层线索的视频显著性检测方法可以进一步划分为基于变换分析的方法、基于信息论的方法、基于稀疏表示的方法、基于视觉先验的方法和其他方法五类，而基于学习的方法可以分为传统学习方法和深度学习方法两类。

本论文提出了一种基于稀疏重建与传播的无监督视频显著性检测算法，利用运动紧致性和运动独特性挖掘运动信息，采用基于稀疏的显著性双向传播方案捕获帧间对应关系，利用全局优化改善整个视频中显著性对象的全局一致性。在视频数据中，运动物体通常比静态物体更容易受到关注。但是，并非所有的运动目标都是显著的，还需要通过分析周围区域和相邻帧进行进一步区分。因此，如何充分利用运动信息来突出显著性目标并抑制背景对视频显著性检测至关重要。目前，常用的运动特征有光流对比度、光流梯度等。然而，这些方法容易受噪音和背景运动影响，稳健性较差。众所周知，稀疏表示是一种判别性较好的数据表达方式，且对噪声具有较好的鲁棒性，已经被用于改善各种推理任务性能，例如对

象跟踪、人脸识别、形状估计等。而且，也有许多显著性检测算法基于稀疏表示进行计算，但多数仅用于处理单张图像数据或视频的单帧数据。在该工作中，稀疏表示技术不仅用于计算单帧显著性，还用于进行具有前向-后向传播策略的帧间显著性计算。首先，在静态线索和运动先验的帮助下，通过稀疏重建模型计算视频序列中每个视频帧的显著性。具体来说，从光流数据特性入手，将颜色空间的紧致性和独特性先验引入运动域，提出了运动紧致性和运动独特性概念，其中运动紧致性描述了光流信息的分布，而运动独特性描述了运动幅度信息的外貌特征。利用这些运动线索，可以有效改善运动显著性测量的精度。实际上，视频序列中相邻帧之间的相似性较高，因此稀疏重建方法也可以用于描述帧与帧之间的对应关系，进而定义帧间显著性图。基于此，设计了一种渐进式的稀疏传播框架，采用前向-后向策略来建模帧间对应关系，并生成空时域上的帧间显著性图。在前向传播中，前一帧视频数据用于构建前向字典，并对视频当前帧进行重建，捕获前向帧间关系。相反，后向传播则从最后一帧逐渐传播至视频第一帧，并利用后一帧视频构成的后向字典对当前帧进行重建，挖掘后向帧间关系。此外，在视频显著性检测中，应考虑时空一致性以获得更加平滑、一致的结果，即要求显著性区域或背景的显著性值不应沿时间轴急剧变化。然而，在大多数现有方法中，输入视频是逐帧处理的，这样会忽略了整个视频序列的全局特性，进而使得显著性结果只能保证整个视频的局部一致性而非全局一致性。因此，提出了一种基于能量函数的全局优化方案，主要包括一元数据项、时空平滑项、空间互斥项和全局一致性项。在三个大规模视频数据集上对该方法进了实验验证，结果表明所提方法获得了最优的定性和定量结果。该部分对应本论文中第7章内容。

CONTENT

Chapter 1 Introduction

1.1 Background and Overview

Human visual system works as a filter to allocate more attention to the attractive and interesting regions or objects for further processing. Humans can exhibit visual fixation, which is maintaining of the visual gaze on a single location. Inspired by this visual perception phenomenon, some visual saliency models focus on predicting human fixations [1]. In addition, driven by computer vision applications, some visual saliency models aim at identifying the salient regions from the image or video [2], which has been applied in image/video segmentation [3,4], image/video retrieval [5,6], image retargeting [7,8], image compression [9], image enhancement [10-12], video coding [13], foreground annotation [14], quality assessment [15,16], thumbnail creation [17], action recognition [18], and video summarization [19]. This thesis mainly focuses on salient object detection task.

The last decade has witnessed the remarkable progress of image saliency detection, and a plenty of methods have been proposed and achieved the superior performances, especially the deep learning based methods have yielded a qualitative leap in performances. In fact, the human visual system can not only perceive the appearance of the object, but also be affected by the depth information from the scene. With the development of imaging devices, the depth map can be acquired conveniently and accurately, which lays the data foundation for RGBD saliency detection [20]. Generally, there are three options for 3D depth imaging, *i.e.*, structured light [21], TOF (Time-of-Flight) [22], and binocular imaging [23]. The structured light pattern (*e.g.*, Kinect) captures the depth information via the change of light signal projected by the camera, which can obtain high-resolution depth map. The TOF system (*e.g.*, Camcube) estimates the depth through the round-trip time of the light pulses, which has good anti-jamming performance and wider viewing angle. The stereo imaging system takes photo pair via stereo camera and calculates the object's disparity based on two-view geometry. Depth map can provide many useful attributes for foreground extraction from the

complex background, such as shape, contour, and surface normal. Some examples of saliency detection with and without depth cue are shown in Fig. 1-1. As can be seen, utilizing the depth cue, RGBD saliency model achieves superior performance with consistent foreground enhancement. However, how to effectively exploit the depth information to enhance the identification of salient object has not yet reached a consensus, and still needs to be further investigated. Considering the ways of using depth information, the RGBD saliency detection model can be divided into depth feature based method and depth measure based method. Depth feature based method focuses on taking the depth information as a supplement to color feature, and depth measure based method aims at capturing comprehensive attributes from the depth map (*e.g.*, shape) through the designed depth measurements. More details will be discussed in Chapter 2.

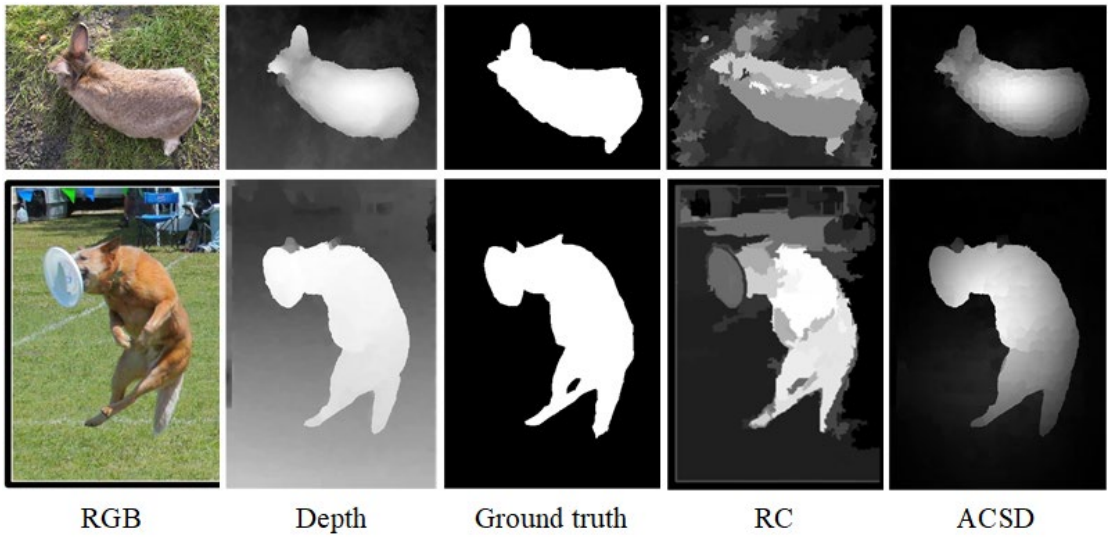

Fig. 1-1 Some illustrations of saliency detection with and without depth cue. The first three columns correspond to the RGB image, depth map, and ground truth, respectively. The fourth column shows the image saliency detection result using the RC method [25]. The fifth column represents the RGBD saliency detection result using the ACSD method [65].

In recent years, with the explosive growth of data volume, human need to process multiple relevant images collaboratively. As an emerging and challenging issue, co-saliency detection gains more and more attention from researchers, which aims at detecting the common and salient regions from an image group containing multiple related images, while the categories, intrinsic attributes, and locations are entirely unknown [24]. In general, three properties should be satisfied by the co-salient object, *i.e.*, (1) the object should be salient in each individual image, (2) the object should be

repeated in most of the images, and (3) the object should be similar in appearance among multiple images. Some visual examples of co-saliency detection are provided in Fig. 1-2. In the individual image, all the cows should be detected as the salient objects. However, only the brown cow is the common object from the image group. Therefore, the inter-image correspondence among multiple images plays a useful role in representing the common attribute. On the whole, co-saliency detection methods are roughly grouped into two categories according to whether the depth cue is introduced, *i.e.*, RGB co-saliency detection and RGBD co-saliency detection. Further, the RGB co-saliency detection methods can be divided into some sub-classes based on different correspondence capturing strategies, *i.e.*, matching based method, clustering based method, rank analysis based method, propagation based method, and learning based method.

Fig. 1-2 Examples of the co-saliency detection model. The first row presents the input images, and the second row shows the co-salient object (brown cow) in this group.

Different from image data, video sequences contain more abundant appearance information and continuous motion cue, which can better represent the characteristics of the target in a dynamic way. However, the clustered backgrounds, complex motion patterns, and changed views also bring new challenges to interpret video content effectively. Video saliency detection aims at continuously locating the motion-related salient object from the given video sequences by considering the spatial and temporal information jointly. The spatial information represents the intra-frame saliency in the individual frame, while the temporal information provides the inter-frame constraints and motion cues. Fig. 1-3 illustrates some examples of video saliency detection. In this camel video, both two camels appeared from 40th frame should be detected as the salient objects through a single image saliency model. However, only the front one is continuously moving and repeating, which is the salient object in this video. The

differences between co-saliency detection and video saliency detection lie in two aspects, *i.e.*, (1) The inter-frame correspondence has the temporal property in video saliency detection rather than in co-saliency detection. For co-saliency detection in an image group, the common salient objects have the consistent semantic category, but are not necessarily the same object. By contrast, the salient objects in video are continuous in the time axis and consistent among different frames; (2) The motion cue is essential to distinguish the salient object from the complex scene in video saliency detection model. However, this cue is not included in co-saliency detection model. Similar to the classification strategy of image saliency detection, the video saliency detection methods are divided into two categories, *i.e.*, low-level cue based method and learning based method. For clarity, the low-level cue based method is further grouped into fusion model and direct-pipeline model according to feature extraction method, and the learning based method is further divided into supervised method and unsupervised method.

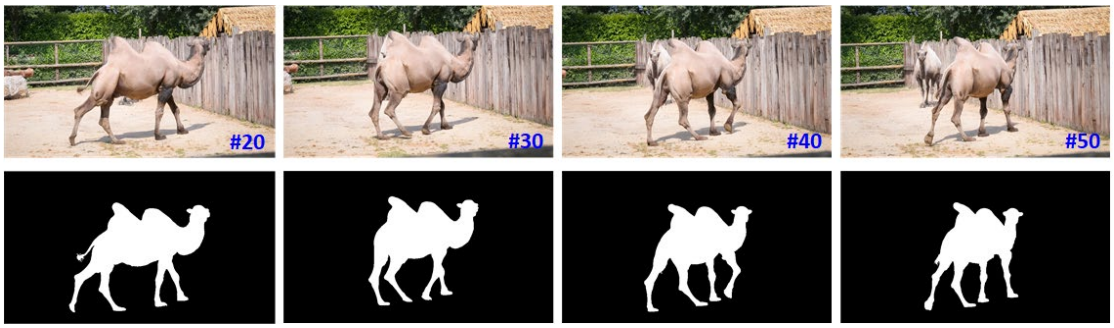

Fig. 1-3 Examples of the video saliency detection model. The first row is the input video frames, and the second row shows the salient object in this video, *i.e.*, the front camel.

As stated above, the major relationships among four different visual saliency detection models are summarized in Fig. 1-4, where the image saliency detection model is the basis for other three models. With the depth cue, RGBD saliency map can be obtained from an image saliency detection model. Introducing the inter-image correspondence, image saliency detection model can be transformed into a co-saliency detection method. Video saliency detection can be derived from an image saliency detection model by combining the temporal correspondence and motion cue, or from a co-saliency detection method by integrating the motion cue. In practice, in order to obtain superior performance, it is necessary to design a specialized algorithm to achieve co-saliency detection or video saliency detection, rather than directly transplanting the image saliency detection algorithms.

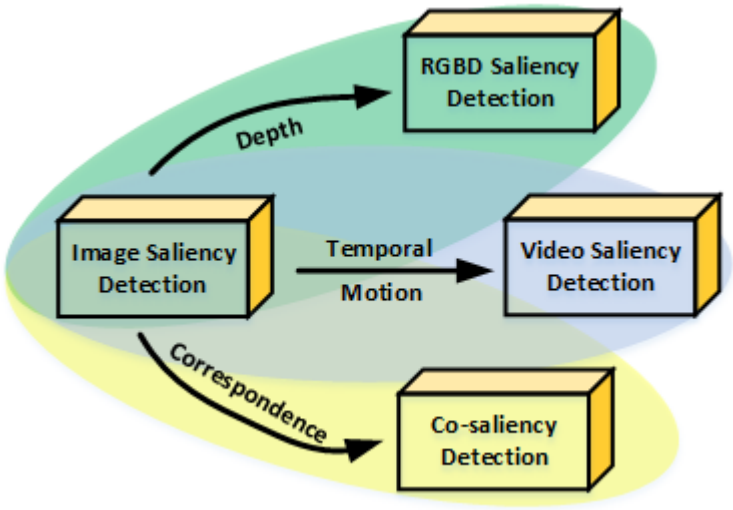

Fig. 1-4 Relationships between different visual saliency detection models.

1.2 Research Contents and Contributions

In this thesis, we explicitly address the challenges of visual saliency detection with comprehensive information including depth cue, inter-image correspondence, and temporal constraint. With these information, we conduct the research on RGBD saliency detection, co-saliency detection, and video saliency detection. The main contributions of this thesis are summarized as follows:

(1) A novel saliency detection method for stereoscopic images is presented based on depth confidence analysis and multiple cues fusion. According to the observation of depth distribution, a depth confidence measure is introduced into the model to reduce the negative influence of poor depth map. Moreover, the compactness saliency is computed by integrating the color and depth information, and the foreground saliency is calculated by using the multiple cue contrast with depth-refined foreground seeds selection mechanism. Finally, these two parts are fused to generate the final saliency result.

(2) An effective model is proposed to detect the co-salient objects from the RGBD images, where the depth information is demonstrated to be served as a useful complement for co-saliency detection. In this work, the similarity matching models at superipixel and image levels are designed to constrain the inter saliency map generation, which is robust to the complex backgrounds. The Cross Label Propagation (CLP) scheme is proposed to optimize the co-saliency model in a cross way. Moreover, a new

dataset for RGBD co-saliency detection with the corresponding pixel-level ground truth is constructed.

(3) An iterative framework for RGBD co-saliency detection is provided, which utilizes the existing single saliency maps as the initialization, and generates the final RGBD co-saliency map by using a refinement-cycle model. In this work, a novel depth descriptor, named depth shape prior, is proposed to capture the shape attributes from the depth map and enhance the identification of co-salient objects from RGBD images. In addition, a superpixel-level common probability function among multiple images is calculated to exploit the inter-image corresponding relationship in the deletion scheme, and the iterative updating strategy is presented to obtain more homogeneous and consistent co-saliency result in the iteration scheme.

(4) A co-saliency detection method for RGBD images is designed based on hierarchical sparsity reconstruction and energy function refinement. In this work, the corresponding relationship among multiple images is simulated as a hierarchical sparsity framework, where the global inter saliency reconstruction model describes the inter-image correspondence from the perspective of the whole image group via a common reconstruction dictionary, and the pairwise inter saliency reconstruction model utilizes a set of foreground dictionaries produced by other images to capture local inter-image information. Moreover, an energy function refinement model is proposed to improve the intra-image smoothness and inter-image consistency.

(5) An effective method to detect the salient objects in video is demonstrated based on sparse reconstruction and propagation. The sparsity-based saliency reconstruction model is designed to generate single-frame saliency map, making the best use of the static and motion priors. The sparsity-based saliency propagation with the forward-backward strategy is presented to capture the correspondence in the temporal space and produce inter-frame saliency map. In order to attain the global consistency of the salient object in the whole video, a global optimization model including the unary data term, spatiotemporal smooth term, spatial incompatibility term, and global consistency term, is formulated.

1.3 Chapter Arrangement

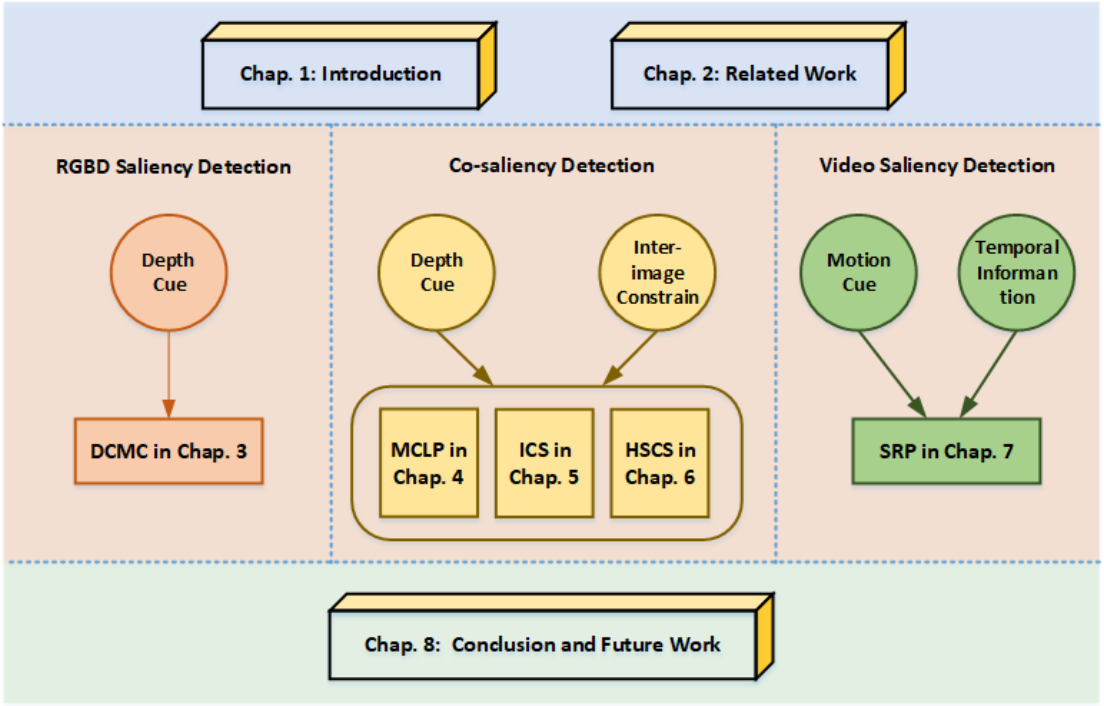

Fig. 1-5 Overview of the organization of this thesis.

This thesis is organised into eight chapters. As shown in Fig. 1-5, the main topics in each chapter are briefly summarized as follows:

Chapter 1 introduces the background and overview of visual saliency detection, and summarizes the main research contents and contributions of this thesis.

Chapter 2 reviews different types of saliency detection algorithms including saliency detection, co-saliency detection, and video saliency detection. Moreover, this chapter also introduces the benchmark datasets and evaluation metrics, which lays the foundation for subsequent researches.

Chapter 3 gives the details of the stereoscopic saliency detection method based on depth confidence analysis and multiple cues fusion. First, the framework of the proposed method is introduced, including depth confidence measure, graph construction, compactness saliency using color and depth cues, foreground saliency using multiple cues contrast, and saliency map generation. Then, experiments on two publicly available stereo datasets demonstrate that the proposed method performs better than other state-of-the-art approaches. Finally, a summary is made at the end of this chapter.

Chapter 4 demonstrates a co-saliency detection method for RGBD images based on multi-constraint feature matching and cross label propagation. The technical details of the framework are provided, which includes the intra saliency detection, inter saliency detection, and optimization. The comprehensive experiments on two RGBD co-saliency datasets, including the comparison with state-of-the-art methods and ablation studies, demonstrate the effectiveness of the proposed model. Finally, the method is summarized at the end of this chapter.

Chapter 5 presents an iterative co-saliency framework for RGBD images. Three main parts of the proposed framework are introduced, including the addition scheme, deletion scheme, and iteration scheme. The comprehensive comparisons and discussions on two RGBD co-saliency datasets demonstrate that the proposed method outperforms other state-of-the-art saliency and co-saliency models. Finally, a summary is made at the end of this chapter.

Chapter 6 provides the detailed explanations on the hierarchical sparsity based co-saliency detection framework for RGBD images. First, the intra saliency calculation, hierarchical inter saliency detection based on global and pairwise sparsity reconstructions, and energy function refinement are detailed. Then, experiments and ablation studies on two RGBD co-saliency benchmarks demonstrate that the proposed method outperforms the state-of-the-art algorithms both qualitatively and quantitatively. Finally, the conclusion is drawn at the end of this chapter.

Chapter 7 describes the video saliency detection method by using sparsity-based reconstruction and propagation. Three progressive steps of the proposed framework are introduced, *i.e.*, single-frame saliency reconstruction, inter-frame saliency propagation, and global optimization. Then, the comprehensive experiments and analyses on three challenging video saliency datasets demonstrate that the proposed method outperforms the state-of-the-art saliency, co-saliency, and video saliency models. At last, the proposed method is summarized at the end of this chapter.

Chapter 8 concludes the thesis and sheds the light on future work in visual saliency detection.

Chapter 2 Literature Review and Preliminary Work

In this chapter, the related works on saliency detection, co-saliency detection, and video saliency detection are firstly reviewed. Then, the benchmark datasets for different saliency detection tasks are introduced. Finally, the evaluation metrics of saliency detection are presented.

2.1 Saliency Detection

2.1.1 RGB Saliency Detection

The last decade has witnessed the remarkable progress of image saliency detection, and a plenty of methods have been proposed and achieved the superior performances, especially the deep learning based methods have yielded a qualitative leap in performances. Following [2], image saliency detection methods can be classified into bottom-up model [25-39] and top-down model [40-55].

Bottom-up model is stimulus-driven, which focuses on exploring low-level vision features. Some visual priors are utilized to describe the properties of salient object based on the visual inspirations from the human visual system, such as contrast prior [25], background prior [29, 36], and compactness prior [33]. In addition, some traditional techniques are also introduced to achieve image saliency detection, such as sparse representation [27], cellular automata [30], random walks [31], low-rank recovery [34], Bayesian theory [35], and frequency domain analysis [39]. Top-down model is task-driven, which utilizes supervised learning with labels and achieves high performance. Especially, deep learning technique has been demonstrated the powerful ability in saliency detection. Some hierarchical deep networks for saliency detection are proposed, such as SuperCNN [44], and DHSNet [47]. In addition, the multi-scale or multi-context deep saliency network is proposed to learn more comprehensive features, such as deep contrast network [46], network with short connections [49], multi-context deep learning framework [52], and multi-scale deep network [53]. The symmetrical network is also introduced in saliency detection, such as the encoder-decoder fully cnvolutional

networks [50]. Moreover, some deep weakly supervised methods for salient object detection are proposed by using the image-level supervision [54] or noisy annotation [55].

2.1.2 RGBD Saliency Detection

Different from image saliency detection, RGBD saliency detection model considers the color information and depth cue together to identify the salient object. As a useful cue for saliency detection, depth information is usually utilized in two ways, *i.e.*, directly incorporating as the feature and designing as the depth measure. Depth feature based method [56-63] focuses on using the depth information as a supplement to color feature. Depth measure based method [64-70] aims at capturing comprehensive attributes from the depth map (*e.g.* shape and structure) through the designed depth measures.

In [58], the color, luminance, texture, and depth features were extracted from the RGBD images to calculate the feature contrast maps, and the fusion and enhancement schemes were utilized to produce the final 3D saliency map. Peng *et al.* [59] calculated the depth saliency through a multi-contextual contrast model, which considers the contrast prior, global distinctiveness, and background cue of depth map. Moreover, a multi-stage RGBD saliency model combining the low-level feature contrast, mid-level region grouping, and high-level prior enhancement was proposed. Recently, deep learning has been successfully applied to RGBD saliency detection. Qu *et al.* [61] designed a CNN to automatically learn the interaction between low-level cues and saliency result for RGBD saliency detection. The local contrast, global contrast, background prior, and spatial prior were combined to generate the raw saliency feature vectors, which are embedded into a CNN to produce the initial saliency map. Finally, Laplacian propagation was introduced to further refine the initial saliency map and obtain the final saliency result. In addition to the multi-modal fusion problem that previous RGBD salient object detection focus on, Han *et al.* [62] firstly exposed the cross-modal discrepancy in the RGBD data and proposed two cross-modal transfer learning strategies to better explore modal-specific representations in the depth modality. This work is the pioneering one that involves the cross-modal transfer learning problem in RGBD salient object detection. In [63], Chen *et al*. innovatively modelled the cross-modal complementary part including the RGB and depth data as a

residual function for RGBD saliency detection. Such a re-formulation elegantly posed the problem of exploiting cross-modal complementarity as approximating the residual, making the multi-modal fusion network to be really complementarity-aware. In this work, the high-level contexts and low-level spatial cues were well-integrated, and the saliency maps were enhanced progressively.

In order to capture the comprehensive and implicit attributes from the depth map and enhance the identification of salient object, some depth measures are designed, such as anisotropic center-surround difference measure [65], local background enclosure measure [67], and depth contrast increased measure [69].

- In [65], Ju *et al*. proposed an Anisotropic Center-Surround Difference (ACSD) measure with 3D spatial prior refinement to calculate the depth-aware saliency map.
- Since the backgrounds always contain the regions that are highly variable in depth map, some high contrast background regions may induce false positives. To overcome this problem, Feng *et al*. [67] proposed a Local Background Enclosure (LBE) measure to directly capture salient structure from depth map, which quantifies the proportion of object boundary located in front of the background.
- The salient objects are always placed at different depth levels and occupy small areas according to the domain knowledge in photography. Based on this observation, Sheng *et al*. [69] proposed a Depth Contrast Increased (DCI) measure to pop-out the salient object through increasing the depth contrast between the salient object and distractors.
- Wang *et al*. [70] proposed a multistage salient object detection framework for RGBD images via Minimum Barrier Distance (MBD) transform and multilayer cellular automata based saliency fusion. The depth-induced saliency map was generated through the FastMBD method, and the depth bias and 3D spatial prior were used to fuse different saliency maps at multiple stages.

Depth feature based method is an intuitive and explicit way to achieve RGBD saliency detection, which uses the depth information as an additional feature to supplement color feature, but ignores the potential attributes (*e.g.*, shape and contour) in the depth map. By contrast, depth measure based method aims at exploiting these implicit information to refine the saliency result. However, how to effectively exploit the depth information to enhance the identification of salient object is a relatively difficult work.

2.2 Co-saliency Detection

In co-saliency detection, the inter-image correspondence is introduced as the common attribute constraint to discriminate the common objects from all the salient objects. To achieve co-saliency detection, some low-level or high-level features are firstly extracted to represent each image unit (*e.g.*, superpixels), where the low-level feature describes the heuristic characteristics (*e.g.*, color, texture, luminance), and the high-level feature captures the semantic attributes through some deep networks. Then, using these features, intra and inter saliency models are designed to explore the saliency representation from the perspectives of the individual image and inter image, respectively. For inter-image constraints capturing, different techniques are introduced, such as clustering, similarity matching, low rank analysis, and propagation. Finally, fusion and optimization schemes are utilized to generate the final co-saliency map.

In this section, we discuss two categories of co-saliency detection methods according to the different data, *i.e.*, RGB co-saliency detection and RGBD co-saliency detection. Obviously, different from the RGB co-saliency detection, RGBD co-saliency detection model needs to combine the depth constraint with inter-image correspondence jointly. In addition, similar to the RGBD saliency detection, the depth cue can be used as an additional feature or a measure in RGBD co-saliency detection methods.

2.2.1 RGB Co-saliency Detection

In this subsection, some RGB co-saliency detection models based on different correspondence capturing strategies are reviewed, *i.e.*, matching based method [71-78], clustering based method [79], rank analysis based method [80,81], propagation based method [82,83], and learning based method [84-87].

In most of the existing methods, ***inter-image correspondence is simulated as a similarity matching process among basic units***. As a pioneering work, Li and Ngan [71] proposed a co-saliency detection model for an image pair, where the inter-image correspondence is formulated as the similarity between two nodes through the normalized single-pair SimRank on a co-multilayer graph. However, this method is only applicable to image pairs. Liu *et al*. [76] proposed a hierarchical segmentation based co-saliency detection model, where the inter-image correspondence is formulated as the global similarity of each region.

Clustering is an effective way to build the inter-image correspondence, where the co-salient regions should be assigned to the same category. A cluster-based co-saliency detection algorithm without heavy learning for multiple images was proposed in [79]. Taking the cluster as the basic unit, an inter-image clustering model was designed to represent the multi-image relationship by integrating the contrast, spatial, and corresponding cues. The proposed method achieved a substantial improvement in efficiency.

Ideally, ***feature representations of co-salient objects should be similar and consistent, thus, the rank of feature matrix should appear low.*** Cao *et al*. [80] proposed a fusion framework for co-saliency detection based on rank constraint, which is valid for multiple images and also works well on single image saliency detection. The self-adaptive weights for fusion process were determined by the low-rank energy. Moreover, this method can be used as a universal fusion framework for multiple saliency maps.

Propagation scheme among multiple images is presented to capture the inter-image relationship. A co-saliency detection method based on two-stage propagation was proposed in [82], where the inter-saliency propagation stage is utilized to discover common properties and generate the pairwise common foreground cue maps, and the intra-saliency propagation stage aims at further suppressing the backgrounds and refining the inter-saliency propagation maps.

Recently, ***learning based methods for RGB co-saliency detection attract more and more attention and achieve competitive performance***. In [84], Zhang *et al*. proposed a co-saliency detection model from deep and wide perspectives under the Bayesian framework. From the deep perspective, some higher-level features extracted by the convolutional neural network with additional adaptive layers were used to explore better representations. From the wide perspective, some visually similar neighbors were introduced to effectively suppress the common background regions. This method is a pioneering work to achieve co-saliency detection by using deep learning, which mainly uses the convolutional network to extract better feature representations of the target. With the FCN framework, Wei *et al*. [85] proposed an end-to-end group-wise deep co-saliency detection model. First, the semantic block with 13 convolutional layers was utilized to obtain the basic feature representation. Then, the group-wise feature representation and single feature representation were captured to

represent the group-wise interaction information and individual image information, respectively. Finally, the collaborative learning structure with the convolution-deconvolution model was used to output the co-saliency map. The overall performance of this method is satisfactory, but the boundary of the target needs to be sharper. The model aims to learn a predictor for each instance through maximizing inter-class distances and minimizing intra-class distances. The theory is to gradually learn from the easy/faithful samples to more complex/confusable ones. Zhang *et al*. [86] proposed a novel framework for co-saliency detection by integrating the Multi-Instance Learning (MIL) regime into Self-Paced Learning (SPL) paradigm. Metric learning was introduced into co-saliency detection in [87], which jointly learns discriminative feature representation and co-salient object detector via a new objective function. This method has the capacity to handle the wide variation in image scene and achieves superior performance.

2.2.2 RGBD Co-saliency Detection

Combining the depth cue with inter-image correspondence, RGBD co-saliency detection can be achieved. Limited by the data sources, only a few of methods are proposed to achieve RGBD co-saliency detection. Fu *et al*. [88] introduced the RGBD co-saliency map into an object-based RGBD co-segmentation model with mutex constraint, where the depth cue is utilized to enhance identification of common foreground objects and provide local features for region comparison. In [89], Song *et al*. proposed a bagging-based clustering method for RGBD co-saliency detection. The inter-image correspondence was explored via feature bagging and regional clustering. Moreover, three depth cues, including average depth value, depth range, and the Histogram of Oriented Gradient (HOG) on the depth map, were extracted to represent the depth attributes of each region. In this thesis, in order to further promote the development of this direction, three new proposed algorithms around this topic will be introduced.

2.3 Video Saliency Detection

Video sequences provide the sequential and motion information in addition to the

color appearance, which benefit for the perception and identification of scene. The salient object in video is defined as the repeated, motion-related, and distinctive target. The repeated attribute constrains the salient object that should appear in most of the video frames. The motion-related characteristic is consistent with the human visual mechanism that the moving object attracts more attention than the static one. The distinctive property indicates the object should be prominent with respect to the background in each frame. Most of the video saliency detection methods are dedicated to exploiting the low-level cues (*e.g.*, color appearance, motion cue, and prior constraint) [90-102]. Only a few works focus on learning the high-level features and extracting the salient object in video through a learning network [103-107].

2.3.1 Low-level Cue Based Video Saliency Detection

According to the way of spatiotemporal extraction, low-level cue based video saliency detection method is classified into fusion model and direct-pipeline model, as shown in Fig. 2-1. For the fusion model, the spatial and temporal features are extracted to generate the spatial saliency and temporal saliency respectively, then they are combined to produce the final spatiotemporal saliency. By contrast, the direct-pipeline model directly extracts the spatiotemporal feature to generate the final spatiotemporal saliency in a straightforward and progressive way without any branches.

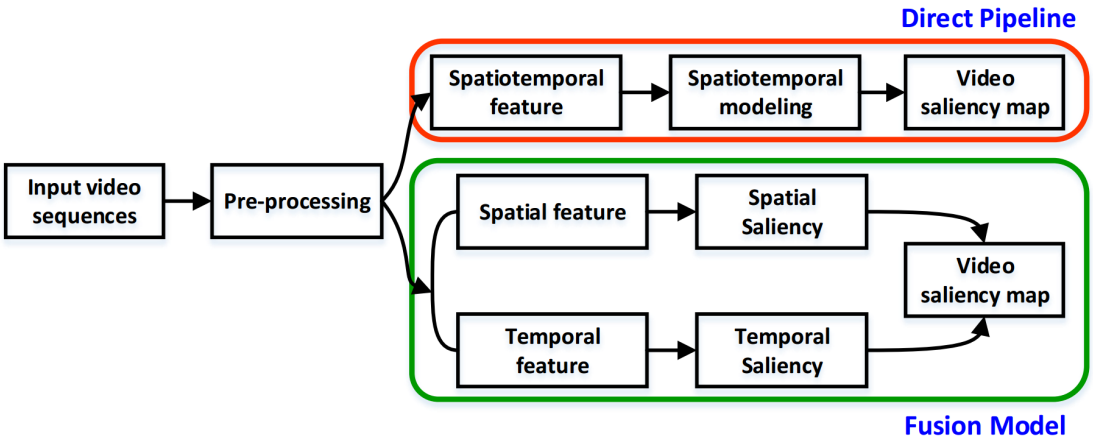

Fig. 2-1 Framework of low-level cue based video saliency detection.

Fusion model fuses the spatial saliency and temporal saliency to achieve video saliency, where the spatial cue represents the intra-frame information in each frame and the temporal cue describes the inter-frame relationship among multiple frames.

For spatial saliency detection, some techniques and priors in image saliency

detection can be used, such as sparse reconstruction, low-rank analysis, center-surround contrast prior, and background prior. For example, sparse reconstruction was utilized to discover the regions with high center-surround contrast [90], the static saliency map was generated via feature contrast in compressed domain [92], the global contrast and spatial sparsity were used to measure the spatial saliency of each superpixel [94], the background prior was utilized to calculate the spatial saliency [99], and color contrast was used to define the color saliency [100].

For temporal saliency, the motion cue is exploited to represent the moving objects in the video. In [90], the target patch was reconstructed by overlapping patches in neighboring frames. Fang *et al*. [92] exploited the motion vectors extracted from the video bitstream to calculate the feature differences between DCT blocks. The superpixel-level temporal saliency was evaluated by motion distinctiveness of motion histograms in [94]. Xi *et al*. [99] used the SIFT flow and bidirectional consistency propagation to define the temporal background prior. In [100], the motion gradient guided contrast computation was utilized to define the temporal saliency.

In most of the fusion based models, fusion strategy is not a key issue. Some simple strategies have been developed, such as a fusion scheme considering the saliency characteristic [92], an adaptive fusion method at the pixel level [94], a simple addition strategy [99]. In [100], Chen *et al*. conducted the modeling-based saliency adjustment and low-level saliency fusion to produce the fusion result. Furthermore, the low-rank coherency guided spatial-temporal saliency diffusion and saliency boosting strategies were adopted to improve the temporal smoothness and saliency accuracy.

Direct-pipeline model directly extracts the spatiotemporal feature to generate the final spatiotemporal saliency in a straightforward and progressive way.

In [91], the stacked temporal slices along X-T and Y-T planes were used to represent the spatiotemporal feature, and the motion saliency was calculated by low-rank and sparse decomposition, where the low-rank component corresponds to the background, and the sparse proportion represents the moving foreground object.\par

Optical flow and its deformations are utilized to define the spatiotemporal feature. Wang *et al*. [95] presented a spatiotemporal saliency model based on gradient flow field and energy optimization, which is robust to complex scenes, various motion patterns, and diverse appearances. The gradient flow field represented the salient regions by incorporating the intra-frame and inter-frame information. In [101], Liu *et al*. presented

a progressive pipeline for video saliency detection, including the superpixel-level graph based motion saliency, temporal propagation, and spatial propagation. The motion saliency was measured by the shortest path on the superpixel-level graph with global motion histogram feature. Guo *et al*. [102] introduced a salient object detection method for video from the perspective of object proposal via a more intuitive visual saliency analysis. The salient proposals were firstly determined by spatial saliency stimuli and contrast-based motion saliency cue. Then, proposal ranking and voting schemes were conducted to screen out non-salient regions and estimate the initial saliency. Finally, temporal consistency and appearance diversity were considered to refine the initial saliency map. It is worth learning that object proposal provides a more comprehensive and high-level representation to detect the salient object.

In addition, motion knowledge is used to capture the spatiotemporal feature. Kim *et al*. [96] exploited the random walk with restart to detect the salient object in video, where the temporal saliency calculated by motion distinctiveness, temporal consistency, and abrupt change is employed as the restarting distribution of random walker. In [97,98], the spatial edge and motion boundary were incorporated as the spatiotemporal edge probability cue to estimate the initial object on the intra-frame graph, and the spatiotemporal saliency was calculated by the geodesic distance on the inter-frame graph.

In summary, the fusion model is a more intuitive method compared with the direct-pipeline model. Moreover, the existing image saliency methods can be directly used to compute the spatial saliency, which lays the foundation for spatiotemporal saliency calculation. Therefore, most of the methods pay more attention to this type.

2.3.2 Learning Based Video Saliency Detection

The supervised learning method aims at learning the spatiotemporal features for video saliency detection by means of a large number of labelled video sequences. Le *et al*. [104] proposed a deep model to capture the SpatioTemporal deep Feature (STF), which consists of the local feature produced by a region-based CNN and the global feature computed from a block-based CNN with temporal-segments embedding. Using the STF feature, random forest and spatiotemporal conditional random field models were introduced to obtain the final saliency map. In [105], Wang *et al*.designed a deep saliency detection model for video, which captures the spatial and temporal saliency

information simultaneously. The static network generated the static saliency map for each individual frame via the FCNs, and the dynamic network employed frame pairs and static saliency map as input to obtain the dynamic saliency result. Le *et al*. [106] proposed an end-to-end 3D fully convolutional network for salient object detection in video, which contains an encoder network and a decoder network.

Compared with supervised learning methods, only a few works focus on unsupervised learning model. As a pioneering work, Li *et al*. [107] proposed an unsupervised approach for video salient object detection by using the saliency-guided stacked autoencoders. First, saliency cues extracted from the spatiotemporal neighbors at three levels (*i.e.*, pixel, superpixel, and object levels) were combined as a high-dimensional feature vector. Then, the stacked autoencoders were learned in an unsupervised manner to obtain the initial saliency map. Finally, some post-processing operations were used to further highlight the salient objects and suppress the distractors. In this method, manual intervention will be further reduced if the hand-crafted saliency cues are automatically learned from the network.

For video saliency detection, motion cue is crucial to suppress the backgrounds and static salient objects, especially in the case of multiple objects. In general, optical flow is a common technique to represent the motion attribute. However, it is time-consuming and sometimes inaccurate, which will degenerate the efficiency and accuracy. Therefore, some deep learning based methods directly embed the continuous multiple frames into the network to learn the motion information and avoid the optical flow calculation. Of course, the video frame and optical flow can be simultaneously embedded into the network to learn the spatiotemporal feature. However, the first option may be better in terms of efficiency. In addition, the salient objects should be consistent in appearance among different frames. Therefore, some techniques, such as energy function optimization, are adopted to improve the consistency of the salient object.

2.4 Benchmark Datasets

In this section, the benchmark datasets for image saliency detection, co-saliency detection, and video saliency detection are introduced, respectively.

For image saliency detection, a number of datasets have been constructed over the

past decade, including some large datasets with pixel-level annotations, such as DUT-OMRON [37], HKU-IS [53], MSRA10K [108], and XPIE [109], as listed in Table 2-1. Benefiting from the growth of data volume, deep learning based RGB saliency detection methods have achieved superior performance. In contrast, the datasets with pixel-wise ground truth annotations for RGBD saliency detection are relatively inadequate, which only consist of NLPR dataset [59] and NJUD dataset [65], as listed in the last two rows of Table 2-1. The NLPR dataset includes 1000 RGBD images with the resolution of 640640, where the depth maps are captured by Microsoft Kinect. The NJUD dataset is released on 2015, which includes 2000 RGBD images with the resolution of 600×600. The depth map in the NJUD dataset is estimated by the stereo images.

Table 2-1 Brief introduction of saliency detection for RGB image and RGBD images.

Dataset	Image number	Max resolution	Depth attribute	Object property	Background property
ACSD [39]	1000	400×400	-	single, moderate	clean, simple
ECSSD [26]	1000	400×400	-	single, large	clean, simple
DUT-OMRON [37]	5168	400×400	-	single, small	complex
MSRA10K [108]	10000	400×400	-	single, large	clean, simple
PASCAL-S [38]	1000	500×500	-	multiple, moderate	simple
HKU-IS [53]	850	400×400	-	multiple, moderate	clean
XPIE [109]	4447	300×300	-	single, moderate	complex
STEREO [56]	797	1200×900	depth estimation	single, moderate	diverse
NLPR [59]	1000	640×640	Kinect capturing	single, moderate	diverse
NJUD [65]	2000	600×600	depth estimation	single, moderate	diverse

For co-saliency detection, five RGB datasets and two RGBD datasets are commonly used for evaluation, as listed in Table 2-2. MSRC [110] is a challenging dataset with complex background, which contains 7 image groups of totally 240 images with manually pixel-wise ground truth. The iCoseg [111] dataset consists of 38 image groups of totally 643 images, and the manually labeled pixel-wise ground-truth masks is also provided. Image Pair [71] dataset only contains image pairs, whereas other datasets usually include more than two images in each group. A larger co-saliency detection dataset named Cosal2015 is constructed in [112], which consists of 2015 RGB

images distributed in 50 image groups with pixel-wise ground truth. INCT2016 [113] is a more challenging dataset with larger appearance variation, indefinite number of targets, and complicated backgrounds, which contains 291 images distributed in 12 categories with pixel-level ground truth. There are two commonly used datasets with pixel-level hand annotations for RGBD co-saliency detection. One is the RGBD Coseg183 dataset [88], which contains 183 RGBD images in total that distributed in 16 image groups. The other one is the RGBD Cosal150 dataset [114], which collects 21 image groups containing a total of 150 RGBD images.

Table 2-2 Brief introduction of co-saliency detection datasets.

Dataset	Image number	Group number	Group size	Resolution	Depth attribute	Object property	Background property
MSRC [110]	240	7	30-53	320×210	-	complex	clean, simple
iCoseg [111]	643	38	4-42	500×300	-	multiple	diverse
Image Pair [71]	210	115	2	128×100	-	single	clustered
Cosal2015 [112]	2015	50	26-52	500×333	-	multiple	clustered
INCT2016 [113]	291	12	15-31	500×375	-	multiple	complex
RGBD Coseg183 [88]	183	16	12-36	640×480	Kinect capturing	multiple	complex
RGBD Cosal150 [114]	150	21	2-20	600×600	depth estimation	single	diverse

For video saliency detection, many datasets have been released, such as ViSal [95], MCL [96], UVSD [101], VOS [107], SegTrackV1 [115], SegTrackV2 [116], and DAVIS [117], as listed in Table 2-3. The DAVIS dataset is a commonly used and challenging dataset, which contains 50 video sequences with the fully-annotated pixel-level ground truth for each frame. The UVSD dataset is a specially designed and newly established dataset for video saliency detection, which consists of 18 unconstrained videos with complicated motion patterns and cluttered scenes, and the pixel-wise ground truth for each frame is available. A very large video saliency detection dataset named VOS is constructed, which consists of 116103 frames in total that distributed in 200 video sequences. In this dataset, 7467 frames are annotated into binary ground truth, which is suitable for training and learning a deep model to extract the salient object in video.

Table 2-3 Brief introduction of video saliency detection datasets.

Dataset	Frame number	Video number	Video size	Resolution	Object property	Background property
SegTrackV1 [115]	244	6	21-71	414×352	single	diverse
SegTrackV2 [116]	1065	14	21-279	640×360	single	diverse
ViSal [95]	963	17	30-100	512×228	single	diverse
MCL [96]	3689	9	131-789	480×270	single, small	complex
DAVIS [117]	3455	50	25-104	1920×1080	multiple	complex
UVSD [101]	6524	18	71-307	352×288	single, small	clustered
VOS [107]	116103	200	~ 500	800×800	single	complex

2.5 Evaluation Metrics

In addition to directly comparing the saliency map with ground truth, some evaluation metrics are developed to quantitatively evaluate the performance of saliency detection methods, such as Precision-Recall (PR) curve, F-measure, Receive Operator Characteristic (ROC) curve, Area Under the Curve (AUC) score, Mean Absolute Error (MAE), and S-measure.

Precision-Recall (PR) curve and F-measure. By thresholding the saliency map with a series of fixed integers from 0 to 255, the binary saliency masks are achieved. Therefore, the precision and recall scores are calculated by comparing the binary mask with the ground truth. The PR curve is drawn under different precision and recall scores, where the vertical axis denotes the precision score, and the horizontal axis corresponds to the recall score. The closer the PR curve is to the coordinate (1,1), the better performance achieves. In order to comprehensively evaluate the saliency map, a weighted harmonic mean of precision and recall is defined as F-measure [2], which is expressed as:

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

where β^2 is generally set to 0.3 for emphasizing the precision as suggested in [39].

Receive Operator Characteristic (ROC) curve and AUC score. The ROC curve describes the relationship between the false positive rate (FPR) and true positive rate (TPR), which is represented as:

$$TPR = \frac{|S_F \cap G_F|}{|G_F|} \quad FPR = \frac{|S_F \cap G_B|}{|G_B|} \tag{2-2}$$

where S_F, G_F, and G_B denote the set of detected foreground pixels in the binary saliency mask, the set of foreground pixels in the ground truth, and the set of background pixels in the ground truth, respectively. The closer the ROC curve is to the upper right, the better performance achieves. AUC score is the area under the ROC curve, and the larger, the better.

Mean Absolute Error (MAE). MAE score directly evaluates the difference between the continuous saliency map S and ground truth G directly:

$$MAE = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} |S(x,y) - G(x,y)| \tag{2-3}$$

where w and h represent the width and height of the image, respectively. The smaller the MAE score is, the more similar to the ground truth, and the better performance achieves.

S-measure. S-measure [118] evaluates the structural similarity between the saliency map and ground truth as:

$$S_m = \alpha \cdot S_o + (1-\alpha) \cdot S_r \tag{2-4}$$

where α is set to 0.5 for assigning equal contribution to both region (S_o) and object (S_r) similarity.

Chapter 3 Saliency Detection for Stereoscopic Images Based on Depth Confidence Analysis and Multiple Cues Fusion

Stereoscopic perception is an important part of human visual system that allows the brain to perceive the depth of scene. However, depth information has not been fully explored in the existing saliency detection models. In this chapter, a novel saliency detection method for stereoscopic images is proposed combining the depth confidence analysis and multiple cues fusion. More details and experiments are introduced in the following sections.

3.1 Introduction

Saliency detection aims to effectively highlight the salient objects and suppress the background regions. Most of the previous works on saliency detection mainly concentrate on RGB color information while ignoring depth/disparity cue [25-55]. In fact, stereoscopic depth information has demonstrated its usefulness for many computer vision tasks [3,7,12,13], including saliency detection [56-70]. However, limited by the depth imaging techniques, sometimes the quality of the depth map is not satisfactory. As we all know, a good depth map benefits for the saliency detection, whereas a poor depth map may degenerate the saliency measurement. Consequently, it is vital that construct a measure to describe the quality of depth map. According to the observation of depth distribution, a confidence measure for depth map is introduced into the model to reduce the negative influence of poor depth map on saliency detection. In addition, in order to make full use of the depth information, a novel compactness saliency model is defined by integrating color and depth information, and design a depth-refined foreground seeds selection mechanism to calculate the foreground saliency by using the multiple cues contrast. Finally, these two complementary saliency models are fused to generate more robust saliency result.

The main contributions of this method can be summarized as follows: (1) According to the observation of depth distribution, a depth confidence measure is

proposed to evaluate the quality of input depth map and reduce the influence of poor depth map on saliency detection; (2) A stereoscopic compactness model integrating color and depth information is put forward to compute the compactness saliency; (3) A depth-refined foreground seeds selection mechanism is presented. The foreground saliency is measured by contrast between the target regions with seed regions, which integrate color, depth, and texture cues.

3.2 Proposed RGBD Saliency Model

The flowchart of the proposed stereo saliency detection method is depicted as Fig. 3-1. First, a depth confidence measure is calculated to evaluate the reliability of depth map, and used in the following processes. The depth confidence measure can reduce the negative influence of poor depth map on saliency detection. Simultaneously, RGB image is abstracted into superpixels and represented as a graph. Then, compactness saliency is calculated by using the color and depth cues. Further, some foreground seeds are determined via a depth-refined foreground seeds selection mechanism. Taking color, depth, and texture cues into consideration, the foreground saliency is calculated through a multiple cues contrast model. Finally, compactness and foreground saliency map are weighted to obtain the final saliency map.

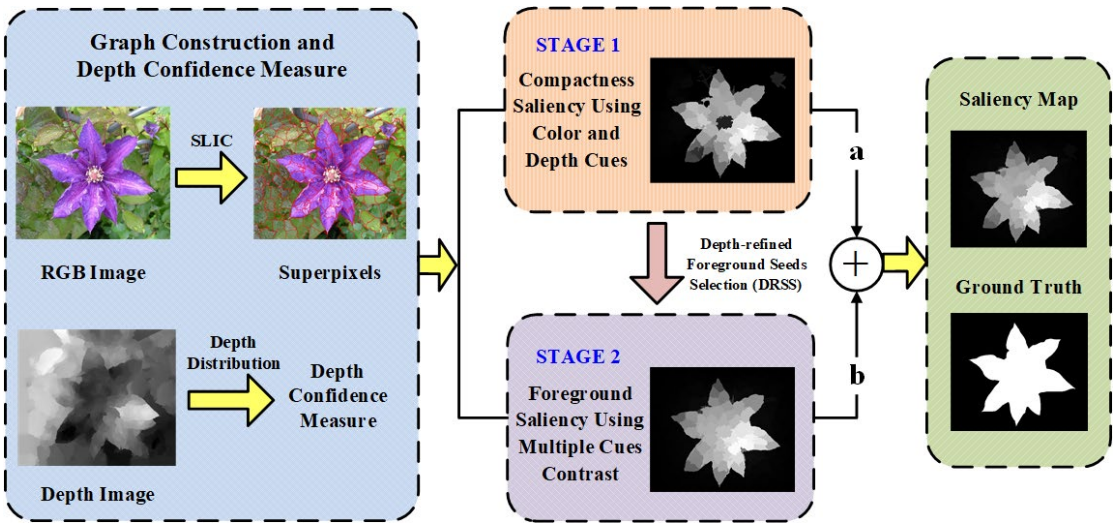

Fig. 3-1 Flowchart of the proposed method.

3.2.1 Depth Confidence Measure

The quality of depth map is very important for the using of depth cue. Specifically, a good depth map can provide accurate depth information which benefits for saliency detection, while the poor depth map may degenerate the saliency measurement. Thus, a depth confidence measure is proposed in this letter to evaluate the reliability of depth map. Observing the different qualities of depth maps, we found that a good depth map often owns clear hierarchy, and the salient objects can be distinctly highlighted from the backgrounds. Fig. 3-2 shows some examples of different qualities of depth maps.

(a) (b) (c)

Fig. 3-2 Different qualities of depth maps. (a) Good depth map, $\lambda_d = 0.8014$. (b) Common depth map, $\lambda_d = 0.3890$. (c) Poor depth map, $\lambda_d = 0.0422$.

The input depth map is roughly ranked into three grades, *i.e.*, good, common, and poor. Based on the observation of depth statistical characteristic, we found that the values of good depth map usually concentrate on a lower range, whereas the values of poor depth map tends to distribute on a relatively larger range. Therefore, the mean value of the whole depth image is an effective parameter to tell them apart. In statistics, coefficient of variation is used to evaluate the dispersion degree of the data. It is observed that the poor depth map appears strong concentration compared with other cases. Thus, the coefficient of variation is introduced in the confidence measure. In addition, there is a more random distribution for a common depth map, therefore, the depth frequency entropy is defined to evaluate the randomness of an input depth map. In summary, the depth confidence measure is defined as:

$$\lambda_d = \exp\left((1-m_d)\cdot CV \cdot H\right) - 1 \tag{3-1}$$

where m_d is the mean value of the depth map, $CV = m_d/\sigma_d$ is coefficient of variation, σ_d is the standard deviation of the depth map, and H is the depth frequency entropy representing the randomness of depth distribution, which is defined as:

$$H = -\sum_{i=1}^{L} P_i \log\left(P_i\right) \tag{3-2}$$

where $P_i = n_i/n_\Sigma$, n_Σ is the number of pixels in the depth map, n_i is the number of

pixels that belong to the region level r_i, and L is the levels of depth map. Note that, the input depth map is firstly normalized to [0, 1]. Then, $L-1$ thresholds, namely T_k, are used to divide the depth map into L levels. A larger λ_d corresponds to more reliable of the input depth map. As shown in Fig. 3-2, the depth confidence measure effectively distinguishes different qualities of depth maps according to the statistical characteristics of depth map.

3.2.2 Graph Construction

The input RGB image is abstracted into some homogenous and compact superpixels using SLIC method [119]. The number of superpixel N is set to 200 in all the experiments. Then, a graph $G=(V,E)$ is constructed, where V represents the set of nodes corresponding to the superpixels, and E denotes the set of links between adjacent superpixels.

The color difference l_{ij} in Lab color space and depth difference d_{ij} between superpixels v_i and v_j are defined as

$$l_{ij} = \left\| \boldsymbol{c}_i - \boldsymbol{c}_j \right\|_2 \tag{3-3}$$

and

$$d_{ij} = \left| d_i - d_j \right| \tag{3-4}$$

where $\boldsymbol{c}_i$ is the mean color value of superpixel v_i, d_i denotes the mean depth value of superpixel v_i, $\|\cdot\|_2$ and $|\cdot|$ represent the ℓ_2-norm functrion and absolute value function, respectively.

Therefore, the similarity between superpixels v_i and v_j is defined as:

$$a_{ij} = \exp\left(-\frac{l_{ij} + \lambda_d \cdot d_{ij}}{\sigma^2} \right) \tag{3-5}$$

where σ^2 is a parameter to control strength of the similarity, which is set to 0.1 in all experiments.

The affinity matrix $\mathbf{W} = \left[w_{ij} \right]_{N \times N}$ is defined as the similarity between two adjacent superpixels.

$$w_{ij} = \begin{cases} a_{ij}, & if \ j \in \Omega_i \\ 0, & otherwise \end{cases} \tag{3-6}$$

where Ω_i is the set of neighbors of superpixel v_i.

3.2.3 Compactness Saliency Using Color and Depth Cues

The salient regions incline to have a small spatial variance, whereas the backgrounds usually have a high spatial variance since their superpixels are often distributed over the entire image [33]. In fact, depth map also exhibits limited compactness, that is, the depth values of salient regions are more likely to have a centralized distribution near the center of image. Motivated by this, the compactness saliency is calculated by using the color and depth cues. The novel stereoscopic compactness saliency is defined as:

$$S_{CS}(i) = \left[1 - norm\left(cc(i) + dc(i)\right)\right] \cdot Obj(i) \tag{3-7}$$

where $cc(i)$ is the color-based compactness of superpixel v_i, $dc(i)$ is the depth-based compactness of superpixel v_i, and $norm(\cdot)$ is the min-max normalization function. Considering the importance of location for saliency detection, objectness measure $Obj(i)$ [120] is introduced to evaluate the probability of superpixel v_i that belongs to an object. The color and depth-based compactness are defined as:

$$cc(i) = \frac{\sum_{j=1}^{N} a_{ij} \cdot n_j \cdot \left\| \boldsymbol{b}_j - \boldsymbol{\mu}_i \right\|_2}{\sum_{j=1}^{N} a_{ij} \cdot n_j} \tag{3-8}$$

and

$$dc(i) = \frac{\sum_{j=1}^{N} a_{ij} \cdot n_j \cdot \left\| \boldsymbol{b}_j - \boldsymbol{p} \right\|_2 \cdot \exp\left(-\frac{\lambda_d \cdot d_i}{\sigma^2}\right)}{\sum_{j=1}^{N} a_{ij} \cdot n_j} \tag{3-9}$$

where a_{ij} is the similarity between two superxpels after manifold ranking [121], n_j denotes the number of pixels that belong to superpixel v_j, which emphasizes the impact of larger region, $\boldsymbol{b}_j = \left[b_j^x, b_j^y\right]$ is the centroid coordinate of superpixel v_j, $\boldsymbol{p} = \left[c_x, c_y\right]$ is the spatial position of the image center, and the spatial mean $\boldsymbol{\mu}_i = \left[\mu_i^x, \mu_i^y\right]$ is defined as:

$$\mu_i^x = \frac{\sum_{j=1}^{N} a_{ij} \cdot n_j \cdot b_j^x}{\sum_{j=1}^{N} a_{ij} \cdot n_j} \tag{3-10}$$

and

$$\mu_i^y = \frac{\sum_{j=1}^{N} a_{ij} \cdot n_j \cdot b_j^y}{\sum_{j=1}^{N} a_{ij} \cdot n_j}. \tag{3-11}$$

3.2.4 Foreground Saliency Using Multiple Cues Contrast

Although the stereoscopic compactness saliency model is active on some level, there are some limitations. For example, when the salient regions have similar appearances with backgrounds, the regions may be wrongly detected. Hence, a foreground saliency model based on multiple cues contrast is proposed to mitigate this problem.

Traditionally, foreground seeds are selected only based on the preliminary saliency map. Considering the effectiveness of depth information, the foreground seeds are selected while constraining them to have larger values of compactness saliency and depth simultaneously. Therefore, a depth-refined foreground seeds selection method (DRSS) is proposed as shown in Fig. 3-3. First, preliminary seeds are determined by thresholding segmentation of compactness saliency map, where the threshold τ is set to 0.5. Then, the mean depth value of preliminary seeds is used to refine the preliminary seeds and obtain the final foreground seeds set.

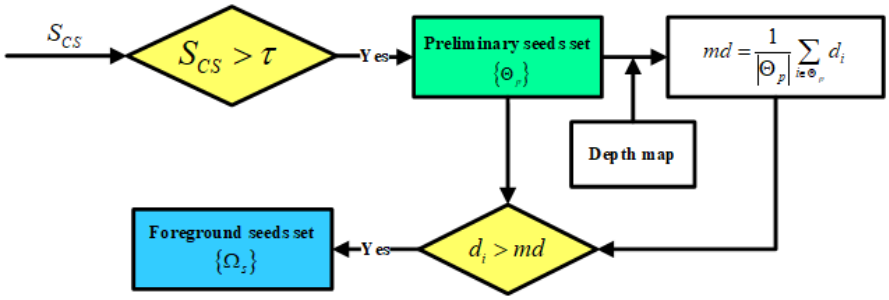

Fig. 3-3 Flowchart of depth-refined foreground seeds selection mechanism.

Next, the feature contrast of each superpixel with the foreground seeds is calculated by using the multiple cues including color, depth, texture, and spatial position. A superpixel is more likely to be salient if it is more similar to the foreground seeds. The foreground saliency is computed as follows:

$$S_{fg}(i) = \sum_{j\in\Omega_s}\left[a_{ij}\cdot D_t(i,j)\cdot\exp\left(-\left\|\boldsymbol{b}_i-\boldsymbol{b}_j\right\|_2/\sigma^2\right)\cdot n_j\right] \tag{3-12}$$

where Ω_s is the set of foreground seeds, $\left\|\boldsymbol{b}_i-\boldsymbol{b}_j\right\|$ denotes the Euclidean distance between positions of two superpixels, and $D_t(i,j)$ is the texture similarity between superpixels using LBP feature [122], which is defined as:

$$D_t(i,j) = \frac{\left|\boldsymbol{k}_i^T\boldsymbol{k}_j\right|}{\left\|\boldsymbol{k}_i\right\|_2\cdot\left\|\boldsymbol{k}_j\right\|_2} \tag{3-13}$$

where k_i is LBP histogram frequency of superpixel v_i. To avoid the problem that saliency map highlights object boundaries rather than the entire region, manifold ranking method is used to propagate the foreground saliency map. At last, the map after propagation is normalized to [0,1], and the final foreground saliency map S_{FS} is obtained.

3.2.5 Saliency Map Integration

The compactness and foreground saliency maps are complementary to each other. Considering the foreground saliency map is based on the compactness saliency result, we integrate these two saliency maps through a weighted-sum method.

$$S = \gamma \cdot S_{CS} + (1-\gamma) \cdot S_{FS} \tag{3-14}$$

where S is the final saliency map, and γ is a weighted coefficient that balances the compactness saliency map S_{CS} and foreground saliency map S_{FS}.

3.3 Experimental Results

The performance of the prposed method is evaluated on two RGBD saliency datasets, *i.e.*, NJU-400 [123] and NJUD [65]. The PR curve, F-measure, and MAE score are introduced as the evaluation metrics. In all experiments, the parameters are set to $L = 3$, $T_1 = 0.4$, $T_2 = 0.6$, and $\gamma = 0.8$, respectively.

3.3.1 Performance Comparison

The proposed method is compared with 8 state-of-the-art 2D methods (RC [25], MR [37], DS [124], MAP [125], DCLC [33], LPS [126], BSCA [30], and RRWR [31]), and 2 stereo saliency detection methods (SS [56] and ACSD [65]). Fig. 3-4 shows the evaluation results of the proposed method with 10 state-of-the-art methods on two datasets. On both datasets, the PR curves show that the proposed method performes better than other methods. Similarly, the proposed method achieves the best performance in terms of the average precision, F-measure, and MAE score compared with other approaches due to the depth confidence analysis and two-stage saliency computation mechanism. Taking the F-measure on the NJUD dataset as an example,

the F-measure of ACSD is 0.5552, and the F-measure of our method reaches 0.6055 with the performance gain of 9%. Fig. 3-5 presents visual comparisons of different saliency detection methods. The proposed method has more similar appearances with ground truth, and owns clear contour and uniform salient regions. For example, the two bottles in the second image are detected more complete and accurate, and the background (*e.g.*, the white box) are effectively suppressed. In the third image, the background regions (*e.g.*, the distant trees and bare gound) are obviously suppressed by the proposed method, and the sculpture is highlighted very well. The qualitative and quantitative comparisons demonstrate the effectiveness of the proposed model.

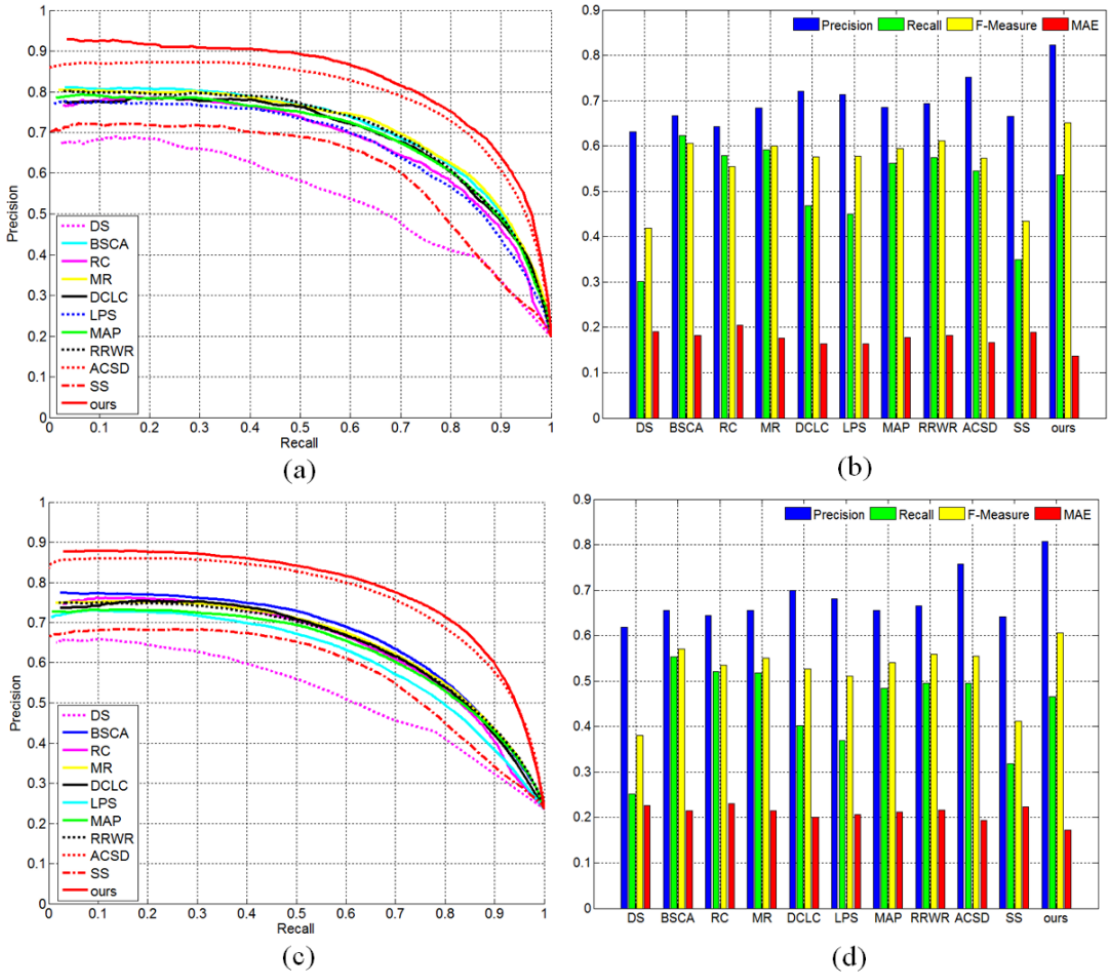

Fig. 3-4 Quantitative comparisons of proposed method with 10 state-of-the-art methods. (a) PR curves of different methods on NJU-400 dataset. (b) Average precision, recall, F-measure, and MAE of different methods on NJU-400 dataset. (c) PR curves of different methods on NJUD dataset. (d) Average precision, recall, F-measure, and MAE of different methods on NJUD dataset.

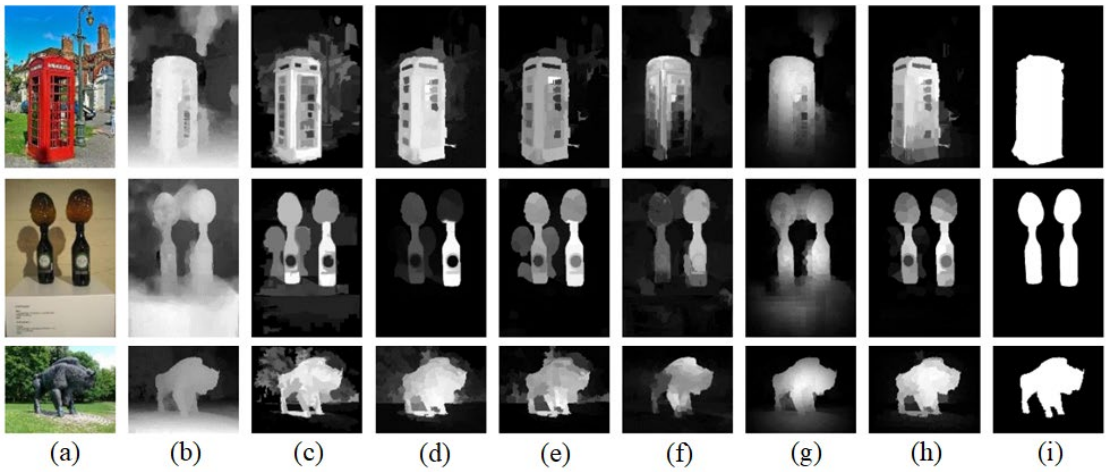

Fig. 3-5 Visual comparisons of saliency maps. (a) Input RGB image. (b) Input depth map. (c) RC. (d) RRWR. (e) DCLC. (f) SS. (g) ACSD. (h) Ours. (i) Ground truth.

3.3.2 Parameter Analysis

In this section, the proposed method under different factors including the depth confidence measure and DRSS scheme is evaluated. The PR curves and quantitative metrics are shown in Fig. 3-6. In order to reduce the influence of poor depth map in stereo saliency detection, depth confidence measure λ_d is introduced. Comparing the black line with the blue line in Fig. 3-6(a), it demonstrates that the performance with λ_d is superior to the result without depth confidence measure. The same conclusion can be drawn from the comparisons of the first two columns in Fig. 3-6(b). At the stage of foreground saliency detection, the DRSS mechanism is proposed to acquire more accurate foreground seeds. As shown in Fig. 3-6, the model with DRSS scheme achieves better performance with higher quantitative metrics.

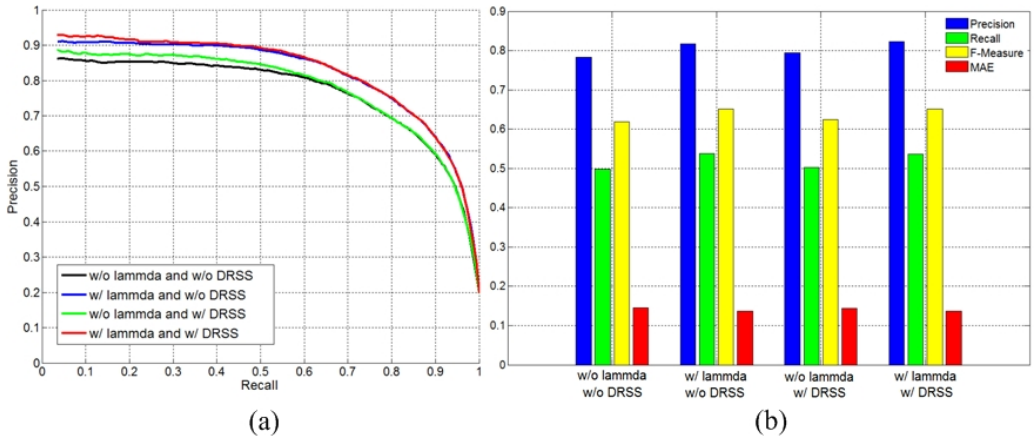

Fig. 3-6 Evaluation of different factors on NJU-400 dataset. (a) PR curves of different factors. (b) Average precision, recall, F-measure, and MAE score of different factors.

3.4 Summary

In this chapter, a novel saliency detection model for stereoscopic images was introduced based on depth confidence analysis and multiple cues fusion. First, the quality of depth map was considered when introduces the depth information into the saliency model, and a depth confidence measure was proposed to evaluate the reliability of depth map. In addition, a novel stereoscopic compactness model integrating color and depth information was proposed to compute the compactness saliency. To achieve more robust saliency detection result, a foreground saliency detection method based on multiple cues contrast was proposed, which includes a depth-refined foreground seeds selection method. At last, weighted-sum method was used to generate the final saliency map. Experimental evaluations on two public benchmarks have validated the advantages of the proposed approach.

Chapter 4 Co-saliency Detection for RGBD Images Based on Multi-constraint Feature Matching and Cross Label Propagation

Co-saliency detection aims at extracting the common salient regions from an image group containing two or more relevant images, which is a newly emerging topic in computer vision community. Different from most of the existing co-saliency detection models focusing on RGB image group, this chapter addresses the co-saliency detection for RGBD images, which utilizes the depth information to enhance the identification of co-saliency. First, the intra saliency map for each image is generated by the single image saliency model, while the inter saliency map is calculated based on the superpixel-level and image-level similarity matching, which represent the corresponding relationship among multiple images. Then, the Cross Label Propagation (CLP) is used to refine the intra and inter saliency maps in a cross way. At last, all the original and optimized saliency maps are integrated to generate the final co-saliency result. Experiments on two RGBD co-saliency datasets demonstrate the effectiveness of the proposed model.

4.1 Introduction

Most of the existing co-saliency detection models mainly focus on RGB image group, which have achieved superior performances [71-84]. However, little work addresses co-saliency detection for RGBD images. In fact, depth information has demonstrated its usefulness for many computer vision tasks including saliency detection, which can reduce the ambiguity with color descriptors and enhance the identification of the object from complex background [56-70]. Motivated by this, depth information is introduced as a supplementary cue of color feature for the co-saliency detection model in this work.

As we all know, it is critical to effectively capture the inter-image correspondence among multiple images in co-saliency detection. In the existing co-saliency detection methods, the inter-image correspondence has been modelled as a similarity matching

[71-78], a cluster process [79], a rank constraint [80,81], a propagation process [82,83] or a learning process [84-87]. In this chapter, the inter saliency of a superpixel is defined as the weighted sum of the intra saliency of corresponding superpixels in other images. In order to explore the inter-image relationship, the similarity matching methods on two levels are designed, where the superpixel-level similarity matching scheme focuses on determining the matching superpixel set for the current superpixel based on three constraints from other images, and the image-level similarity measurement provides a global relationship between two images on the whole image scale. Introducing the multiple constraints into feature matching, the inter-image relationship will become more stable and robust.

In summary, there are two main issues that need to be focused on: (1) effectively capture the corresponding relationship among multiple images, and (2) introduce the depth cue into co-saliency detection. Therefore, a novel co-saliency detection model for RGBD image is proposed, which integrates the depth cue to enhance the identification of co-saliency. The similarity matching on superpixel and image levels is designed to capture the corresponding relationship and constrain the inter saliency map generation. In addition, a Cross Label Propagation (CLP) method is proposed to optimize the intra and inter saliency maps in a cross way.

4.2 Proposed Co-saliency Detection Method for RGBD Images

In this section, the proposed co-saliency detection model for RGBD images is introduced, and the flowchart is shown in Fig. 4-1. First, the basic intra saliency map is generated by single saliency detection method associating with the depth cue on the individual image. Then, two similarity matching methods on different scales are presented to acquire the corresponding relationship among the multiple images. According to the corresponding relationship and intra saliency map, the inter saliency map of each image is generated. In order to further improve the consistency of salient objects and suppress the background regions, a CLP optimization scheme is designed to refine the intra and inter saliency maps in a cross way. At last, the weighted fusion is used to produce the final co-saliency result.

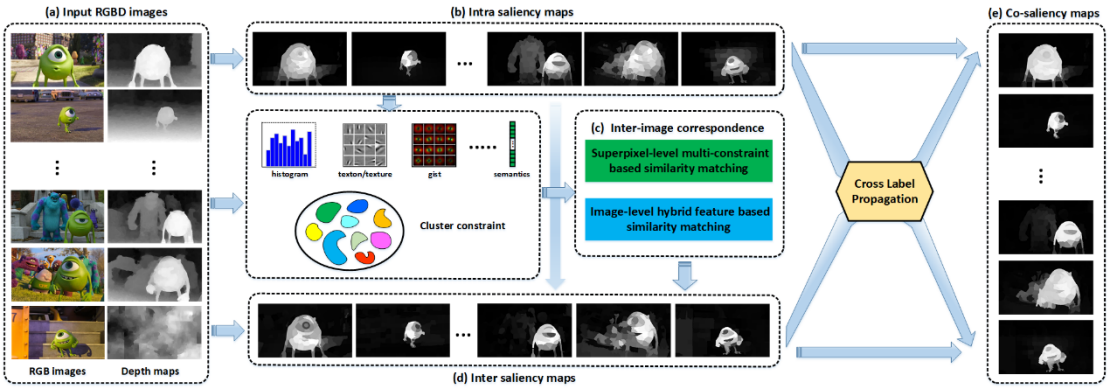

Fig. 4-1 Flowchart of the proposed RGBD co-saliency detection model. (a) The input RGBD images. (b) The intra saliency maps produced by existing single saliency method collaborating with depth information. (c) The inter-image correspondence is obtained by superpixel-level multi-constraint based similarity matching and image-level hybrid feature based similarity matching. Using the corresponding relationship and intra saliency maps, the inter saliency maps (d) are generated. At last, the final co-saliency results (e) are achieved based on CLP scheme.

4.2.1 Intra Saliency Detection

Given N input RGB images $\{I^i\}_{i=1}^N$ and the corresponding depth maps $\{D^i\}_{i=1}^N$. Each RGB image I^i is firstly abstracted into superpixels $R^i = \{r_m^i\}_{m=1}^{N_i}$ by using the SLIC method [119], where N_i denotes the number of superpixels in image I^i. Then, the previous work introduced in Chapter 3 (*i.e.*, DCMC method) is exploited to generate the intra saliency map for each RGBD images. The intra saliency value of a superpixel r_m^i in image I^i is assigned with the mean value of all pixels that belong to superpixel r_m^i in the corresponding intra saliency map, which is denoted as $S_{\text{intra}}(r_m^i)$. Note that, any single saliency method can be utilized to generate the intra saliency map. In general, the more accurate the intra saliency map is, the better it is for the co-saliency computation using the proposed model. The experimental comparisons of different intra saliency maps are discussed in Section 4.3.5.

4.2.2 Inter Saliency Detection

For co-saliency detection, we should determine which salient object is the common one that appears in most of the images. Thus, acquiring the corresponding relationship among multiple images is the key point of co-saliency detection model. In the proposed model, the matching methods on two levels are designed to represent the correspondence among multiple images. The first one is the superpixel-level multi-constraint based similarity matching scheme, which focuses on determining the

matching superpixel set for the current superpixel by using three constraints from other images. The second is the image-level hybrid feature based similarity measurement, which provides a global relationship on the whole image scale. With the corresponding relationship, the inter saliency of a superpixel is defined as the weighted sum of the intra saliency of corresponding superpixels in other images.

1) Superpixel-level multi-constraint based similarity matching

At the superpixel level, the correspondence is represented as the multi-constraint based matching relationship between superpixels among the multiple images, which combines the similarity constraint, saliency consistency, and cluster-based constraint.

Similarity constraint. In this work, the color and depth cues are simultaneously considered to represent the similarity constraint. However, for some RGBD images, the depth map is seriously noisy, which may degenerate the accuracy of the measurement. To address this issue, the depth confidence measure λ_d is introduced to evaluate the reliability of depth map as presented in Chapter 3. A larger λ_d corresponds to more reliable of the input depth map. The detailed definition is not described here again. In the model, λ_d is used as a controller for the introduction of depth information. Then, the similarity matrix $\boldsymbol{S}=\left[s\left(r_m^i,r_n^j\right)\right]_{N_i\times N_j}$ between two superpixels from the images I^i and I^j is defined as:

$$s\left(r_m^i,r_n^j\right)=\exp\left(-\frac{\left\|\boldsymbol{c}_m^i-\boldsymbol{c}_n^j\right\|_2+\min(\lambda_d^i,\lambda_d^j)\cdot\left|d_m^i-d_n^j\right|}{\sigma^2}\right) \tag{4-1}$$

where $\boldsymbol{c}_m^i$ is the mean color vector of superpixel r_m^i in the Lab color space, d_m^i denotes the mean depth value of superpixel r_m^i, λ_d^i represents the depth confidence measure of depth map D^i, $\|\cdot\|_2$ is the ℓ_2-norm function, σ^2 is a parameter to control strength of the similarity, which is fixed to 0.1. Based on the similarity matrix in Eq. (4-1), the $K_{\max}$ nearest superpixels in each of other images for superpixel r_m^i can be determined. Further, all these superpixels for r_m^i are composed as the similarity constraint set $\Phi_1\left(r_m^i\right)$.

Saliency consistency. Considering the task of co-saliency detection, the saliency consistency is introduced as another important cue to constrain the feature matching. Thus, the saliency similarity between the target superpixel r_m^i and other superpixels $\left\{r_n^j\right\}_{n=1}^{N_j}$ in image I^j is calculated, and the superpixels with consistent saliency values are selected to generate the saliency consistency set for r_m^i as:

$$\Phi_2\left(r_m^i\right)=\left\{r_n^j \,\middle|\, \left|S_{\text{intra}}\left(r_m^i\right)-S_{\text{intra}}\left(r_n^j\right)\right|\le T_1\right\} \tag{4-2}$$

where $n=\{1,2,\cdots,N_j\}$, and T_1 is a threshold to control strength of the saliency similarity, which is fixed to 0.3 in all experiments.

Cluster-based constraint. Inspired by the fact that the matching superpixels should be grouped into the same cluster, therefore, the cluster-based constraint is introduced to build the cluster-level correspondence. First, K-means++ clustering [127] is used to group the superpixels $\{r_m^i\}_{m=1}^{N_i}$ into K clusters $\{C_k^i\}_{k=1}^{K}$ with the cluster centers $\{c_k^i\}_{k=1}^{K}$. Then, the Euclidean distance is utilized to measure and determine the cluster-level superpixel matching relationship. Specifically, for each superpixel r_m^i, one superpixel with the minimum distance in each of other images is determined. Supposed that the superpixel r_m^i belongs to the cluster C_p^i, and the superpixel r_n^j belongs to the cluster C_q^j. The cluster-level nearest superpixels for superpixel r_m^i are denoted as the set $\Phi_3\left(r_m^i\right)$:

$$\Phi_3\left(r_m^i\right)=\left\{r_n^j \,\middle|\, \arg\min_{C_q^j, q\in[1,K]} Ed\left(c_p^i, c_q^j\right)\right\} \tag{4-3}$$

where $Ed(\cdot)$ denotes the Euclidean distance function, c_p^i and c_q^j are the cluster centers of clusters C_p^i and C_q^j, respectively.

Similarity matching. Three corresponding sets Φ_1, Φ_2, and Φ_3 are combined to determine the matching relationship for each superpixel. The matching matrix $\boldsymbol{ML}^{ij}=\left[ml\left(r_m^i, r_n^j\right)\right]_{N_i\times N_j}$ is defined as:

$$ml\left(r_m^i, r_n^j\right)=\begin{cases}1, & if\ r_n^j\in\left\{\Phi_1\left(r_m^i\right)\cap\Phi_2\left(r_m^i\right)\cap\Phi_3\left(r_m^i\right)\right\}\\ 0, & otherwise\end{cases}. \tag{4-4}$$

2) Image-level hybrid feature based similarity matching

Enlightened by the observation that the greater similarity between two images means the greater likelihood of finding the matching regions, thus, a full-image size similarity descriptor is designed as the weighted coefficient for inter saliency calculation. To evaluate the image similarity, three types of features are used to represent the image property from different aspects and guarantee the completeness and generality of feature selection. First, the color feature, as the common basic feature in most of the saliency detection methods, is introduced in the proposed method. Inspired by the fact that similar images should have approximate depth distributions and similar appearances in salient objects, the depth and saliency histograms are added in the

feature pool. At last, these feature distances are integrated through a self-adaptive weighted strategy to evaluate the similarity between two images.

The features used in the proposed method are listed in Table 4-1. The details of the features are described as follows: the color histogram in the RGB color space is utilized to represent the color distribution; the texton histogram is used to express the texture feature; and the GIST feature [128] is introduced to describe the spatial structure of the scene. In addition, the deep feature produced by VGG network [129] is used to describe the semantic information of the image. Specifically, the fc7 feature with pre-trained VGG16 model on ImageNet is directly extracted as the semantic feature. Moreover, the depth and saliency histograms are used to describe the distributions of the depth map and single saliency map.

Table 4-1 Image property descriptor and the feature distance.

	feature	description	dim	distance
col	$\boldsymbol{h}_c$	RGB histogram	512	$d_{c1}=\chi^2\left(\boldsymbol{h}_c^i,\boldsymbol{h}_c^j\right)$
	$\boldsymbol{t}$	texton histogram	15	$d_{c2}=\chi^2\left(\boldsymbol{t}^i,\boldsymbol{t}^j\right)$
	$\boldsymbol{s}$	semantic feature	4096	$d_{c3}=1-\cos\left(\boldsymbol{s}^i,\boldsymbol{s}^j\right)$
	$\boldsymbol{g}$	GIST feature	512	$d_{c4}=1-\cos\left(\boldsymbol{g}^i,\boldsymbol{g}^j\right)$
dep	$\boldsymbol{h}_d$	depth histogram	512	$d_d=\chi^2\left(\boldsymbol{h}_d^i,\boldsymbol{h}_d^j\right)$
sal	$\boldsymbol{h}_s$	saliency histogram	512	$d_s=\chi^2\left(\boldsymbol{h}_s^i,\boldsymbol{h}_s^j\right)$

Then, the feature distances between two images are summarized in the last column of Table 4-1, where $\chi^2(\cdot)$ represents the Chi-square distance, and $\cos(\cdot)$ denotes the cosine distance. Finally, these feature distances are fused to evaluate the image similarity as:

$$\varphi^{ij}=1-\left(\alpha_c\cdot\sum_{i=1}^{4}d_{ci}/4+\alpha_d\cdot d_d+\alpha_s\cdot d_s\right) \tag{4-5}$$

where φ^{ij} denotes the similarity measurement between images I^i and I^j, α_c, α_d, and α_s are the coefficients for color, depth, and saliency feature distances, respectively. A larger φ^{ij} corresponds to higher similarity between the two input images. The coefficients are set based on three criteria: (1) The sum of coefficients should be 1, as $\alpha_c+\alpha_d+\alpha_s=1$. (2) The color and saliency distances are assigned the same weight for simplicity. (3) The poor depth map, like a noise, may have a negative influence on the

measurement. Therefore, a self-adaptive weighted coefficient for depth distance is designed according to the depth confidence measure λ_d.

$$\alpha_d = \begin{cases} \lambda_d^{\min}, & if\ \lambda_d^{\min} = \min(\lambda_d^i, \lambda_d^j) \le T_2 \\ 1/3, & otherwise \end{cases} \tag{4-6}$$

and

$$\alpha_c = \alpha_s = \frac{1}{2} \cdot (1 - \alpha_d) \tag{4-7}$$

where T_2 is a threshold to distinguish the degenerated depth map, which is set to 0.2 in the experiments.

3) Inter saliency calculation

After obtaining the corresponding relationship among multiple images through the superpixel-level feature matching and image-level similarity matching, the inter saliency of a superpixel is computed as the weighted sum of the intra saliency of corresponding superpixels in other images. The superpixel-level feature matching result provides the corresponding relationship between the superpixels among different images, and the weighted coefficient is defined as the image-level similarity measurement.

With the matching matrix $\boldsymbol{ML}^{ij}$, image similarity φ^{ij}, and intra saliency map S_{intra}, the inter saliency value of each superpixel is assigned as:

$$S_{\text{inter}}\left(r_m^i\right) = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \frac{\varphi^{ij}}{N_j} \sum_{n=1}^{N_j} S_{\text{intra}}\left(r_n^j\right) \cdot ml\left(r_m^i, r_n^j\right) \tag{4-8}$$

where N represents the number of images in the group, N_j denotes the number of superpixels in image I^j, and φ^{ij} is the similarity measurement between images I^i and I^j.

4.2.3 Optimization and Propagation

In the proposed method, the optimization of saliency map is casted as a "label propagation" problem, where the uncertain labels are propagated by using two types of certain seeds, *i.e.*, background and salient seeds. The proposed CLP method is used to optimize the intra and inter saliency maps in a cross way, which means the propagative seeds are crosswise interacted. The cross seeding strategy optimizes the intra and inter saliency maps jointly, and improves the accuracy and robustness.

First, a graph for each image $G^i = \left(V^i, E^i\right)$ is constructed, where V^i represents the

set of nodes in image I^i which corresponds to the superpixels, and E^i denotes the set of links between adjacent nodes in image I^i.

The affinity matrix $\boldsymbol{W}^i = \left[w_{uv}^i\right]_{N_i \times N_i}$ is defined as the similarity between two adjacent superpixels in image I^i:

$$w_{uv}^i = \begin{cases} \exp\left(-\dfrac{\left\|\boldsymbol{c}_u^i - \boldsymbol{c}_v^i\right\|_2 + \lambda_d^i \cdot \left|d_u^i - d_v^i\right|}{\sigma^2}\right), & if\ v \in \Psi_u^i \\ 0, & otherwise \end{cases} \tag{4-9}$$

where Ψ_u^i is the set of neighbors of superpixel r_u^i.

Taking the optimization of intra saliency map as an example, the detailed procedures are described as follows. The certain seeds, including foreground labeled seeds F and background labeled seeds B, are selected to update and optimize the saliency of unlabeled nodes U. Two thresholds are designed to determine these labeled seeds.

$$TF(S) = \max\left(\frac{2}{N_i}\sum_{m=1}^{N_i}\left|S\left(r_m^i\right)\right|, T_{\min}\right) \tag{4-10}$$

$$TB(S) = \min\left(\frac{1}{N_i}\sum_{m=1}^{N_i}\left|S\left(r_m^i\right)\right|, T_{\max}\right) \tag{4-11}$$

where $S\left(r_m^i\right)$ denotes the intra or inter saliency score of the superpixel r_m^i, $TF(S)$ is a threshold of saliency map S for foreground seeds selection, $T_{\min}$ is the minimum threshold for $TF(S)$, $TB(S)$ is a threshold for background seeds selection, and $T_{\max}$ is the maximum threshold for $TB(S)$. Then, these thresholds are used to determine a set of labeled seeds and initialize the saliency score of superpixels. The saliency scores of superpixels in CLP method are initialized as follows.

$$V_0^{CLP}\left(r_m^i\right) = \begin{cases} 1, & if\ S_{\text{inter}}\left(r_m^i\right) \geq TF\left(S_{\text{inter}}\right) \\ 0, & if\ S_{\text{inter}}\left(r_m^i\right) \leq TB\left(S_{\text{inter}}\right) \\ S_{\text{intra}}\left(r_m^i\right), & otherwise \end{cases}. \tag{4-12}$$

Once the initialization is completed, the saliency is propagated using the labeled seeds on the graph according to the equation:

$$V_{\text{intra}}^{CLP}\left(r_m^i\right) = \sum_{n=1}^{N_i} w_{mn}^i V_0^{CLP}\left(r_n^i\right). \tag{4-13}$$

After normalization, the optimized intra saliency map $S_{\text{intra}}^{CLP} = norm\left(V_{\text{intra}}^{CLP}\right)$ is achieved, where $norm(\cdot)$ is the min-max normalization function. The optimized inter saliency map S_{inter}^{CLP} is generated using the same procedure conducted on the inter

saliency map. It should be noted that the inter saliency map is firstly optimized using the intra saliency map in CLP method. Then, the optimized inter saliency map is used to update the intra saliency map, since the inter saliency is generated by the intra saliency map. In order to guarantee the optimization performance, the inter saliency map is firstly optimized in the proposed model.

4.2.4 Co-saliency Detection

Finally, the initial intra/inter saliency maps and the optimized intra/inter saliency maps are integrated to generate the final co-saliency map.

$$S_{co}^{CLP} = \gamma_1 \cdot S_{\text{intra}} + \gamma_2 \cdot S_{\text{inter}} + \gamma_3 \cdot S_{\text{intra}}^{CLP} + \gamma_4 \cdot S_{\text{inter}}^{CLP} \tag{4-14}$$

where γ_i is the weighted coefficient with $\sum_{i=1}^{4} \gamma_i = 1$. Without loss of generality, these four parameters are all set to 0.25 in experiments.

4.3 Experimental Results

First, the experimental settings including the datasets, implementation details, and evaluation metrics are introduced. Then, the qualitative and quantitative comparisons are presented in Section 4.3.2. The evaluation of different parameters are analyzed in Sections 4.3.3-4.3.5. At last, some degenerated cases are discussed in Section 4.3.6.

4.3.1 Experimental Settings

The proposed co-saliency model is evaluated on the RGBD Coseg183 dataset [88] and RGBD Cosal150 dataset [114] by using four criteria including the PR curve, the Precision and Recall scores, F-measure, and Mean Absolute Error (MAE) are calculated. In the proposed method, the number of superpixels is set to 200, the number of clusters K used in K-means++ is set to 10, the maximum matching number $K_{\max}$ for feature matching is set to 40, and the thresholds for seeds selection are assigned to $T_{\min} = 0.6$ and $T_{\max} = 0.2$. The proposed method is implemented in MATLAB 2014a, and all the experiments are performed on a Quad Core 3.5GHz workstation with 16GB RAM. The proposed algorithm costs average 41.03 seconds to process one image on the RGBD Cosal150 dataset. According to statistics, the intra saliency calculation costs 27.86%

running time, the inter saliency model takes 71.94% running time, and the optimization stage consumes 0.20% running time.

4.3.2 Comparison with State-of-the-art Methods

The proposed method is compared with some state-of-the-art saliency/co-saliency detection methods, *i.e.*, RC [25], HS [26], BSCA [30], DRFI [41], ACSD [65], DCMC [130], SCS [77], CCS [79], and LRMF [81], in which the first six methods are single saliency detection models and the last three ones are co-saliency methods for RGB image. In addition, other two optional optimization mechanisms, namely Label Propagation (LP) and Shared Label Propagation (SLP), are reported as the baselines in the experiments. The main difference of these three methods lies in the selection of certain seeds. The LP scheme determines the seeds from their own intra or inter saliency map. For the SLP mechanism, the intra and inter saliency maps are combined to determine the common seeds. Then, the selected seeds are shared to optimize the intra and inter saliency maps simultaneously. By contrast, the propagative seeds are crosswise interacted for the CLP scheme, and it bridges the gap between the intra saliency and inter saliency in the process of optimization. Therefore, in principle, the CLP scheme is more suitable for co-saliency detection due to the interactive information from intra and inter saliency maps.

Some visual comparisons of different methods on two datasets are illustrated in Fig. 4-2, which contain five image groups: woman, sculpture, car, yellow flashlight, and computer. From the figure, even though the images own complex and variable backgrounds or the salient objects exhibit large variations in shape and direction, the proposed method effectively highlights the common salient objects from the image group. Furthermore, the results produced by our model are more accurate and uniform than other methods. For example, in the woman group, the eyes and mouth of the woman are wrongly detected as background regions through the SCS model [77], and some background regions (*e.g.*, the sheds) are also detected as salient regions due to their complex textures. The same situation is faced to the LRMF method [81], where the body of the woman is missed and the background regions are wrongly detected. By contrast, the woman in different images are uniformly detected by the proposed method with clearly contour, and the background regions are effectively suppressed.

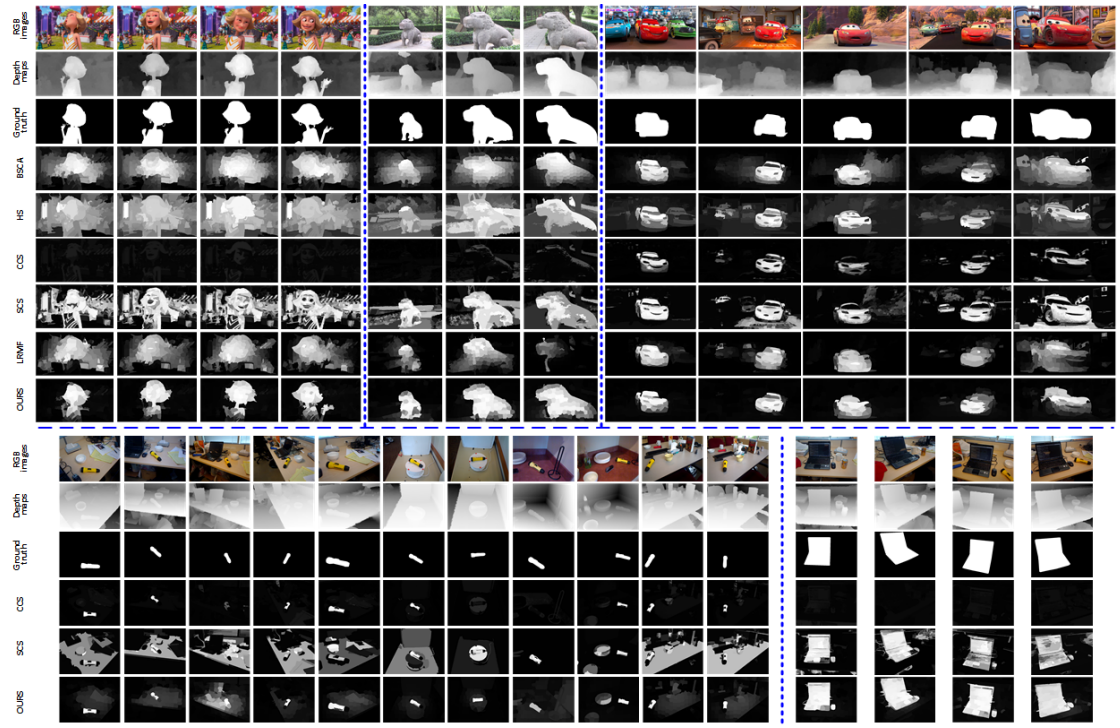

Fig. 4-2 Visual examples of different saliency and co-saliency detection methods on two datasets.

The quantitative comparison results in terms of the PR curves, precision and recall scores, F-measure, and MAE scores are reported in Fig. 4-3. Before comparing with other methods, the results of the proposed method in different stages are first analyzed, which include intra and inter saliency modeling, and co-saliency generation with three different optimization schemes (*i.e.*, LP, SLP, and CLP). It can be observed that (1) the inter saliency map performs a better result compared with other existing co-saliency methods, and (2) the performance of co-saliency result with the optimization model is obviously improved compared with the intra and inter saliency maps individually. Moreover, consistent with the theoretical analysis, the CLP scheme achieves favorable performance on the two datasets. For example, the proposed method with CLP optimization achieves the best performance on the RGBD Cosal150 dataset according to the comprehensive measures (F-measure = 0.8403, and MAE = 0.1370), and it also performs the best result on RGBD Coseg183 dataset in terms of MAE measure (F-measure = 0.6365, and MAE = 0.0979). For the SLP optimization strategy, its performance is slightly worse than other methods, where the MAE score is 0.1430 on the RGBD Cosal150 dataset, and 0.1052 on the RGBD Coseg183 dataset. The main reason is that the shared way for seed selection enables reduction in the number of seeds, which in turn degrades the propagation performance.

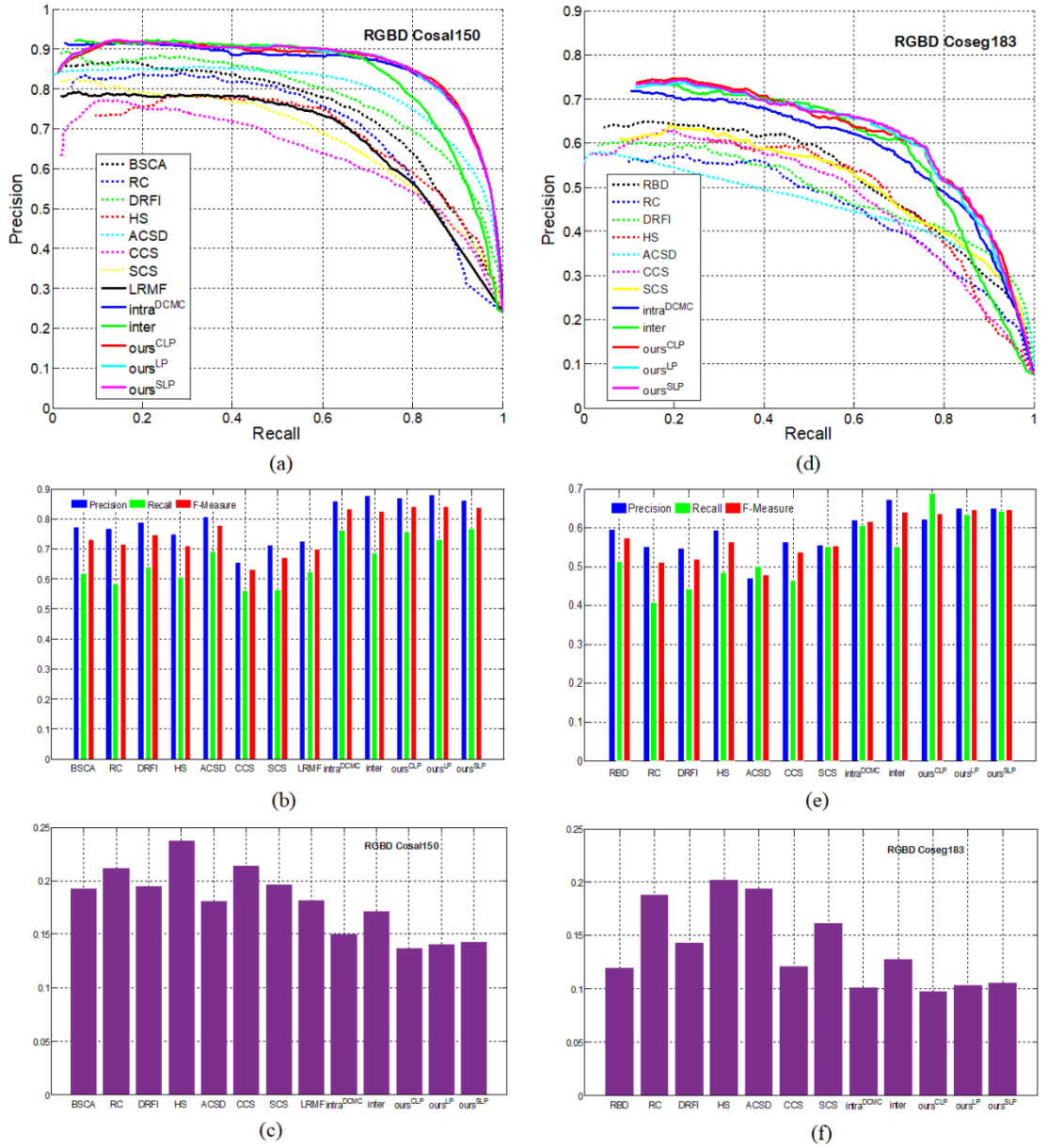

Fig. 4-3 Quantitative performances of different methods on two datasets. Notice that "our*" means implementing our method using different optimization approaches, where *={LP, SLP, CLP}. (a)-(c) PR curves, Precision and Recall scores, F-measure, and MAE scores on the RGBD Cosal150 dataset. (d)-(f) PR curves, Precision and Recall scores, F-measure, and MAE scores on the RGBD Coseg183 dataset.

Compared with other single saliency and co-saliency methods, the proposed model achieves the highest precisions of the whole PR curves on both of the RGBD Coal150 and Coseg183 datasets. In addition, the proposed co-saliency model achieves the best performance on both two datasets with the highest F-measure and the smallest MAE score. In terms of the F-measure, the proposed co-saliency model achieves a maximum percentage gain of 30.67% compared to other saliency results on the RGBD Cosal150 dataset, and the minimum percentage gain also reaches 5.88%. The maximum and

minimum percentage gains of the MAE score achieve 46.00% and 24.14%, respectively. On the RGBD Coseg183 dataset, the proposed method also obtains obvious performance gains. For example, the F-measure and MAE score are at least increased by 10.67% and 18.21%, respectively.

In summary, benefiting from the two-level similarity matching and CLP optimization, the proposed co-saliency detection model achieves superior performance. The visual comparisons and quantitative analyses demonstrate the effectiveness of the proposed model. The influence of some parameters will be discussed in the next subsections.

4.3.3 Evaluation of the Maximum Matching Number

Some experiments on the RGBD Cosal150 dataset are conducted to analyze the influence of the maximum matching number $K_{\max}$ in the procedure of inter saliency calculation. Considering the parameter $K_{\max}$ is only used to calculate the inter saliency map, the F-measures of the inter saliency maps are evaluated under different $K_{\max}$, as shown in Fig. 4-4. From the curve of F-measure with different $K_{\max}$, the inter saliency maps with different maximum matching numbers achieve the comparable performance except for the result with $K_{\max}=10$. The main reason is that $K_{\max}=10$ is too small to obtain a relatively stable and accurate matching result. The performance will become stable when the $K_{\max}$ reaches 30. Considering the number of superpixel in an image is set to 200 in the experiments, the maximum matching number $K_{\max}$ is set to 40 for balancing the computational complexity and accuracy. In conclusion, the performance of inter saliency map is not highly sensitive to the parameter $K_{\max}$. In general, due to the saliency consistency and cluster-based constraint are introduced in our model, the maximum matching number between superpixels among two images in the process of feature matching can be set to a larger value, such as the one-fifth of the superpixel number in an image.

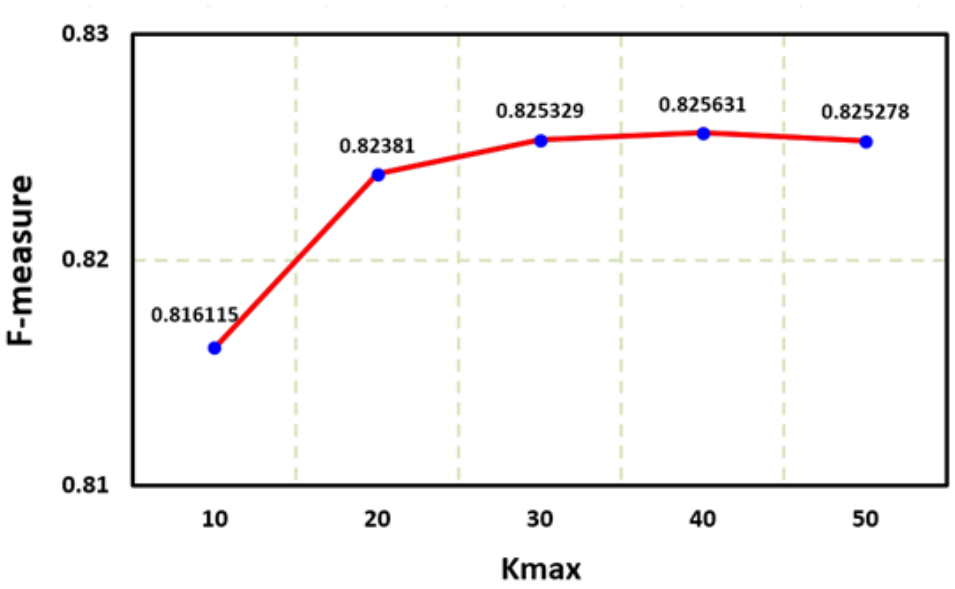

Fig. 4-4 F-measure of inter saliency map on the RGBD Cosal150 dataset under different maximum matching numbers.

4.3.4 Evaluation of the Depth Cue

In the proposed model, the depth cue is introduced to assist the identification of co-salient regions. To evaluate the effect of depth cue on the whole framework, an experiment on the RGBD Cosal150 dataset is conducted, and the results are shown in Fig. 4-5 and Table 4-2.

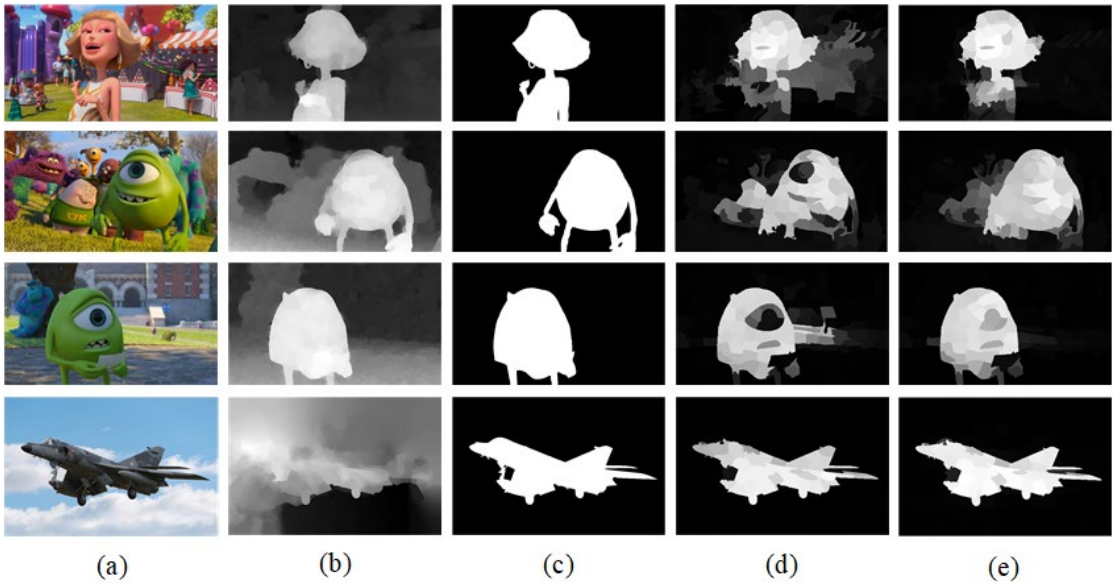

Fig. 4-5 Visual examples with and without depth information through the proposed co-saliency detection method. (a) RGB images. (b) Depth maps. (c) Ground truth. (d) Saliency maps without depth information. (e) Saliency maps with depth information.

For the depth map, in most cases, it is clear and effective, and exhibits great power in improving the saliency performance, such as the first three rows in Fig. 4-5. The depth information can be regarded as an effective cue to distinguish the foreground from the complex background. Therefore, utilizing the depth cue, the complex and cluttered background regions (such as the stores, other cartoons, and lawns) are suppressed obviously, and the salient objects are better highlighted. However, in some cases, the depth map has poor quality and may degrade the performance. To address this problem, the depth confidence measure is introduced as a weight to control the contribution of depth information. In this poor quality case, as shown in the last row of Fig. 4-5, even with the noisy depth information, the proposed model still achieves relatively satisfying performance similar to the RGB co-saliency model. From the quantitative measures reported in Table 4-2, without the depth cue, the F-measure of the proposed model achieves 0.7639, and the MAE score reaches 0.1599. With the depth cue, the overall performance of F-measure is increased to 0.8403, and the MAE score is also improved to 0.1370. In summary, all the experiments demonstrate that the using of the depth information in the proposed model is useful and effective.

Table 4-2 Quantitative evaluations with and without depth cue through the proposed co-saliency detection method on the RGBD Cosal150 dataset.

	without depth	with depth	percentage gain
F-measure	0.7639	0.8403	10.00%
MAE	0.1599	0.1370	14.32%

4.3.5 Evaluation of the Different Intra Saliency Methods

The proposed method focuses on designing an opened framework to make the existing saliency detection methods work well in co-saliency scenarios. Therefore, the experiments on the RGBD Cosal150 dataset are conducted to evaluate the performance with different intra saliency initializations. In the experiment, five different single saliency maps produced by HS [26], BSCA [30], DRFI [41], ACSD [65], and DCMC [130] are used as the intra saliency maps. The PR curves are illustrated in Fig. 4-6, and some quantitative measures, including the F-measure and MAE score, are reported in Table 4-3.

In Fig. 4-6, the PR curves of the results using the proposed co-saliency framework

are higher than those from the original saliency maps. The consistent conclusion can be drawn from the quantitative comparisons listed in Table 4-3. The F-measure is improved by the proposed co-saliency model, and the MAE scores also achieves better performances compared with the previous saliency results. Taking the F-measure as an example, the proposed co-saliency model achieves a maximum percentage gain of 4.41% compared to the corresponding intra saliency result, and the average percentage gain also reaches 2.59%. Similarly, the maximum percentage gains of the MAE scores achieves 15.75%. Moreover, in general, the better the single saliency map (intra saliency map) achieves, the better performance of the co-saliency map is. In brief, the results demonstrate that the proposed model can effectively improve the performance of the existing single saliency models, and make them work well for co-saliency detection.

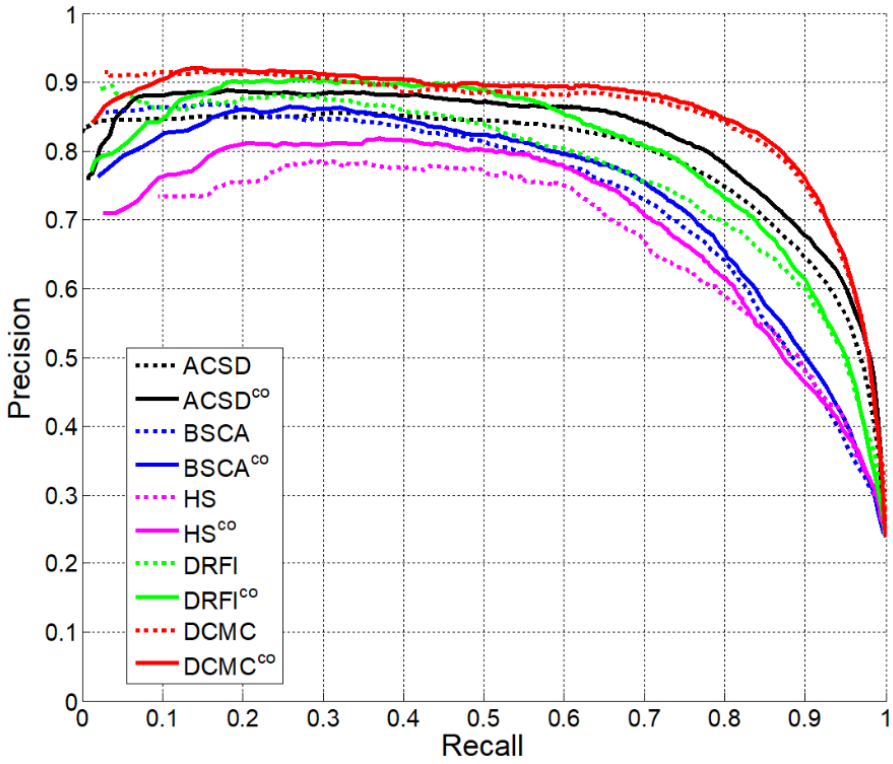

Fig. 4-6 PR curves of co-saliency results on the RGBD Cosal150 dataset with different intra saliency maps. The subscript of "co" means the co-saliency result produced by the proposed co-saliency model with the CLP optimization and corresponding intra saliency initialization.

Table 4-3 Quantitative evaluations of co-saliency results with different intra saliency models on the RGBD Cosal150 dataset.

		F-measure	MAE
ACSD	intra saliency	0.7788	0.1806
	our co-saliency	**0.8039**	**0.1529**
BSCA	intra saliency	0.7318	0.1925
	our co-saliency	**0.7470**	**0.1731**
HS	intra saliency	0.7101	0.2375
	our co-saliency	**0.7283**	**0.2063**
DRFI	intra saliency	0.7484	0.1949
	our co-saliency	**0.7814**	**0.1642**
DCMC	intra saliency	0.8348	0.1498
	our co-saliency	**0.8403**	**0.1370**

4.3.6 Discussion

Some challenging cases of the proposed RGBD co-saliency model are shown in Fig. 4-7. For the bike group, the salient foreground is very trivial and includes lots of stuff regions, such as the spokes and back seat. These regions are difficult to detect completely and accurately through the proposed co-saliency model. For the soda can group, the scene is relatively complex and cluttered, and the soda can is too small to be detected as the salient object compared with the computer in each image. Thus, the small scale object is not detected successfully through the proposed model, especially in the complex and cluttered scenes.

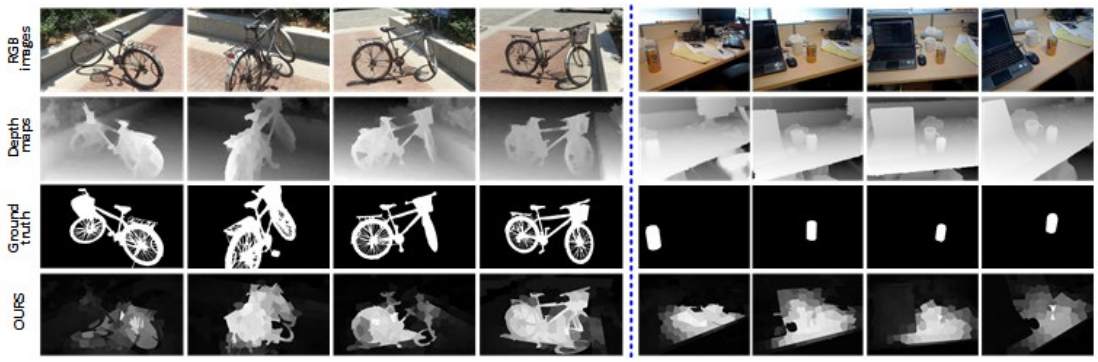

Fig. 4-7 Some challenging examples for the proposed RGBD co-saliency detection model. The left group contains the bike as the common salient object, and the right group includes the orange can

as the common salient object.

4.4 Summary

In this chapter, a co-saliency detection model for RGBD images was presented, which focuses on exploring the inter-image correspondence constraint among multiple images and introducing the depth cue to enhance the identification of co-salient objects. The similarity constraint, saliency consistency, and cluster-based constraint were introduced in feature similarity matching to obtain more stable and accurate corresponding relationship at superpixel level. The image-level similarity descriptor was designed as the weighted coefficient for inter saliency calculation. In addition, a CLP optimization strategy was proposed to optimize the intra and inter saliency maps in a cross way. The comprehensive comparisons and ablation studies on two RGBD co-saliency detection datasets have demonstrated that the proposed method outperforms other state-of-the-art saliency and co-saliency models, and verified the effectiveness of improving the existing saliency models in co-saliency application.

Chapter 5 An Iterative Co-saliency Framework for RGBD Images

The existing co-saliency detection methods often generate the co-saliency map through a direct forward pipeline which is based on the designed cues or initialization, but lack the refinement-cycle scheme. In this chapter, making full use of depth information, an iterative RGBD co-saliency framework is proposed, which utilizes the existing single saliency maps as the initialization, and generates the final RGBD co-saliency map by using a refinement-cycle model. Three schemes are employed in the proposed RGBD co-saliency framework, which include the addition scheme, deletion scheme, and iteration scheme. The addition scheme is used to highlight the salient regions based on intra-image depth propagation and saliency propagation. The deletion scheme captures the inter-image constraint to suppress the background regions and filter the non-common salient regions. The iteration scheme is proposed to obtain more homogeneous and consistent co-saliency map in a cycle way. Note that, a novel descriptor, named depth shape prior, is proposed in the addition scheme to introduce the depth information to enhance identification of co-salient objects. The proposed method can effectively exploit any existing 2D saliency model to work well in RGBD co-saliency scenarios. The comprehensive experiments on two RGBD co-saliency datasets demonstrate the effectiveness of the proposed framework.

5.1 Introduction

The co-salient object in an image group should satisfy two properties simultaneously, *i.e.*, (1) the objects should be salient in each individual image, and (2) the objects should repeatedly appear in most of the images. In other words, inter-image correspondence plays more important role in co-saliency detection [24]. For the first property of co-saliency detection, the single saliency map produced by the existing saliency model can be directly considered as an initialization without the need to design a new algorithm. Moreover, the existing co-saliency detection methods mainly rely on

the designed cues or initialization, and lack the refinement-cycle. In this chapter, an effective refinement-cycle framework for RGBD co-saliency detection is proposed, which integrates the addition scheme, deletion scheme, and iteration scheme. The addition scheme is used to enrich the saliency regions through the depth and saliency propagations. Furthermore, the depth information from the RGBD images has been demonstrated the power and usefulness for many computer vision tasks [3,7,12,13,58,62,65]. However, in the existing methods, the information of the relevant and similar objects in a sequence of images is ignored and not exploited. In the addition scheme, a novel depth descriptor, named depth shape prior, is proposed to capture the shape attributes from the depth map to improve the co-saliency detection performance. In the deletion scheme, the inter saliency model is formalized as common probability function to capture the inter-image correspondence. The iterative optimization scheme is designed to achieve more superior co-saliency result in a cycle way.

In summary, most of the existing co-saliency methods aim to design a single forward pipeline which generates the co-saliency map based on the designed cues directly, but lack the refinement-cycle scheme and ignore the depth information for RGBD images. Therefore, an iterative RGBD co-saliency framework is proposed, which utilizes the additional depth information and employs the existing RGB saliency map as the initialization in a refinement-cycle model to produce the final RGBD co-saliency map.

5.2 Proposed Iterative Framework

Fig. 5-1 shows the proposed RGBD co-saliency framework framework. The proposed method is firstly initialized by the existing 2D saliency maps, and then three schemes are employed to generate the final RGBD co-saliency map. The addition scheme is used to grow the initialized saliency map from the perspective of intra-image, the deletion scheme is designed to suppress the non-common regions from the perspective of inter-image, and the iteration scheme is exploited to obtain more homogeneous and consistent co-saliency map in a cycle way.

Notations: Given N input RGB images $\{I^i\}_{i=1}^{N}$ and the corresponding depth maps $\{D^i\}_{i=1}^{N}$. The M_i single saliency maps for image I^i produced by the existing single image saliency models are represented as $S^i=\{S_j^i\}_{j=1}^{M_i}$. In our method, the

superpixel-level region is regarded as the basic unit for processing. Thus, each RGB image I^i is firstly abstracted into superpixels $R^i = \{r_m^i\}_{m=1}^{N_i}$ by using the SLIC algorithm [119], where N_i is the number of superpixels in image I^i.

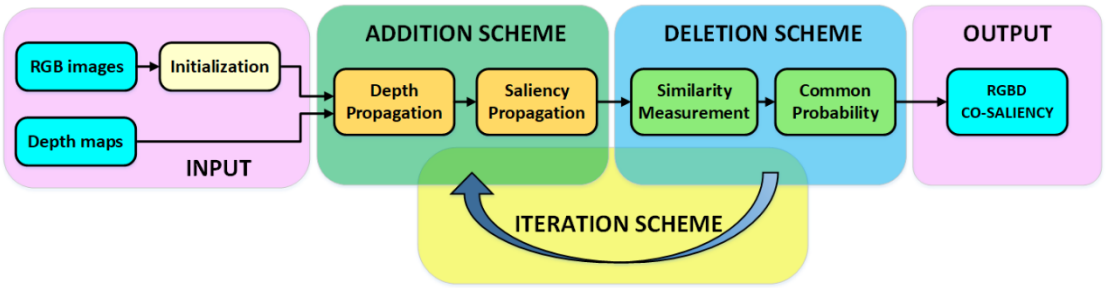

Fig. 5-1 Flowchart of the proposed RGBD co-saliency framework.

5.2.1 Initialization

The proposed co-saliency framework aims at discovering the co-salient objects from multiple images in a group with the assistance of some existing 2D saliency maps. Therefore, the framework is initialized by some existing saliency maps produced by the 2D saliency models. It is well known that different saliency methods own different superiority in detecting salient regions. In a way, these saliency maps are complementary in some regions. Thus, the fused result can inherit the merits of the multiple saliency maps, and produce more robust and superior detection baseline. In our method, the simple average function is used to generate a more generalized initialization result. The initialized saliency map for image I^i is denoted as:

$$S_f^i\left(r_m^i\right) = \frac{1}{M_i}\sum_{j=1}^{M_i} S_j^i\left(r_m^i\right) \tag{5-1}$$

where $S_j^i\left(r_m^i\right)$ denotes the saliency value of superpixel r_m^i produced by j^{th} saliency method for image I^i, and M_i is the number of saliency maps for image I^i. In the experiments, five saliency methods, including RC [25], HS [26], BSCA [30], RRWR [31], and DCLC [33], are used to produce the initialized saliency map. Some examples of the initialized saliency map are shown in Fig. 5-2(c). From the figure, the initialized result provides an impressive baseline for later co-saliency detection.

Fig. 5-2 Some examples of the proposed method. (a) RGB image. (b) Depth map. (c) The initialized saliency map. (d) The co-saliency map without iteration. (e) The final co-saliency map with iteration.

5.2.2 Addition Scheme

The addition scheme is designed to extend the saliency region based on the intra-image constraint with two propagation strategies. First, a novel depth descriptor, named depth shape prior, is proposed to deeply capture the depth cue and produce an RGBD saliency result in depth propagation. Then, saliency propagation is utilized to further optimize and improve the saliency result.

1) Depth propagation

After initialization, the merits of the different saliency maps are inherited into the initialized saliency map. The depth information is introduced into the framework to enhance the identification of salient objects due to its usefulness in saliency detection. In general, the depth map owns the following properties:

(i) The salient object appears higher depth value compared to the backgrounds.

(ii) The high quality depth map can provide sharp and explicit object boundary.

(iii) The interior depth value of the object should be smooth and consistency.

Inspired by these observations, a depth descriptor, namely depth shape prior (DSP), is proposed to capture the shape attributes from the depth map and improve the performance of the co-saliency detection by using the depth consistency and shape attributes. The proposed DSP descriptor is based on depth propagation and region grow. Several identified superpixels are selected as the seeds first, and then the DSP map can be calculated via depth constraints.

For each image I^i, the top K superpixels with higher initialized saliency values are selected as the root seeds, which are represented as $\{r_{rk}^i\}_{k=1}^K$, and the corresponding DSP map DSP_k^i is initialized as zero.

For each root seed, a set of child nodes $\{r_{cp}^i\}$ are determined to depict the depth

shape based on the depth smoothness and consistency constraints. In the l-loop diffusion, the superpixels direct neighboring the $(l-1)$-loop child nodes are selected as the l-loop child nodes only if they satisfy the following two constraints:

(a) Depth smoothness: the depth difference between the neighbor superpixel and $(l-1)$-loop child seeds is less than a certain threshold T_1, as $\left|d_{nq}^i - d_{c,l-1}^i\right| \le T_1$, where d_{nq}^i is the depth value of the neighbor superpixel r_{nq}^i, and $d_{c,l-1}^i$ is the average depth value of $(l-1)$-loop child seeds.

(b) Depth consistency: the depth difference between the neighbor superpixel and root seed should be smaller than a specific threshold T_2, as $\left|d_{nq}^i - d_{rk}^i\right| \le T_2$, where d_{rk}^i is the depth value of the root seed r_{rk}^i.

Be noted that the child node in the first loop diffusion is initialized by the root seed in the proposed method, and the two thresholds are set to 0.1 and 0.2, respectively. The DSP value of the child node r_{cp}^i in the l-loop is defined as:

$$DSP_k^i\left(r_{cp}^i\right) = 1 - \min\left(\left|d_{cp,l}^i - d_{c,l-1}^i\right|, \left|d_{cp,l}^i - d_{rk}^i\right|\right) \tag{5-2}$$

where $d_{cp,l}^i$ denotes the depth value of the child node r_{cp}^i in the l-loop, $d_{c,l-1}^i$ is the average depth value of $(l-1)$-loop child node set, and $|\cdot|$ represents the absolute value function. Then, the next loop diffusion will be continued until there is no neighboring superpixel satisfies the depth constraints.

In the proposed method, the top K root seeds are selected for each image I^i to improve the robustness, and K DSP maps are obtained for each image. Therefore, the final DSP map is defined as:

$$DSP^i\left(r_m^i\right) = \frac{1}{K}\sum_{k=1}^{K} DSP_k^i\left(r_m^i\right) \tag{5-3}$$

where K is the number of the root seeds, which is fixed to 10 in all the experiments.

To achieve more superior and stable saliency result, the initialized RGB saliency and the DSP map are combined. Because the bad depth map may degenerate the accuracy of DSP map generation, the depth confidence measure λ_d is introduced to evaluate the quality of the depth information. Thus, the RGBD saliency that integrates initialized saliency map and DSP map weighted by depth confidence measure according to the depth quality is defined as:

$$S_{dp}^i\left(r_m^i\right) = \left(1-\lambda_d^i\right)\cdot S_f^i\left(r_m^i\right) + \lambda_d^i \cdot S_f^i\left(r_m^i\right)\cdot DSP^i\left(r_m^i\right) \tag{5-4}$$

where λ_d^i is the depth confidence measure for image I^i, $S_f^i\left(r_m^i\right)$ represents the initialized saliency value of superpixel r_m^i, and $DSP^i\left(r_m^i\right)$ denotes the DSP value of

superpixel r_m^i. The obtained saliency map is normalized into [0,1]. With this depth confidence measure, the poor-quality depth map will be limited in the combination with RGB feature to avoid the degradation of the RGBD co-saliency result. Fig. 5-3 shows some examples of the depth propagation. Comparing with the RGB saliency map, some background regions around the salient object are effectively suppressed through depth propagation, such as the lawns in the first image, the roads in the second image, and the buildings in the third image. Moreover, the RGBD saliency model is more robust. Even if the quality of depth map is bad, such as the last raw in Fig. 5-3, the proposed model still achieves better result by highlighting the RGB saliency component while DSP descriptor can not exploit accurate shape attributes from the poor-quality depth map.

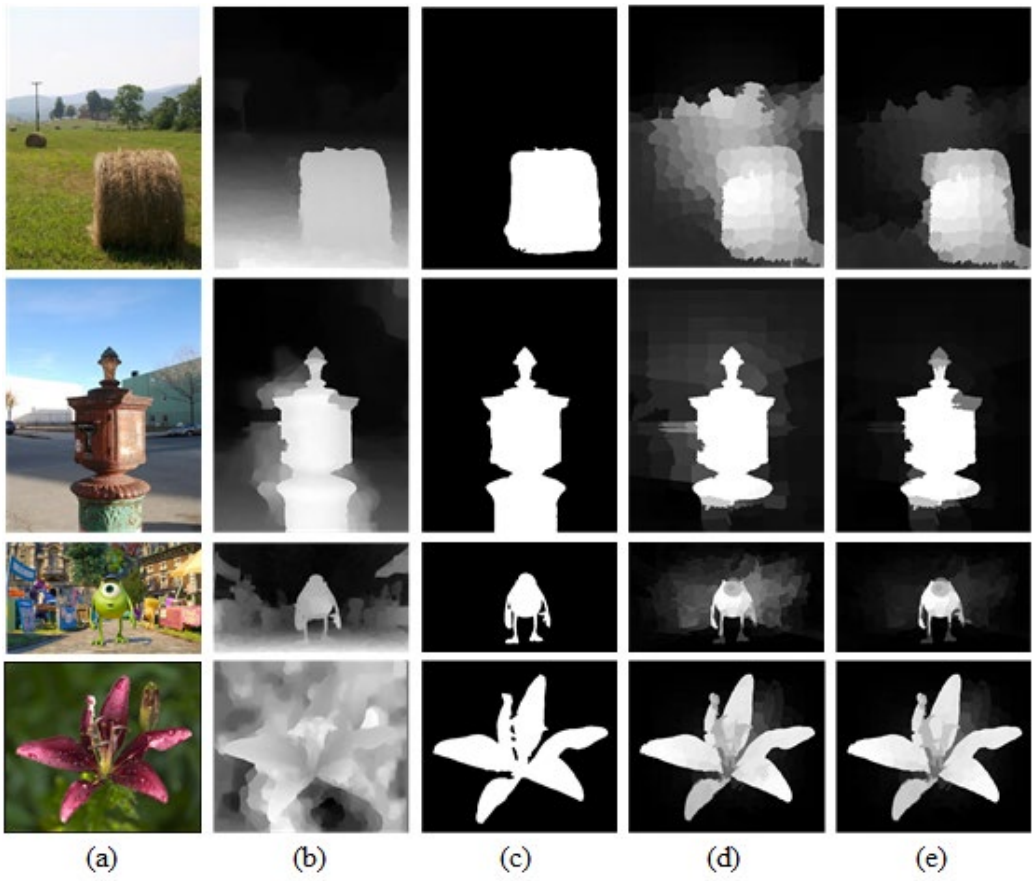

Fig. 5-3 Examples of the depth propagation. (a) RGB image. (b) Depth map. (c) Ground truth. (d) The RGB saliency result. (e) The RGBD saliency result with DSP descriptor.

2) Saliency propagation

With the obtained RGBD saliency map, the saliency propagation is conducted to further optimize the result. First, the superpixels are classified into three groups based on the saliency value, which is denoted as the saliency seed superpixels, background

seed superpixels, and the unknown superpixels. Then, saliency propagation is used to propagate the saliency of unknown superpixels on the graph from the saliency and background seeds.

For image I^i, a graph $G^i=(\upsilon^i,\varepsilon^i)$ among superpixels is constructed, where υ^i denotes the node set corresponding to the superpixels, and ε^i is the link set among the adjacent nodes. The affinity matrix $\boldsymbol{W}^i=\left[w_{uv}^i\right]_{N_i\times N_i}$ is defined as the similarity between two adjacent superpixels:

$$w_{uv}^i=\begin{cases}\exp\left(-\dfrac{\left\|\boldsymbol{c}_u^i-\boldsymbol{c}_v^i\right\|_2+\lambda_d^i\cdot\left|d_u^i-d_v^i\right|}{\sigma^2}\right), & if\ r_v^i\in\Omega_u^i\\ 0, & otherwise\end{cases} \tag{5-5}$$

where $\boldsymbol{c}_u^i$ and d_u^i denote the mean Lab color vector and depth value of the superpixel r_u^i, Ω_u^i represents the neighbor set of superpixel r_u^i, $\|\cdot\|_2$ is the 2-norm of vector, λ_d^i denotes the depth confidence measure, and σ^2 is a constant control parameter.

In the proposed method, the seed superpixels are selected based on RGBD saliency value produced by depth propagation. Then, the top κ superpixels with higher saliency values are considered as the saliency seeds, and the bottom κ superpixels with lower saliency values are treated as the background seeds. In the experiments, κ is set to 10. The initialized propagation score of the superpixel is defined as:

$$S_0^i\left(r_n^i\right)=\begin{cases}1, & if\ r_n^i\in\Psi_F\\ 0, & if\ r_n^i\in\Psi_B\\ S_{dp}^i\left(r_n^i\right), & otherwise\end{cases} \tag{5-6}$$

where Ψ_F and Ψ_B represent the saliency and background seed sets, respectively.

Using the labeled seeds, the saliency is propagated on the graph, and the score with saliency propagation is achieved by:

$$S_{sp}^i\left(r_m^i\right)=\sum_{n=1}^{N_i}w_{mn}^iS_0^i\left(r_n^i\right) \tag{5-7}$$

where w_{mn}^i is the element of the affinity matrix.

5.2.3 Deletion Scheme

The addition scheme is used to improve and optimize the saliency map from the perspective of intra-image. On the other hand, the inter-image information should be captured to determine the common attribute of the objects. Therefore, a deletion scheme is designed to explore the corresponding relationship among multiple images, which

aims to suppress the common and non-common backgrounds, and highlight the common salient regions from the perspective of multiple images. In the deletion scheme, a superpixel-level similarity measurement is constructed to represent the similarity relationship between two superpixels. Then, a common probability function using the similarity measurement is defined to calculate the likelihood of each superpixel belonging to the common regions.

1) Multiple cues based similarity measurement

In the deletion scheme, the color, depth, and saliency cues are combined into a measurement to evaluate the similarity between two superpixels.

RGB similarity. The color histogram and texture histogram [131,132] are used to represent the RGB feature on the superpixel level, which are denoted as HC_m^i and HT_m^i, respectively. Then, the Chi-square measure is employed to compute the feature difference. Thus, the RGB similarity is defined as:

$$s_c\left(r_m^i, r_n^j\right) = 1 - \frac{\chi^2\left(HC_m^i, HC_n^j\right) + \chi^2\left(HT_m^i, HT_n^j\right)}{2} \tag{5-8}$$

where r_m^i and r_n^j are the superpixels in image I^i and I^j, respectively, and $\chi^2(\cdot)$ denotes the Chi-square distance function.

Depth similarity. Two depth consistency measurements, namely depth value consistency and depth contrast consistency, are composed of the final depth similarity measurement, which is defined as:

$$S_d\left(r_m^i, r_n^j\right) = \exp\left(-\frac{W_d\left(r_m^i, r_n^j\right) + W_c\left(r_m^i, r_n^j\right)}{\sigma^2}\right) \tag{5-9}$$

where $W_d\left(r_m^i, r_n^j\right)$ is the depth value consistency measurement to evaluate the inter image depth consistency, due to the fact that the common regions should appear similar depth values

$$W_d\left(r_m^i, r_n^j\right) = \left|d_m^i - d_n^j\right|. \tag{5-10}$$

$W_c\left(r_m^i, r_n^j\right)$ describes the depth contrast consistency, because the common regions should represent more similar characteristic in depth contrast measurement.

$$W_c\left(r_m^i, r_n^j\right) = \left|D_c\left(r_m^i\right) - D_c\left(r_n^j\right)\right| \tag{5-11}$$

with

$$D_c\left(r_m^i\right) = \sum_{k \neq m} \left|d_m^i - d_k^i\right| e^{-\left\|\boldsymbol{p}_m^i - \boldsymbol{p}_k^i\right\|_2 / \sigma^2} \tag{5-12}$$

where $D_c\left(r_m^i\right)$ denotes the depth contrast of superpixel r_m^i, $\boldsymbol{p}_m^i$ represents the position

of superpixel r_m^i, and σ^2 is a constant.

Saliency similarity. Inspired by the prior that the common regions should appear more similar in single saliency map compared to other regions, the output saliency map from the addition scheme is used to define the saliency similarity measurement in this work:

$$s_s\left(r_m^i, r_n^j\right) = \exp\left(-\left|S_{sp}^i\left(r_m^i\right) - S_{sp}^j\left(r_n^j\right)\right|\right) \tag{5-13}$$

where $S_{sp}^i\left(r_m^i\right)$ is the saliency score of superpixel r_m^i based on Eq. (5-7).

Combination similarity. Based on these cues, the combination similarity measurement is defined as the average of the three similarity measurements.

$$s_M\left(r_m^i, r_n^j\right) = \frac{s_c\left(r_m^i, r_n^j\right) + s_d\left(r_m^i, r_n^j\right) + s_s\left(r_m^i, r_n^j\right)}{3} \tag{5-14}$$

where $s_c\left(r_m^i, r_n^j\right)$, $s_d\left(r_m^i, r_n^j\right)$, and $s_s\left(r_m^i, r_n^j\right)$ are the normalized RGB, depth, and saliency similarities between superpixel r_m^i and r_n^j, respectively. A larger $s_M\left(r_m^i, r_n^j\right)$ value corresponds to greater similarity between two superpixels.

2) Common probability

For co-saliency detection, it is necessary to discriminate whether the selected salient objects are common or not. The common object means the object with repeated occurrence in multiple images. Based on this definition, the common probability function is used to evaluate the likelihood that a superpixel belongs to the common regions, and it is defined as the sum of maximum matching probability among different images. For each superpixel r_m^i, only the most matching superpixel r_k^j in image I^j is selected for calculation, which is denoted as:

$$r_k^j = \arg\max_{n \in [1, N_j]} S_M\left(r_m^i, r_n^j\right) \tag{5-15}$$

where r_k^j is the most matching/similar superpixel in image I^j for superpixel r_m^i based on the maximum combination similarity score, and N_j represents the number of superpixles in image I^j.

Then, these selected superpixels from different images are used to calculate the common probability:

$$P_c^i\left(r_m^i\right) = \frac{1}{N-1}\sum_{j=1, j\neq i}^{N} S_M\left(r_m^i, r_k^j\right) \tag{5-16}$$

where r_k^j is the most matching superpixel in image I^j for superpixel r_m^i, and N denotes the number of images in an image group. Finally, the updated co-saliency map

of deletion scheme is denoted as:

$$S_{del}^{i}\left(r_{m}^{i}\right)=S_{sp}^{i}\left(r_{m}^{i}\right)\cdot P_{c}^{i}\left(r_{m}^{i}\right) \tag{5-17}$$

where $S_{sp}^{i}\left(r_{m}^{i}\right)$ is the saliency score of superpixel r_{m}^{i} produced by the addition scheme. Fig. 5-3(d) shows the co-saliency map after addition and deletion schemes. Compared with the initialized saliency map shown in Fig. 5-3(c), the co-salient object appears to be more consistency and the backgrounds are effectively suppressed.

5.2.4 Iteration Scheme

In order to obtain more superior co-saliency map, an iterative scheme is designed in the proposed framework, as shown in Fig. 5-1. The iterative scheme works as a refinement model to combine the addition and deletion steps and refine the co-saliency map in loop. In the iteration scheme, a heuristic termination strategy is set by checking the maximum iteration number $I_{\max}$ and the difference between two iterations. Specifically, the second termination condition is introduced to check whether the saliency result becomes stable or not, which is formulated as the average difference between two iteration results.

$$D_{t}^{i}=\left(\frac{1}{\prod}\sum\left|S_{del}^{i}\left(t\right)-S_{del}^{i}\left(t-1\right)\right|\right)\le\varsigma \tag{5-18}$$

where $S_{del}^{i}(t)$ denotes the co-saliency map produced after the t^{th} iteration optimization, $\prod$ represents the number of pixels in the co-saliency map, and ς is a given threshold to determine whether the iteration should be terminated or not, which is set to 0.1 in all experiments. Until $D_{t}^{i}\le\varsigma$, the iteration will be terminated and output the final co-saliency map, otherwise, the iteration will continue. Some visual examples of the iteration scheme are shown in Fig. 5-4. The third column shows the original co-saliency result, and the first iteration and the final co-saliency maps are shown in the last two columns of Fig. 5-4. From the figure, the initial co-saliency map is improved obviously with the iteration processing. For example, the cartoon with blue hair (named Sulley) is suppressed effectively since it is not a common object in the image group. Similarly, the background regions around the red car are also suppressed through the iteration scheme.

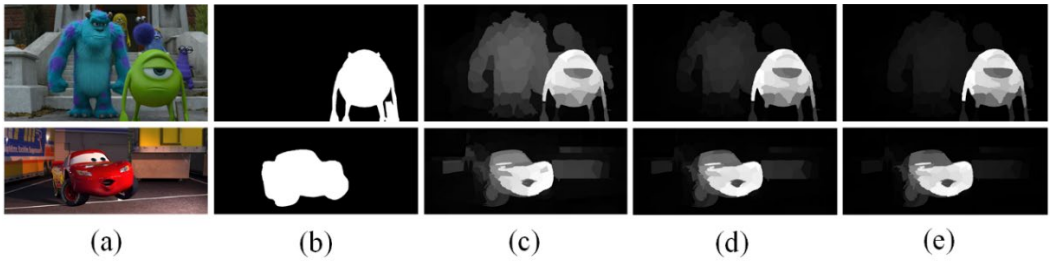

Fig. 5-4 Some examples of the iteration scheme. (a) RGB image. (b) Ground truth. (c) The initial saliency map through the addition and deletion scheme. (d) The saliency map after the first iteration. (e) The final saliency result.

5.3 Experimental Results and Discussion

In this section, the proposed framework is evaluated on two RGBD co-saliency datasets with the qualitative and quantitative comparisons. In addition, the ablation studies are conducted, which include the analysis of each module in the framework and the discussion of one-for-one option co-saliency framework.

5.3.1 Experimental Settings

The proposed co-saliency framework is evaluated on two RGBD benchmarks, *i.e.*, RGBD Coseg183 dataset [88] and RGBD Cosal150 dataset [114]. Three quantitative criteria are adopted to evaluate the co-saliency map, which include the Precision-Recall (PR) curve, F-measure, and AUC score. In the proposed method, the number of superpixels for each image is set to 200, the maximum iteration number is set to 5 for balancing the computational complexity and performance. the propsoed method is implemented in MATLAB 2014a on a Quad Core 3.5GHz workstation with 16GB RAM, which costs average 42.67 seconds to process one image.

5.3.2 Comparison with State-of-the-art Methods

The proposed method is compared with 8 state-of-the-art methods, which include RC [25], HS [26], BSCA [30], RRWR [31], DCLC [33], SCS [77], CCS [79], and LRMF [81]. The first five single image saliency methods are regarded as the input of the proposed framework, and the last three methods are the state-of-the-art co-saliency methods.

For subjective evaluation, some visual examples on two datasets are shown in Figs.

5-5 and 5-6, which consist of three image groups on RGBD Cosal150 dataset (*i.e.*, cartoon named Mike, red car, and statue), as well as three groups on RGBD Coseg183 dataset (*i.e.*, white cap, computer, and red flashlight). From Fig. 5-5, the single image saliency methods (*e.g.*, RC, HS, and RRWR) fail to discover the co-salient objects effectively and accurately. Taking the group Mike as an example, the common salient object is the green cartoon with big eye. However, many non-common objects, such as the cartoon with blue hair and the purple snake, are detected as the salient objects in the single saliency models. In addition, some background regions are not effectively suppressed, such as the trees in the statue group and non-salient cars in the red car group. In a word, the single saliency detection methods fail to detect the common salient objects in co-saliency scenarios. Therefore, it is essential that a co-saliency framework should be designed to convert the single saliency map into co-saliency result. The co-saliency map produced by our framework is shown in the last row of Fig. 5-5. Compared with the single saliency maps, the common salient regions are highlighted more consistent and accurate, and the backgrounds are suppressed effectively. To comprehensive evaluate the proposed method, three state-of-the-art co-saliency methods (*i.e.*, SCS [77], CCS [79], and LRMF [81]) are introduced for comparison. From the figures, the proposed approach can effectively highlight the common salient regions from the image group, and robustly suppress the background regions even when the salient regions exhibit large variations in shape and direction or the background is very complex and interferential.

Fig. 5-5 Visual comparison of different saliency and co-saliency detection methods on the RGBD Cosal150 dataset.

In contrast, the RGBD Coseg183 dataset is more difficult and challenging for co-saliency detection, and some visual examples are shown in Fig. 5-6. As can be seen, the proposed method achieves better performance compared with the other saliency and co-saliency detection methods. For example, the non-common objects (*e.g.* the white bowl and yellow cup) are effectively suppressed in the white cap group compared with other methods. Moreover, in the computer group, the consistency and homogeneity of the salient object is obviously improved compared with others. In the red flashlight group, the red flashlight is highlighted by the proposed method more effective than others. However, some backgrounds are still retained in the final result due to the small size and complex scene.

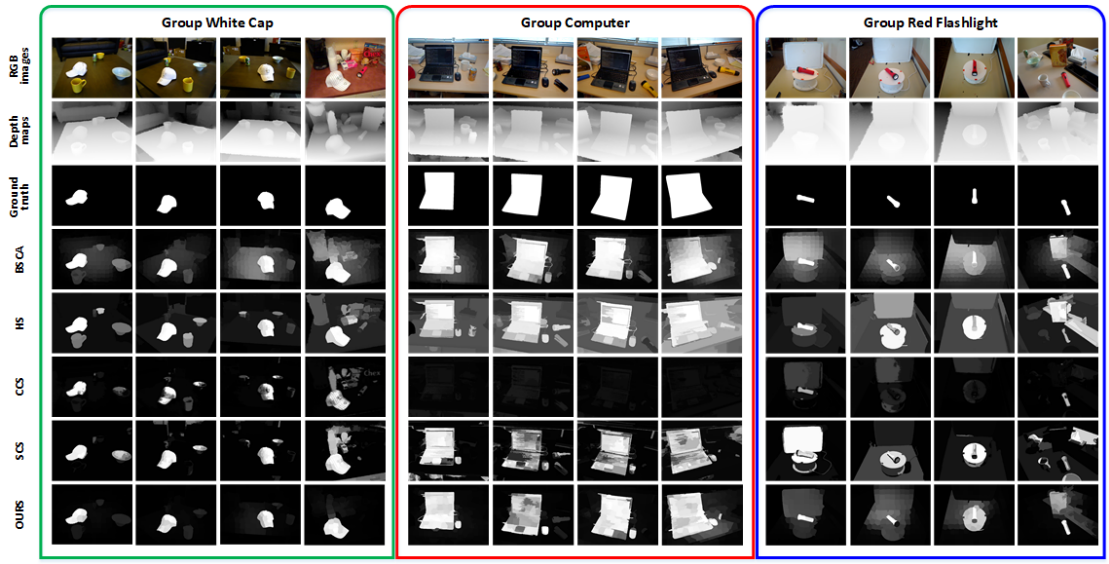

Fig. 5-6 Visual comparison of different saliency and co-saliency detection methods on the RGBD Coseg183 dataset.

The quantitative comparison results including the PR curves, F-measure, and AUC scores are reported in Fig. 5-7. As can be seen, on the RGBD Cosal150 dataset, the proposed method achieves the highest precisions of the whole PR curves, the largest F-measure and AUC score compared with other methods. The same conclusion can be drawn from the results on the RGBD Coseg183 dataset. From the PR curves on both two datasets, it can be seen that the final co-saliency result (*i.e.*, the red line) reaches the highest level in all curves, and the performance of co-saliency framework is obviously superior to the five original single image saliency models. It also demonstrates that the proposed co-saliency framework achieves the goal of converting the single saliency results into co-saliency scenarios. The F-measure and AUC scores

also support the conclusion. The proposed RGBD co-saliency detection framework aims to design a many-for-one structure, *i.e.*, multiple single saliency maps input and one co-saliency map output, to synthesize the superiority of different single saliency maps. In order to prove the effectiveness and versatility of the proposed algorithm, another one-for-one option is also implemented and evaluated, and the relevant results will be discussed in Section 5.3.5.

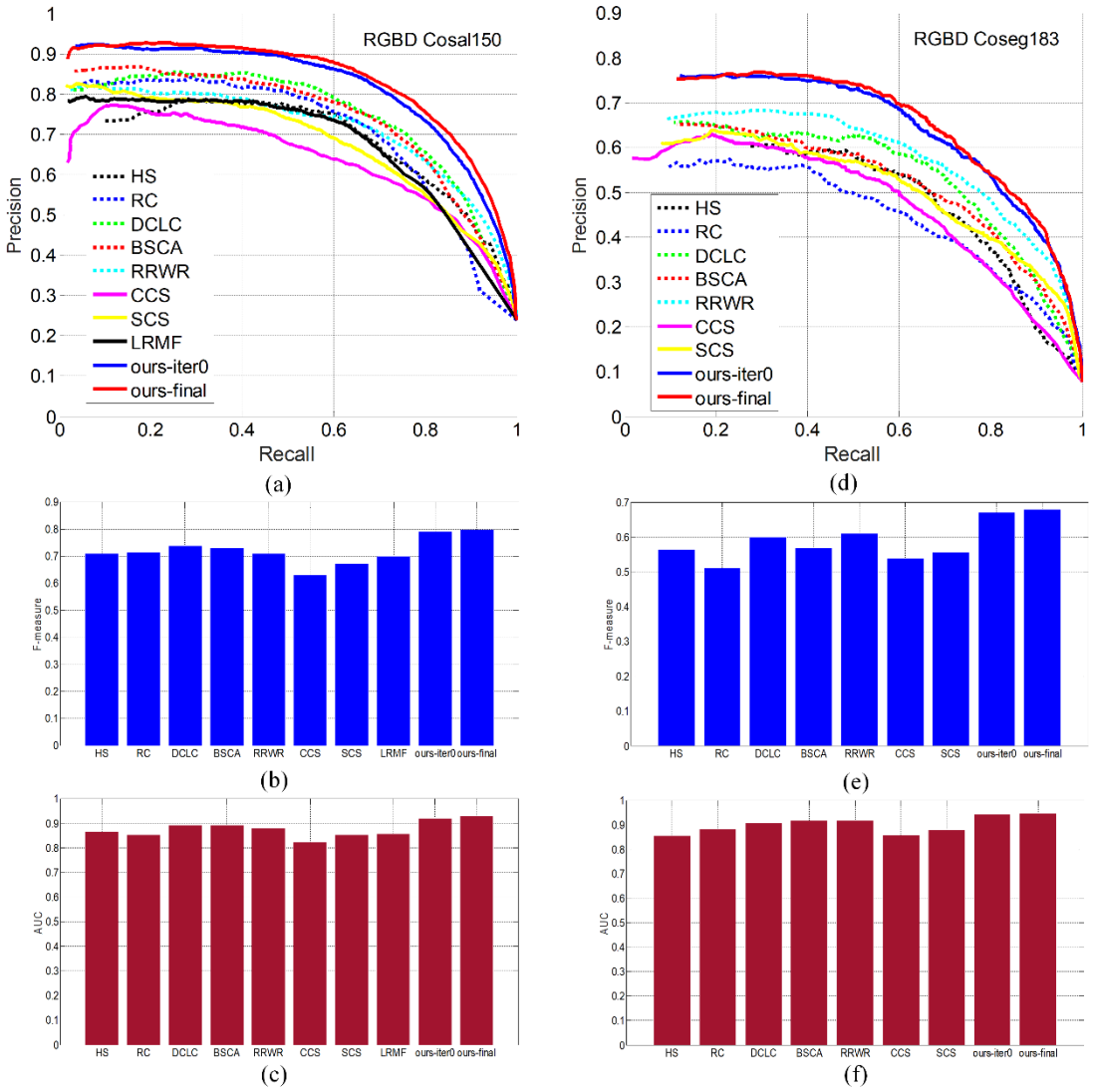

Fig. 5-7 Quantitative comparisons between the proposed method and the state-of-the-art methods on two datasets. Notice that "ours-iter0" means the co-saliency without iteration scheme, and "ours-final" denotes the co-saliency result with iteration scheme. (a)-(c) PR curves, F-measure, and AUC scores on the RGBD Cosal150 dataset. (d)-(f) PR curves, F-measure, and AUC scores on the RGBD Coseg183 dataset.

5.3.3 Module Analysis

In this subsection, each module of the proposed framework including the initialization, addition scheme, deletion scheme, and iteration scheme is comprehensively evaluated on the RGBD Cosal150 dataset. The quantitative comparisons are shown in Fig. 5-8, and the evaluation result of the iteration scheme is represented in Fig. 5-9.

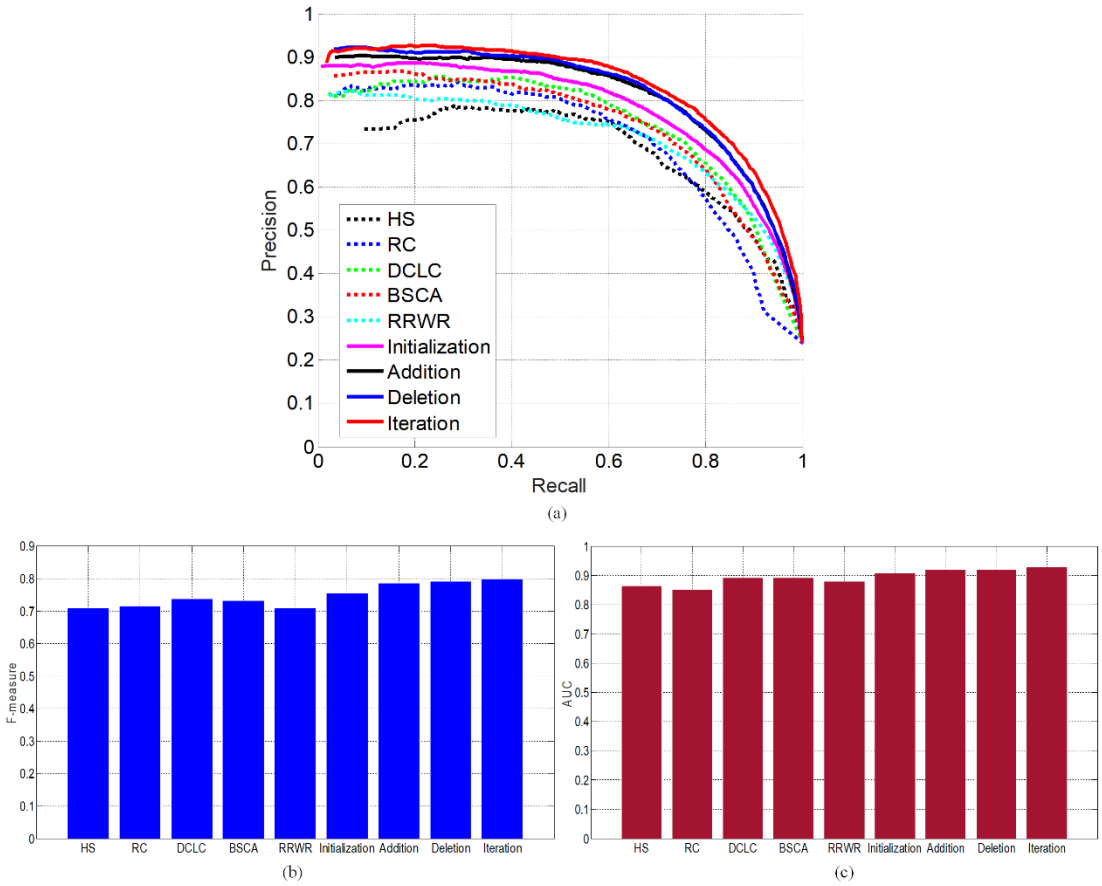

Fig. 5-8 Quantitative comparisons of the each part of the proposed framework on RGBD Cosal150 dataset. (a) PR curves. (b) F-measure. (c) AUC score.

In the initialization process, the original five saliency maps are integrated to produce a baseline for co-saliency detection, and its PR curve is marked as carmine in Fig. 5-8(a). Compared with the PR curves and F-measure of the original single saliency results, it indicates that the initialization result achieves better performance and produces a preferable baseline for later co-saliency detection. In the addition scheme, depth shape prior is proposed to introduce the depth information into the framework, and the label propagation is used to further optimize the saliency result. As shown in

Fig. 5-8, the saliency map through the addition scheme (the black line in PR curves) is improved significantly compared to the initialization result (the carmine line), and the F-measure and AUC score also achieve higher scores. Then, the deletion scheme is conducted to introduce the inter-image corresponding information into the framework and produce the initial co-saliency map. All the quantitative measurements in Fig. 5-8 show that the initial co-saliency result without iteration scheme obtains the best performance compared to other modules. In addition, an iteration scheme is designed to further update the co-saliency map and achieve more consistent result. The PR curves shown in Fig. 5-8(a) demonstrate that the performance is obviously improved using the iteration scheme, in which the blue line denotes the initial co-saliency result and the red line represents the final co-saliency result with iteration scheme. With the iteration scheme, the performance of co-saliency detection is continually optimized according to the F-measure and AUC score.

In order to verify the rationality of the iteration termination condition, an experiment of ten iterations without termination conditions are conducted on the RGBD Cosal150 dataset, and the detailed quantitative comparison results are shown in Fig. 5-9. From the average precision curve, the performance of the algorithm with the iterative progress tends to be stable gradually. In general, the termination conditions will not be satisfied after the first iteration, and its improved level is most noticeable. Moreover, most of the images will satisfy the termination condition after 3~4 iterations, that is, the co-saliency map no longer appears obvious changes. Thus, it also demonstrates that the maximum iteration number of 5 is reasonable in the experiments.

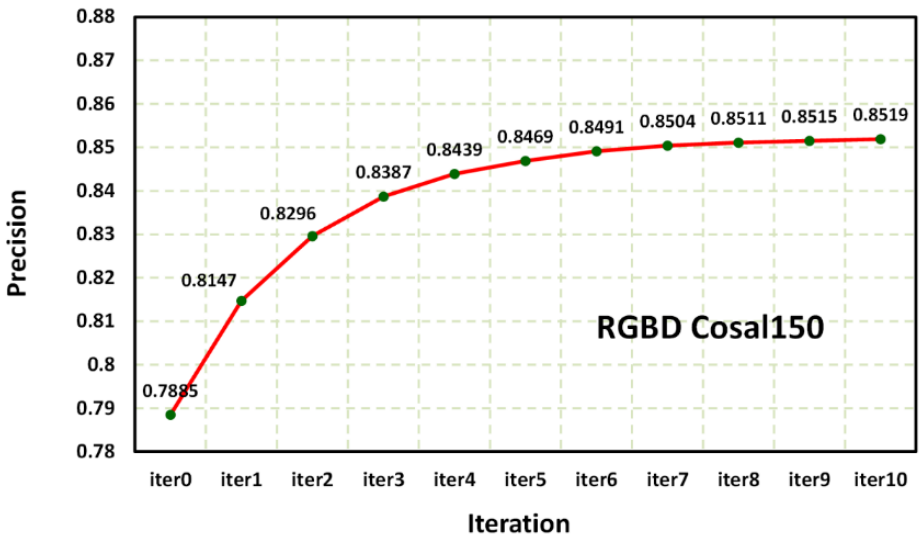

Fig. 5-9 Average precision of each iteration on the RGBD Cosal150 dataset.

5.3.4 Evaluation of Depth Shape Prior

In the framework, a novel depth descriptor, namely depth shape prior (DSP), is proposed to introduce the depth information to assist the identification of the co-salient objects. Introducing the depth shape prior into RGB saliency model, the 2D saliency model will turn into a RGBD saliency model and achieve a better performance. In this subsection, the performance of DSP is evaluated on NJU-400 dataset [123], and the relevant results are shown in Fig. 5-10 and Table 5-1. Five different 2D saliency maps produced by BSCA, RC, HS, RRWR, and DCLC methods are used as the original saliency maps.

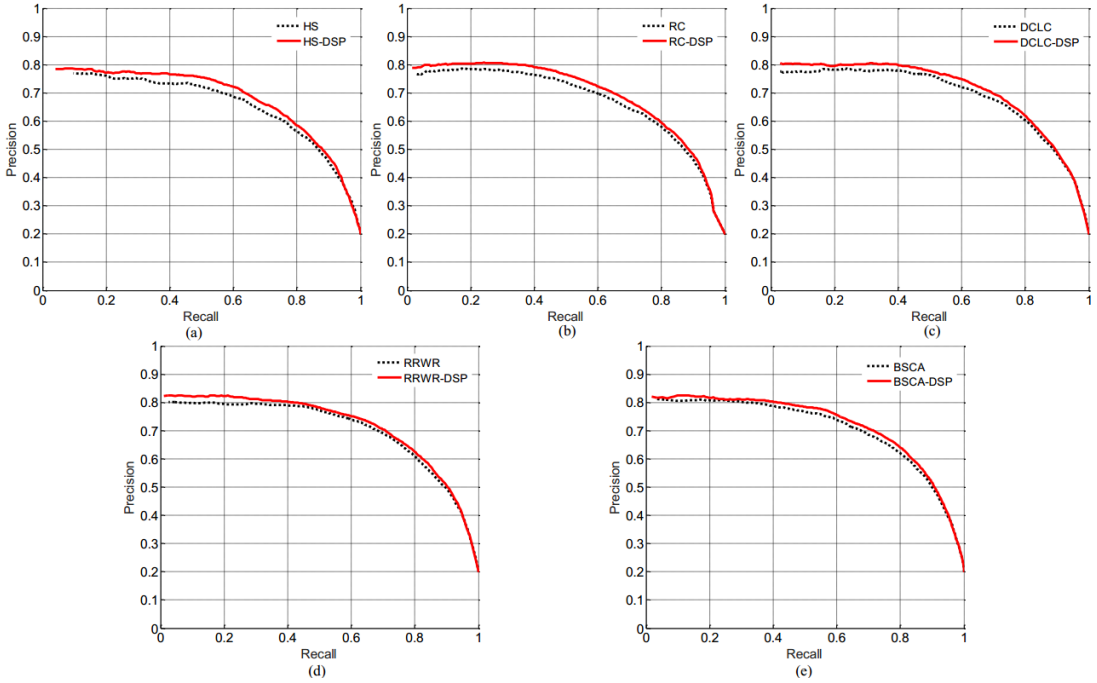

Fig. 5-10 Quantitative evaluation of the DSP on NJU-400 dataset. The black line in each PR curve denotes the original RGB saliency result, and the red line represents the saliency result with depth shape prior. The PR curves for (a) HS, (b) RC, (c) DCLC, (d) RRWR, and (e) BSCA methods.

In the PR curves shown in Fig. 5-10, the black line denotes the original RGB saliency result, and the red line represents the saliency result with depth shape prior. From the PR curves, it can be seen that the saliency result with DSP achieves the higher precisions of the whole PR curves compared with the 2D saliency results, and the F-measure also arrives at the consistent conclusion from Table 5-1. For the F-measure, the maximum percentage gain achieves 3.65% for HS method, and the average percentage gain achieves 2.38%.

Table 5-1 F-measure of the DSP on the NJU-400 Dataset.

	HS	RC	DCLC	RRWR	BSCA
without DSP	0.6661	0.6732	0.6914	0.7040	0.7022
with DSP	0.6904	0.6914	0.7094	0.7132	0.7136
percentage gain	3.65%	2.70%	2.60%	1.31%	1.62%

In order to further illustrate the effectiveness of DSP descriptor in the proposed model, an experiment is conducted on the RGBD Cosal150 dataset, and the results are reported in Table 5-2. From the table, the F-measure achieves the maximum percentage gain of 4.10% for RC method, and the average percentage gain reaches 3.09%. These experiments demonstrate that the depth information could improve the performances of the co-saliency. In other words, DSP can be used as an independent descriptor that converts the 2D saliency map into an RGBD saliency map.

Table 5-2 F-measure of the DSP on the RGBD Cosal150 Dataset.

	HS	RC	DCLC	RRWR	BSCA
without DSP	0.7101	0.7163	0.7385	0.7106	0.7318
with DSP	0.7294	0.7457	0.7642	0.7294	0.7502
percentage gain	2.72%	4.10%	3.48%	2.65%	2.51%

5.3.5 Discussion

The proposed RGBD co-saliency detection framework is designed as a many-for-one model, that is, multiple single 2D saliency maps input and one RGBD co-saliency map output. In fact, the proposed framework can achieve one-for-one model. In other words, if there is only one saliency map is embedded into the framework, it also can output one RGBD co-saliency map. The experimental comparison is reported in Fig. 5-11. The PR curves and F-measure demonstrate that the one-for-one option also achieves the transformation from single image saliency map to RGBD co-saliency map, and obtains better performance of co-saliency detection. In general, the better the saliency map is, the better the co-saliency map achieves. This is, of course, the reason why the multiple saliency maps are fused at first in the proposed framework. It can provide a better baseline for later detection in order to achieve more accurate and stable co-saliency result. However, as the results shown in Fig. 5-11, the proposed framework acquires satisfying result when only one saliency map is embedded.

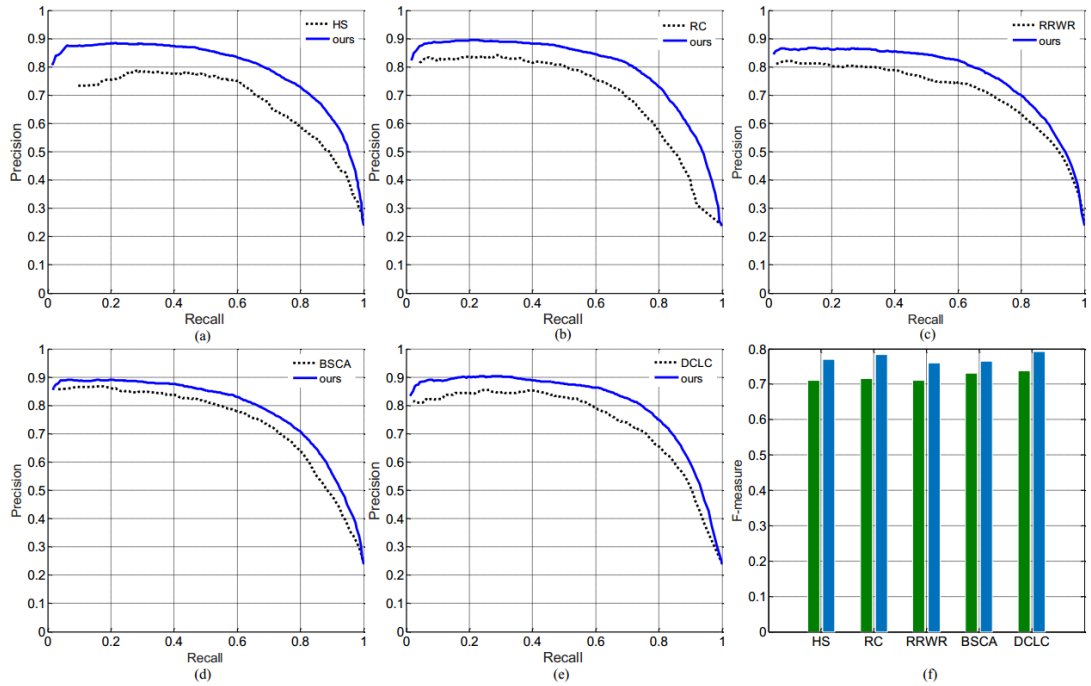

Fig. 5-11 Quantitative evaluation of one-for-one option for the proposed framework on the RGBD Cosal150 dataset. The black line in each PR curve denotes the original RGB saliency result, and the blue line represents the final co-saliency result using the proposed framework. The PR curves with different input saliency maps, *i.e.*, (a) HS, (b) RC, (c) RRWR, (d) BSCA, and (e) DCLC. (f) F-measure of the one-for-one framework.

5.4 Summary

In this chapter, an iterative RGBD co-saliency framework was proposed to convert the existing 2D saliency model into RGBD co-saliency scenario. Three schemes were integrated into the framework, including the addition scheme, deletion scheme, and iteration scheme. The addition scheme aimed to optimize the single saliency map and introduce the depth information into the framework using the depth shape prior descriptor. The deletion scheme focused on capturing the inter-image constraints and suppressing the non-common regions using a common probability function, which is formulated as the likelihood of each superpixel belonging to the common regions. The iterative scheme generated more homogeneous and consistent co-saliency map in a cycle way. The comprehensive comparisons and discussions on two RGBD co-saliency datasets have demonstrated that the proposed method outperforms other state-of-the-art saliency and co-saliency models.

Chapter 6 Hierarchical Sparsity Based Co-saliency Detection for RGBD Images

In this chapter, a novel co-saliency detection method for RGBD images is proposed based on hierarchical sparsity reconstruction and energy function refinement. With the assistance of the intra saliency map, the inter-image correspondence is formulated as a hierarchical sparsity reconstruction framework. The global sparsity reconstruction model with a ranking scheme focuses on capturing the global characteristics among the whole image group through a common foreground dictionary. The pairwise sparsity reconstruction model aims to explore the corresponding relationship between pairwise images through a set of pairwise dictionaries. In order to improve the intra-image smoothness and inter-image consistency, an energy function refinement model is proposed, which includes the unary data term, spatial smooth term, and holistic consistency term. Experiments on two RGBD co-saliency detection benchmarks demonstrate that the proposed method outperforms the state-of-the-art algorithms both qualitatively and quantitatively.

6.1 Introduction

With the recent explosive growth of data volume, people need to process multiple relevant images collaboratively. Co-saliency detection model needs to consider the common attributes of salient objects in an image group through the inter-image constraint. In other words, the co-salient objects should not only be prominent with respect to the backgrounds in each individual image, but also should recur throughout the whole image group [24]. In addition to the color appearance, human can perceive the distance mapping of a scene, which is known as depth information. Moreover, depth information has been proven to be useful for many computer vision tasks, such as segmentation [3], image retargeting [7], enhancement [10], and saliency detection [58,62,65]. However, most of the existing methods focus on handling the RGBD images rather than the RGBD image group. In this work, the depth feature is not only served

as a constraint in the inter-image correspondence modeling, but also used as a color information supplement in the refinement component.

Moreover, in addition to the saliency attribute in an individual image, the repetitiveness constraint across the whole image group is also crucial to suppress the background and non-common salient regions [24]. In the existing methods, the inter-image correspondence is simulated as a similarity matching [71-78], a cluster process [79], a rank constraint [80,81], a propagation process [82,83] or a learning process [84-87]. However, the matching- and propagation-based methods are often time consuming, while the clustering based methods are sensitive to the noise. To overcome these problems, the sparsity-based technique is a good choice and has demonstrated the potential to improve the performance of many tasks, including saliency detection. For the sparsity-based saliency detection methods, the background or foreground dictionary is used to reconstruct each processing unit, and the saliency is measured by the reconstruction error. In addition to describing the saliency of an individual image, sparsity representation can be used to constrain the inter-image correspondence capturing and to achieve inter saliency detection. In this work, a hierarchical sparsity reconstruction model is innovatively proposed to capture a more comprehensive inter-image relationship by considering the global and local inter-image information. The hierarchical sparsity property includes two complementary aspects:

(1) The co-salient objects in the whole image group should belong to the same category and have similar appearance. Therefore, a global foreground dictionary with a ranking scheme is built to reconstruct each image and to capture the global inter-image correspondence, which is called global sparsity reconstruction.

(2) The relationship among multiple images can be decomposed into a combination of multiple pairwise correspondences. Therefore, a set of foreground dictionaries constructed by other images are utilized to reconstruct the current image and obtain multiple pairwise inter saliency maps from the local perspective.

In addition, the co-salient objects in different images of the same group should be similar and consistent in appearance. Thus, a superior co-saliency detection model should guarantee the local smoothness in each individual image and global consistency in the whole image group. In this chapter, an energy function refinement model is proposed to attain a more consistent and accurate co-saliency result, which includes the unary data term, spatial smooth term, and holistic consistency term. The data term

constrains the updating degree of the refinement algorithm, and the smooth term favors that all the spatially adjacent regions with similar appearance should be assigned to consistent saliency scores. In addition to these two traditional terms, a holistic consistency term is specifically designed for the co-saliency detection task, which imposes the appearances of co-salient objects to be consistent in the whole image group.

In summary, an effective and efficient co-saliency detection method for RGBD images is provided based on hierarchical sparsity reconstruction and energy function refinement. The details will be introduced in the following sections.

6.2 Proposed Hierarchical Sparsity Model

The flowchart of the proposed hierarchical sparsity based co-saliency detection method for RGBD images is shown in Fig. 6-1, which includes intra saliency calculation, hierarchical inter saliency detection based on global and pairwise sparsity reconstructions, and energy function refinement.

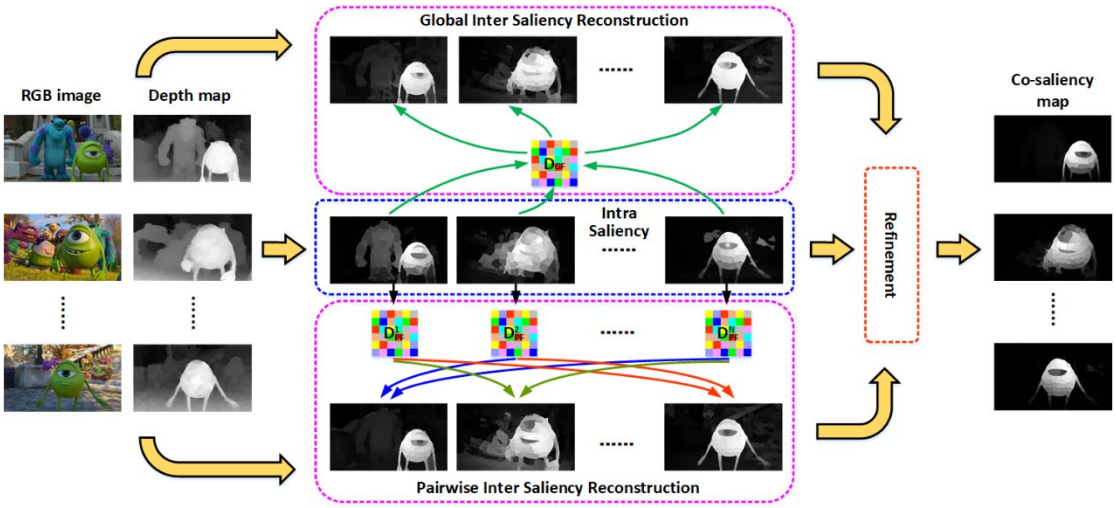

Fig. 6-1 Flowchart of the proposed RGBD co-saliency detection method.

According to the definition of co-saliency detection, the co-salient objects should be prominent in an individual image. Thus, the intra saliency map is firstly calculated for each individual image. The input RGB images in a group is denoted as $\{I^i\}_{i=1}^N$, and the corresponding depth maps as $\{D^i\}_{i=1}^N$, where N is the number of images in the group. For computational efficiency and structural representation, each RGB image I^i is abstracted into some superpixels $R^i = \{r_m^i\}_{m=1}^{N_i}$ through the SLIC algorithm [119], where N_i represents the number of superpixels in image I^i. In light of the

effectiveness and robustness, the DCMC method proposed in the Chapter 3 is used as the basic method for intra saliency detection, and the intra saliency value of superpixel r_m^i is denoted as $S_a\left(r_m^i\right)$.

The background of each image may be diverse within the same image group, while the co-salient objects tend to have a similar appearance in all images. Therefore, the co-salient regions can be reconstructed better than the background regions through a sparsity framework with the foreground dictionary. In this work, the corresponding relationship among multiple images is simulated as a hierarchical sparsity framework considering the global and pairwise sparsity reconstructions. The global inter saliency reconstruction model describes the inter-image correspondence from the perspective of the whole image group via a common reconstruction dictionary, while the pairwise inter saliency reconstruction model utilizes a set of foreground dictionaries produced by other images to capture local inter-image information.

Finally, an energy function refinement model, including the unary data term, spatial smooth term, and holistic consistency term, is proposed to improve the intra-image smoothness and inter-image consistency and to generate the final co-saliency map. The spatial smooth term is used to optimize the intra-image smoothness, and the holistic consistency term is specifically designed for co-saliency detection task to update the inter-image consistency. Hierarchical inter saliency detection based on global and pairwise sparsity reconstructions, as well as the energy function refinement, are detailed in the following sections.

6.2.1 Global Inter Saliency Reconstruction

The co-salient objects in a whole image group should belong to the same category and have a similar appearance. Therefore, a global foreground dictionary is built to reconstruct each image and capture the global inter-image correspondence. First, some initial foreground seeds are selected based on all the intra saliency maps in the image group. Then, a ranking filter is designed to eliminate the interference seeds and to determine the optimal foreground seeds. Next, the feature vectors of foreground seeds are extracted to construct the global foreground dictionary. Finally, the reconstruction error produced by the sparsity framework is utilized to measure the global inter saliency.

1) Initial foreground seeds selection

The intra saliency map provides effective single image saliency description. Assuming that most of the co-salient objects can be included in these saliency maps, the top K superpixels in image I^i with larger intra saliency values are selected as the foreground seeds. Then, all these seeds from different images are combined into an initial foreground seed set $\Phi_{init} = \Phi_{init}^1 \cup \Phi_{init}^2 \cup \ldots \cup \Phi_{init}^N$, where Φ_{init}^i denotes the foreground seed set of image I^i, and N is the number of images.

2) Ranking based seeds filtering

Since the intra saliency result is not completely accurate, some disturbances may be wrongly included in the initial foreground seed set, such as the backgrounds and non-common salient regions, which may degenerate the reconstruction accuracy. Therefore, a ranking scheme is designed to filter the interferences and refine the foreground seeds.

In general, the co-salient objects satisfy three constraints, *i.e.*, (a) the category should be same, (b) the color appearance should be similar, and (c) the depth distribution should be approximate. Combining these three constraints, a novel measure is designed to evaluate the local consistency of superpixels belonging to the initial foreground seed set. First, all the initial seed superpixels are grouped into five clusters by using the K-means++ clustering [127], and each superpixel is assigned to a corresponding cluster center $\{\boldsymbol{c}_i\}_{i=1}^{N\cdot K}$, respectively. Then, introducing the clustering, color, depth, and saliency constraints, the consistency measure is defined as follows:

$$mc(r_m) = \left[\sum_{n=1, n\neq m}^{N\cdot K} \left(1 - \left\|\boldsymbol{c}_m - \boldsymbol{c}_n\right\|_2\right) \cdot w_{mn} \right] \cdot S_a(r_m) \tag{6-1}$$

and

$$w_{mn} = \exp\left(-\frac{\chi^2(\boldsymbol{h}_m, \boldsymbol{h}_n) + \lambda_{\min} \cdot \left| d_m - d_n \right|}{\sigma^2} \right) \tag{6-2}$$

where $r_m, r_n \in \Phi_{init}$, r_m is the cluster center of superpixel r_m, w_{mn} represents the feature similarity between superpixel r_m and superpixel r_n, $S_a(r_m)$ is the intra saliency value of superpixel r_m, $N\cdot K$ is the total number of initial foreground seeds, $\|\cdot\|_2$ is the ℓ_2-norm function, and σ^2 is a parameter to control strength of the similarity, which is set to 0.1 in all experiments. $\boldsymbol{h}_m$ denotes the color histogram of superpixel r_m in the Lab color space, $\chi^2(\cdot)$ represents the Chi-square distance, and d_m is the mean depth value of superpixel r_m. $\lambda_{\min} = \min(\lambda_m, \lambda_n)$ denotes the minimum depth confidence measure of two input depth maps, where λ_m is the depth confidence measure of the

input depth map D^m. A larger mc corresponds to higher consistency with respect to other foreground seeds. In other words, the larger the consistency measure is, the higher the probability of the superpixel being the foreground seed. Finally, the top 80% of initial seeds with larger consistency measure values are reserved as the final foreground seeds, which is denoted as Φ_{fin}.

Some illustrations of the foreground seeds are shown in Fig. 6-2, where the third row presents the visualization of the initial foreground seeds marked in red, and the final foreground seeds marked in yellow after ranking scheme are shown in the last row. As can be seen, some backgrounds (*e.g.*, the lawns located by the blue arrows) in the second and third images are wrongly selected as the foregrounds in the initial seeds set. With the ranking scheme, the correct foreground seeds (*i.e.*, the dark bird) are successfully reserved, while the backgrounds are effectively eliminated.

Fig. 6-2 Some examples of ranking scheme for foreground seeds selection. The first second rows are the RGB images and depth maps, the third row shows the initial foreground seeds marked in red, and last row presents the final foreground seeds marked in yellow after ranking scheme.

3) Sparsity-based global reconstruction

Four types of low-level features, including color components, depth attribute, spatial location, and texture distribution, are utilized to describe each superpixel r_m^i as $\boldsymbol{f}_m^i = \left[\boldsymbol{l}_m^i \; d_m^i \; \boldsymbol{p}_m^i \; \boldsymbol{t}_m^i\right]^T$, where $\boldsymbol{l}$ is the 9-dimensional color components in the RGB, Lab, and HSV color spaces, d denotes the depth value, $\boldsymbol{p}$ corresponds to the 2-dimensional spatial coordinates, and $\boldsymbol{t}$ represents the 15-dimensional texton histogram [131]. The feature representations of the stacking superpixels in the final foreground seeds set Φ_{fin} are constructed as the global foreground dictionary, which is denoted as $\mathbf{D}_{GF}$.

Under the same reconstruction dictionary, the reconstruction error between foreground and background regions should be different. Thus, the image saliency can be measured by the reconstruction error [27]. The reconstruction error is computed by the sparsity representation with a global foreground dictionary, and each superpixel r_m^i is encoded by:

$$\alpha_m^{i\,*} = \arg\min_{\alpha_m^i} \left\| \boldsymbol{f}_m^i - \mathbf{D}_{GF} \cdot \alpha_m^i \right\|_2^2 + \xi \cdot \left\| \alpha_m^i \right\|_1 \tag{6-3}$$

where $\alpha_m^{i\,*}$ is the optimal sparse coefficient for superpixel r_m^i, $\mathbf{D}_{GF}$ denotes the global foreground dictionary, $\boldsymbol{f}_m^i$ is the feature representation of superpixel r_m^i, $\|\cdot\|_1$ is the ℓ_1 -norm function, and ξ is set to 0.01 as suggested in [27].

The foreground dictionary is used to achieve global reconstruction, thus, the superpixel with the smaller reconstruction error should be assigned to a greater saliency value and vice versa. The global inter saliency of superpixel r_m^i is defined as:

$$S_{gr}\left(r_m^i\right) = \exp\left(-\varepsilon_m^i / \sigma^2\right) = \exp\left(-\left\| \boldsymbol{f}_m^i - \mathbf{D}_{GF} \cdot \alpha_m^{i\,*} \right\|_2^2 / \sigma^2\right) \tag{6-4}$$

where $S_{gr}\left(r_m^i\right)$ is the inter saliency of superpixel r_m^i through the global reconstruction, ε_m^i denotes the reconstruction error of superpixel r_m^i, and σ^2 is a weighted constant.

6.2.2 Pairwise Inter Saliency Reconstruction

The global reconstruction aims to describe the inter-image correspondence from the perspective of the whole image group. In fact, the relationship among multiple images can be decomposed into a combination of multiple pairwise correspondences, which benefits capturing the local inter-image information. In order to deeply explore a more comprehensive inter-image corresponding relationship, a sparsity-based

pairwise reconstruction method is proposed to calculate the pairwise inter saliency. First, a foreground dictionary for each image is constructed based on the corresponding intra saliency map, respectively. In this way, the N foreground dictionaries in an image group are obtained, where N denotes the number of images in the group. Then, each image is reconstructed by the $N-1$ foreground dictionaries derived from other images in the group, respectively. Finally, these $N-1$ reconstructed results are fused to generate the pairwise inter saliency map.

For each image I^k, the top K superpixels with larger intra saliency values are selected as the foreground seeds. Similar to the sparsity-based global reconstruction, a 27-dimensional feature vector is used to represent each superpixel. Then, the feature representations of the stacking foreground superpixels in each image are constructed as the pairwise foreground dictionary, which is denoted as $\mathbf{D}_{PF}^k$. As mentioned earlier, the foreground pairwise dictionaries generated by other images can be utilized to reconstruct the current image and capture the local inter-image relationship. Using the pairwise foreground dictionary N produced by the image I^k, the image I^i can be constructed and the saliency is measured as:

$$S_{pr}^k\left(r_m^i\right)=\exp\left(-\varepsilon_m^{k,i}/\sigma^2\right)=\exp\left(-\left\|\boldsymbol{f}_m^i-\mathbf{D}_{PF}^k\cdot\alpha_m^{k,i*}\right\|_2^2/\sigma^2\right) \tag{6-5}$$

where $S_{pr}^k\left(r_m^i\right)$ is the inter saliency through the pairwise reconstruction using the dictionary $\mathbf{D}_{PF}^k$, $\varepsilon_m^{k,i}$ denotes the reconstruction error of superpixel r_m^i, $\alpha_m^{k,i*}$ is the optimal sparse coefficient of superpixel r_m^i, $k\in[1,2,\ldots,N]$, $k\neq i$ represents the index of pairwise foreground dictionary, and σ^2 is a weighted parameter. Therefore, $N-1$ saliency maps for each image can be obtained through different pairwise dictionaries. At last, all these maps are fused to generate the final pairwise inter saliency map by:

$$S_{pr}\left(r_m^i\right)=\frac{1}{N-1}\cdot\sum_{\substack{k=1\\k\neq i}}^{N}S_{pr}^k\left(r_m^i\right) \tag{6-6}$$

The global inter saliency map describes the global inter-image correspondence from the whole image group, while the pairwise inter saliency map captures the local relationship from the pairwise images. Finally, these two inter saliency maps are combined as the hierarchical sparsity based inter saliency:

$$S_r\left(r_m^i\right)=\frac{1}{2}\cdot\left(S_{gr}\left(r_m^i\right)+S_{pr}\left(r_m^i\right)\right) \tag{6-7}$$

where $S_r\left(r_m^i\right)$ denotes the hierarchical sparsity based inter saliency of superpixel r_m^i.

6.2.3 Energy Function Refinement

In order to achieve a superior and globally consistent saliency map, a refinement model with an energy function is designed in our work. Three terms are included in the energy function: the unary data term T_u constrains the similarity between the final saliency map and initial saliency map; the spatial smooth term T_s favors that all the similar and spatially adjacent superpixels in an individual image should be assigned to consistent saliency scores; and the holistic consistency term T_h enforces that the appearance of the salient objects should be consistent within the whole image group. Therefore, the energy function is defined as:

$$E = T_u + T_s + T_h = \sum_m \left(\bar{s}_m - s_m\right)^2 + \sum_{(m,n)\in\Omega} w_{mn} \cdot \left(\bar{s}_m - \bar{s}_n\right)^2 + \sum_m g_m \cdot \bar{s}_m^2 \tag{6-8}$$

where $\bar{s}_m$ denotes the refined saliency value of superpixel r_m, $s_m = S_a\left(r_m\right) \cdot S_r\left(r_m\right)$ is the initial saliency value of superpixel r_m by combining the intra and inter saliencies, Ω represents the spatially adjacent set in an individual image, w_{mn} denotes the similarity between two superpixels, which is defined in the same way as Eq. (6-2), and $g_m = \chi^2\left(\boldsymbol{h}_m, \boldsymbol{h}_g\right)$ is the color difference between the superpixel r_m and global foreground model via the chi-square distance of Lab color histograms. The top 20 superpixels with larger initial saliency value in each image are regarded as the foreground samples to represent the global foreground distribution.

Let $\boldsymbol{s} = [s_m]_{\aleph\times 1}$, and $\bar{\boldsymbol{s}} = [\bar{s}_m]_{NA\times 1}$, where $\aleph = \sum_{i=1}^{N} N_i$ is the total number of superpixels in the whole image group. Then, the energy function is rewritten in the matrix forms as:

$$\mathbf{E} = \left(\bar{\boldsymbol{s}} - \boldsymbol{s}\right)^T \cdot \left(\bar{\boldsymbol{s}} - \boldsymbol{s}\right) + \bar{\boldsymbol{s}}^T \cdot \left(\mathbf{D} - \mathbf{W}\right) \cdot \bar{\boldsymbol{s}} + \bar{\boldsymbol{s}}^T \cdot \mathbf{G} \cdot \bar{\boldsymbol{s}} \tag{6-9}$$

where $\mathbf{W} = \left[w_{mn}\right]_{\aleph\times\aleph}^{(m,n)\in\Omega}$ is the spatial color similarity matrix, $\mathbf{D} = diag\left(d_1, d_2, \cdots, d_{\aleph}\right)$ represents the degree matrix, $d_i = \sum_{j=1,(i,j)\in\Omega}^{\aleph} w_{ij}$, and $\mathbf{G} = diag\left(g_1, g_2, \ldots, g_{\aleph}\right)$ is the difference matrix between the superpixels and global foreground model.

The minimization of the above energy function can be solved by setting the derivative with respect to $\bar{\boldsymbol{s}}$ to be 0, which is represented as:

$$\frac{\partial \mathbf{E}}{\partial \bar{\boldsymbol{s}}} = 2\left(\bar{\boldsymbol{s}} - \boldsymbol{s}\right) + 2\left(\mathbf{D} - \mathbf{W}\right) \cdot \bar{\boldsymbol{s}} + 2\mathbf{G} \cdot \bar{\boldsymbol{s}} = 0 \tag{6-10}$$

Combining the like terms, the solution is given by:

$$\bar{\boldsymbol{s}} = \left[\mathbf{I} + (\mathbf{D} - \mathbf{W}) + \mathbf{G}\right]^{-1} \cdot \boldsymbol{s} \tag{6-11}$$

where $\mathbf{I}$ is an identity matrix with the size of $N \times N$.

6.3 Experimental Results

In this section, the proposed RGBD co-saliency detection method is evaluated on the RGBD Cosal150 dataset and RGBD Coseg183 dataset. The qualitative and quantitative comparisons with some state-of-the-art methods are presented, and some discussions and analyses are conducted.

6.3.1 Experimental Settings

In experiments, two public RGBD co-saliency detection datasets, *i.e.*, RGBD Cosal150 dataset [114] and RGBD Coseg183 dataset [88], are used to evaluate the effectiveness of the proposed method. For quantitative evaluation, three criteria including the PR curve, F-measure, and MAE score are introduced. In this work, the number of superpixels for each image is set to 400, and the number of initial foreground seeds is set to 40.

6.3.2 Comparison with State-of-the-art Methods

The proposed HSCS method is compared with 17 state-of-the-art methods, including DSR [27], BSCA [30], HDCT [32], DCLC [33], SMD [34], DCL [46], DSS [49], R3Net [133], ACSD [65], DF [61], CTMF [62], PCFN [63], SCS [77], CCS [79], LRMF [81], ICS [134], and MCLP [114], where DCL, DSS, R3Net, DF, CTMF, and PCFN are the deep learning based methods. The visual comparisons are shown in Fig. 6-3, and the quantitative evaluations are reported in Fig. 6-4 and Table 6-1.

In Fig. 6-3, four image groups, including the green cartoon in the virtual scene, sculpture in the outdoor scene, and the red and yellow flashlights in the indoor scene, are illustrated for visual comparison. Due to the lack of a high-level feature description and inter-image constraints, the unsupervised single image saliency detection methods (*e.g.*, DSR [27], HDCT [32]) only roughly highlight the salient regions , while the

background regions cannot be suppressed effectively (such as the street in the green cartoon group and the trees in the sculpture group). Benefitting from the strong learning ability of deep learning, the DCL [46] method achieves better performance with more consistent salient regions. However, there are still some wrongly detected backgrounds, such as the white object in the second image of the last group. Combining the depth cue and deep learning, the DF [61] method suppresses the background effectively, but it ignores the completeness of salient objects, such as the third image in the green cartoon group. For the RGB co-saliency detection methods (CCS [79] and SCS [77]), some foregrounds (such as the third image in the green cartoon group) are wrongly suppressed by the CCS method, and some backgrounds (such as the white board in the red flashlight group) are also inaccurately highlighted by the SCS method. Compared with the above methods, RGBD co-saliency detection methods (ICS [134] and MCLP [114]) achieve relatively superior performance with tangible salient objects. However, they still fail to suppress some common backgrounds, such as the ground in the sculpture group and the white board in the red flashlight group. By contrast, benefitting from the hierarchical reconstruction and global refinement, the proposed method can consistently highlight the salient objects and effectively suppress the backgrounds.

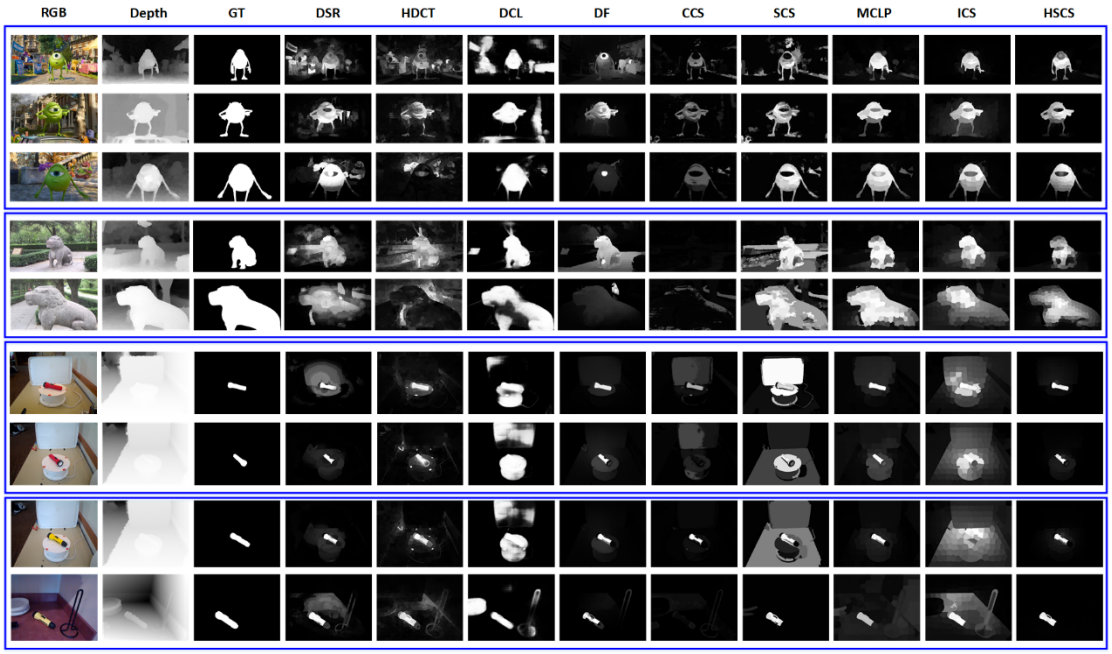

Fig. 6-3 Some visual examples of different methods.

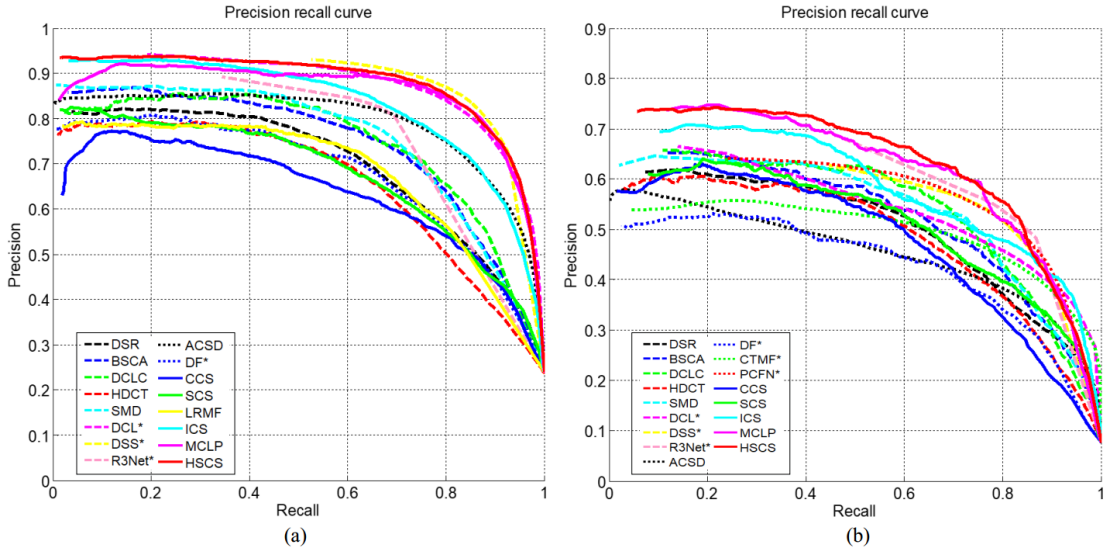

Fig. 6-4 PR curves of different methods on two RGBD co-saliency detection datasets, where "*" denotes the deep learning based methods. (a) PR curves on the RGBD Cosal150 dataset. (b) PR curves on the RGBD Coseg183 dataset.

Table 6-1 Quantitative comparisons with different methods on two datasets, where "*" denotes the deep learning based methods.

	RGBD Cosal150 Dataset		RGBD Coseg183 Dataset	
	F-measure	MAE	F-measure	MAE
DSR	0.6956	0.1867	0.5496	0.1092
BSCA	0.7318	0.1925	0.5678	0.1877
DCLC	0.7385	0.1728	0.5994	0.1097
HDCT	0.6753	0.2146	0.5447	0.1307
SMD	0.7494	0.1774	0.5760	0.1229
DCL*	0.8345	0.1056	0.5531	0.0967
DSS*	0.8540	0.0869	0.5972	0.0782
R3Net*	0.7812	0.1296	0.6190	0.0678
ACSD	0.7788	0.1806	0.4787	0.1940
DF*	0.6844	0.1945	0.4840	0.1077
CTMF*	-	-	0.5316	0.1259
PCFN*	-	-	0.6049	0.0782
CCS	0.6311	0.2138	0.5383	0.1210
SCS	0.6724	0.1966	0.5553	0.1616
LRMF	0.6995	0.1813	-	-
ICS	0.7915	0.1790	0.6011	0.1544
MCLP	0.8403	0.1370	0.6365	0.0979
ours	0.8500	0.1030	0.6466	0.0787

PR curves of different methods on two datasets are shown in Fig. 6-4. As can be seen, the proposed HSCS method reaches a higher precision on the whole PR curves. Moreover, the proposed method is even superior to some deep learning based methods (*e.g.*, DCL [46], R3Net [133], DF [61], CTMF [62], and PCFN [63]). The quantitative measurements, including F-measure and MAE score, are reported in Table 6-1. From the table, the proposed method achieves the competitive performance compared with 17 other state-of-the-art methods. On the RGBD Cosal150 dataset, the F-measure of the proposed method reaches 0.8500, and the maximum percentage gain reaches 34.7% compared with other methods. Especially, the proposed HSCS method also achieves the percentage gain of 8.8% compared with the deep learning based method (*e.g.*, R3Net [133]). On the RGBD Coseg183 dataset, the proposed method achieves the best performance in terms of F-measure, and the performance gains against others are more remarkable. The maximum percentage gain of the proposed method also reaches 35.1% in terms of F-measure. All these visual examples and quantitative measures demonstrate the effectiveness of the proposed method.

6.3.3 Module Analysis

The key points of the proposed hierarchical sparsity-based co-saliency detection method for RGBD images include a hierarchical sparsity based inter saliency model and an energy function refinement model. For hierarchical sparsity based inter saliency generation, the global and pairwise sparsity reconstructions are used to capture the inter-image constraints from two aspects. Each module is comprehensively evaluated on the RGBD Cosal150 dataset, and the F-measures are presented in Table 6-2.

Table 6-2 F-measure of the main modules on the RGBD Cosal150 dataset.

Modules	F-measure
global reconstruction	0.8145
pairwise reconstruction	0.7628
hierarchical inter saliency	0.8198
energy function refinement	0.8500

The global inter saliency reconstruction captures the global corresponding relationship throughout the whole image group and achieves the F-measure of 0.8145.

As a supplement, the multiple images relationship is formulated as pairwise correspondences by using the pairwise reconstruction model with a set of pairwise dictionaries, and the F-measure reaches 0.7628. Combining these two aspects, the hierarchical inter saliency structure can explore a more comprehensive inter-image relationship, and reaches 0.8198 in terms of F-measure, which is superior to most of the existing (co-)saliency detection methods (*e.g.*, DSR [27], SMD [34], DF [61], SCS [77], LRMF [81], and ICS [134]). Finally, the co-saliency detection with energy function refinement achieves the best performance, and the percentage gain reaches 3.7% compared with the inter saliency models.

6.3.4 Evaluation of Depth Cue and Ranking Scheme

In this work, the depth cue is not only served as a constraint in the inter-image correspondence modeling, but also used as a supplement of color information in the refinement component. In order to attain more robust and accurate foreground seeds for global dictionary construction, a ranking scheme is designed to filter the interferences and to obtain optimal foreground seeds. Some experiments are conducted on the RGBD Cosal150 dataset to evaluate the influence of these two constraints, and the F-measures are reported in Table 6-3.

Compared with the first and the third rows, introducing the depth cue into the model, the performance is obviously improved with a percentage gain of 8.4%. Shown in the second and third rows, the performance with the ranking scheme is better than the model without the ranking scheme. In addition, some illustrations are shown in Fig. 6-2. As can be seen, with the ranking scheme, the correct foreground seeds are successfully reserved, while the backgrounds (such as the lawn in the second and third images) are effectively eliminated. All these data demonstrate the effectiveness of the depth information and ranking scheme.

Table 6-3 Evaluation of depth and ranking scheme on the RGBD Cosal150 dataset, where "w/o" means "without", and "w/" corresponds to "with".

	F-measure
w/o depth and w/ ranking	0.7839
w/ depth and w/o ranking	0.8439
w/ depth and w/ ranking	0.8500

6.3.5 Parameter Discussion

In this section, the influence of different numbers of initial foreground seeds and superpixels is discussed, and the tendency chart of the F-measure is shown in Fig. 6-5.

From Fig. 6-5(a), selecting 20 initial foreground seeds for each image is not enough to represent the common saliency attributes completely and degenerates the inter reconstruction result. As the seed number increase, the performance improves and reaches the optimum when K is set to 40. When K reaches 50, the performance of the algorithm begins to drop. The main reason for the drop after 50 is that too many seeds contain background regions and decrease the reconstruction accuracy. As mentioned above, the performance is not highly sensitive to the parameter K, and it is set to 40 in all experiments. In addition to the number of initial foreground seeds, the influence of different numbers of superpixels is further discussed in the experiments. From the curve shown in Fig. 6-5(b), when the number of superpixels is set to 400, the result achieves the best performance. In fact, the performance in different numbers of superpixels are similar, indicating that the proposed algorithm is insensitive to the number of superpixels.

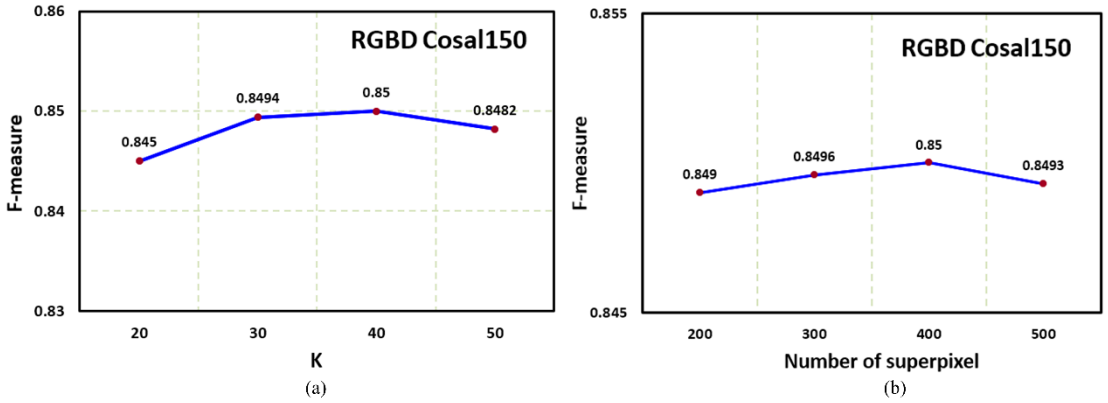

Fig. 6-5 F-measure of different parameters on the RGBD Cosal150 dataset. (a) The influence of different number of initial foreground seeds. (b) The influence of different number of superpixels.

6.3.6 Running Time

The running time of the proposed method is compared with others on a Quad Core 3.7GHz workstation with 16GB RAM and implemented using MATLAB 2014a. The average running time is listed in Table 6-4.

Table 6-4 Comparisons of the average running time (seconds per image) on the RGBD Cosal150 dataset.

Method	DCLC	SMD	DF	CCS	SCS	MCLP	ICS	HSCS
Time	1.96	7.49	12.95	2.65	2.94	41.03	42.67	8.29

In general, compared with the image saliency detection method, co-saliency detection algorithm often requires more computation time, especially for the matching based methods (such as MCLP [114], ICS [134]). For the three RGBD co-saliency detection methods, under the same conditions, the MCLP method takes 41.03 seconds for one image, the ICS method takes 42.67 seconds, and the proposed HSCS method takes an average of 8.29 seconds to process one image. Since the commonly used superpixel-level matching process is replaced by the hierarchical sparsity based reconstruction to capture the inter-image correspondence, the computational efficiency of the proposed algorithm is clearly improved.

6.4 Summary

In this chapter, a novel co-saliency detection method for RGBD images based on hierarchical sparsity reconstruction and energy function refinement was proposed. The major contribution lay in the hierarchical sparsity based inter saliency modeling, where the global inter-image model with a ranking scheme is used to capture the global characteristic among the whole image group through a common foreground dictionary, and the pairwise inter-image model is devoted to exploring the local corresponding relationship through a set of pairwise foreground dictionaries. In addition, an energy function refinement model was proposed to further improve the intra-image smoothness and inter-image consistency. The comprehensive comparisons and discussions on two RGBD co-saliency detection datasets have demonstrated that the proposed method outperforms other state-of-the-art methods both qualitatively and quantitatively.

Chapter 7 Video Saliency Detection via Sparsity-based Reconstruction and Propagation

Video saliency detection aims to continuously discover the motion-related salient objects from the video sequences. Since it needs to consider the spatial and temporal constraints jointly, video saliency detection is more challenging than image saliency detection. In this chapter, a novel method is proposed to detect the salient objects in video based on sparse reconstruction and propagation. With the assistance of novel static and motion priors, a single-frame saliency model is firstly designed to represent the spatial saliency in each individual frame via the sparsity-based reconstruction. Then, through a progressive sparsity-based propagation, the sequential correspondence in the temporal space is captured to produce the inter-frame saliency map. Finally, these two maps are incorporated into a global optimization model to improve spatiotemporal smoothness and global consistency of the salient object in the whole video. Experiments on three large-scale video saliency datasets demonstrate that the proposed method outperforms the state-of-the-art algorithms both qualitatively and quantitatively.

7.1 Introduction

In the past few decades, saliency detection for static image has gained much attention and achieved encouraging performances on the public benchmarks. By contrast, video saliency detection still remains as a relatively challenging and emerging issue. Different from the image saliency detection, video saliency detection aims to continuously locate the motion-related salient object from the video sequences by considering both the spatial and temporal information jointly, where the spatial information represents the intra-frame saliency in the individual frame, and the temporal information provides the inter-frame constraints and motion cues. Moreover, the salient objects in video are continuous in time axis and consistent among different frames, and the motion information is essential to distinguish the salient object from a complex scene.

In video data, the moving objects often attract more attention than the static ones. However, not all moving objects are salient targets and need to be further discriminated by the surrounding regions and adjacent frames. Therefore, how to make full use of the motion information to highlight the salient regions and suppress the backgrounds is essential to video saliency detection. Some motion-based features, such as optical flow contrast and optical flow gradient, have been utilized to separate the foreground regions from the video directly. Nevertheless, these methods are fragile due to the noises and moving backgrounds. In this work, the motion compactness and motion uniqueness are introduced as the motion cues to improve the motion saliency measurement, where the motion compactness describes the distribution of the optical flow, and the motion uniqueness represents the appearance characteristics of the motion amplitude information.

Exhibiting robustness to noise, sparsity-based techniques have been demonstrated to yield discriminative representations that have potential to improve the performances in a variety of inference tasks, such as object tracking, face recognition, and shape estimation. In addition, several saliency detection methods [27,135,136] construct the sparse models from the image and report satisfactory results against complex backgrounds. In [135], a weighted sparse coding framework on different data inputs was proposed to locate the salient objects. Recently, Yuan *et al*. [136] combined the deep neural network (DNN) and dense and sparse labeling (DSL) framework for saliency detection. By contrast, only a few studies [90, 91] employed the sparse representations to achieve video saliency detection. However, these methods only use sparse representations to capture the spatial information from individual frames, thus do not generalize well on the temporal space. To address this, a progressive sparse propagation framework with the forward-backward strategy is developed to model the inter-frame correspondence and generate the inter-frame saliency map. For the forward pass, the previous frame is utilized to build the forward dictionary and reconstruct the current frame. On the contrary, the backward pass processes the video from the last frame to the first frame, and the current frame is reconstructed by the backward dictionary constructed by the latter frame. Through the bidirectional propagation processes, the inter-frame relationship is exploited and the inter-frame saliency is achieved.

Generally, spatiotemporal consistency should be considered in video saliency

models to achieve more homogeneous result, *i.e.*, the saliency value of the salient region or background should not change drastically along the time axis. Moreover, in most of the existing methods, the input video is processed frame by frame without considering a global measure across the whole video sequence. In this way, the saliency result can only guarantee the local consistency rather than the global consistency. Therefore, a global optimization scheme based on energy function is proposed to obtain more homogeneous and consistent saliency result, which includes the unary data term, spatiotemporal smooth term, spatial incompatibility term, and global consistency term.

7.2 Proposed Sparsity Reconstruction and Propagation Model

Motivated by the inherent aspects of salient objects in video, three progressive steps are proposed to achieve video saliency detection, *i.e.*, single-frame saliency reconstruction, inter-frame saliency propagation, and global optimization. The flowchart is shown in Fig. 7-1. First, with the intuition that salient objects in the video should be salient in each individual frame, a single-frame saliency model is designed to capture the spatial saliency using the sparse reconstruction with the static and motion saliency priors. Then, the inter-frame saliency propagation with forward-backward strategy is utilized to model the sequential correspondence in the temporal space and generate the inter-frame saliency map. Finally, a global optimization model is designed to guarantee the global consistency of the salient object across the whole video and achieve more homogeneous saliency result. Each of these steps will be explained in the next subsections.

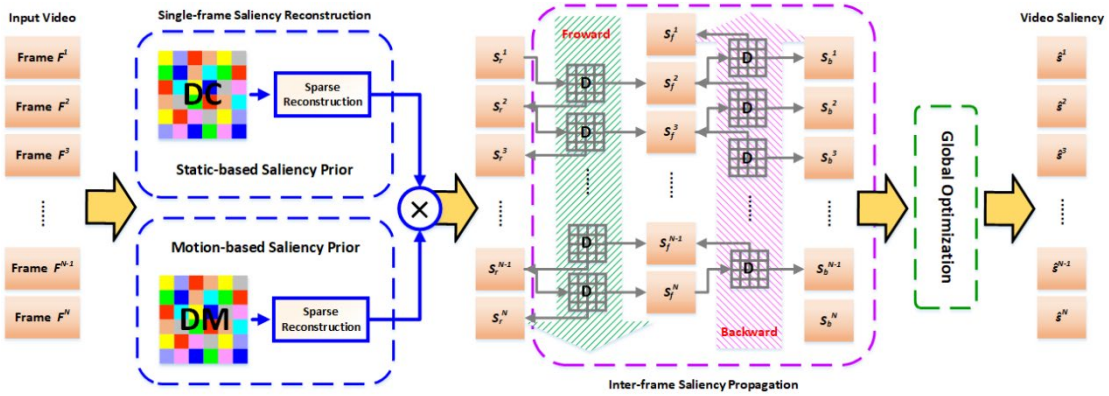

Fig. 7-1 Flowchart of the proposed video saliency detection framework.

7.2.1 Single-frame Saliency Reconstruction

For video saliency detection, the detected object should be salient with respect to the background and underlying motion in each frame. To this end, a sparse reconstruction model with two saliency priors is used to detect the salient object in each individual frame. The first one is the static saliency prior, which utilizes three color saliency cues to construct a color-based reconstruction dictionary (DC). The second one is the motion-based saliency prior, which integrates the motion uniqueness cue and motion compactness cue to build a motion-based dictionary (DM).

Given a video sequence $\mathbf{F}=\left\{F^t\right\}_{t=1}^{N}$ including N frames, some homogeneous superpixels $\mathbf{R}^t=\left\{r_k^t\right\}_{k=1}^{N^t}$ are firstly derived by using the SLIC algorithm [119] for each frame F^t, where N^t is the number of superpixels. In addition, the large displacement optical flow [137] is calculated to represent the pixel-level motion vector. The motion vector $\boldsymbol{v}_k^t$ of superpixel r_k^t is defined as the mean value of pixel-level motion vector in the superpixel.

1) Static-based saliency prior

The static-based saliency prior measures the static saliency in each frame by incorporating the background dictionary into a sparse representation framework. Three color-based cues, including background cue, compactness cue, and uniqueness cue, are integrated to select the background seeds and build the dictionary for reconstruction.

Background Cue. It is generally accepted that in video production, the important objects are close to the image center rather than the boundaries, which is a natural response of the cameraman operating the imaging system. Thus, the superpixels located at the image boundaries are more likely to be the background seeds, and this observation has been applied to many saliency detection models [27,29]. In this work, the superpixels along the image boundaries are selected as the background candidate set Φ_{SB}^t that represents the spatial location attribute of the background regions.

Static Compactness Cue. The salient regions incline to have a small spatial variance, whereas the backgrounds usually have a high spatial variance since their superpixels are often distributed over the entire image. Therefore, the compactness cue is introduced to describe the spatial distribution of the background regions. Following the DCLC method [33], the spatial variance of superpixel r_k^t is calculated by:

$$v_s\left(r_k^t\right)=\frac{\sum_{l=1}^{N^t} a_{kl}^t \cdot n_l^t \cdot \left\|\boldsymbol{p}_l^t-\boldsymbol{\mu}_k^t\right\|_2}{\sum_{l=1}^{N^t} a_{kl}^t \cdot n_l^t} \tag{7-1}$$

where $a_{kl}^t=\exp\left(-\left\|\boldsymbol{lc}_k^t-\boldsymbol{lc}_l^t\right\|_2/\sigma^2\right)$ denotes the color similarity between two superpixels, $\boldsymbol{lc}_k^t$ is the mean Lab color value of superpixel r_k^t, n_l^t represents the number of pixels that belong to the superpixel r_l^t, $\boldsymbol{p}_l^t$ denotes the spatial coordinates of superpixel r_l^t, $\boldsymbol{\mu}_k^t$ is the spatial mean, $\|\cdot\|_2$ is the ℓ_2-norm function, and σ^2 is a parameter to control degree of the similarity, which is set to 0.1. Then, the top Q_1 superpixels with larger spatial variances are selected as the compactness-based background candidate set Φ_{SC}^t.

Static Uniqueness Cue. The third cue represents the global appearance of the background regions in which the salient object shows different properties in appearance compared with the background. In this work, a cluster-based method is proposed to define the uniqueness cue. First, K-means++ clustering [127] is used to group the superpixels into K clusters $\left\{C_i^t\right\}_{i=1}^K$ with the cluster centers $\left\{c_i^t\right\}_{i=1}^K$, where the cluster number is set to 20 in the experiments. Then, two clusters with the largest Euclidean distance are selected by:

$$\left\{C_p^t, C_q^t\right\}=\underset{m,n\in\{1,2,\cdots,K\}}{\arg\max}\ E_d\left(c_m^t, c_n^t\right)\cdot e^{-\left|v_s\left(C_m^t\right)-v_s\left(C_n^t\right)\right|} \tag{7-2}$$

where $E_d\left(c_m^t, c_n^t\right)$ is the Euclidean distance between the two cluster centers, and $v_s\left(C_m^t\right)$ denotes the mean spatial variance of the cluster C_m^t. The selected two clusters correspond to one foreground cluster and one background cluster. Finally, a decision scheme considering the spatial variance and background probability is designed to determine the uniqueness-based background candidate set Φ_{SU}^t as:

$$\Phi_{SU}^t=\begin{cases}\left\{C_p^t\right\}, \textit{if}\left[v_s\left(C_p^t\right)>v_s\left(C_q^t\right)\right]\cap\left[P_b\left(C_p^t\right)>P_b\left(C_q^t\right)\right]\\ \left\{C_q^t\right\}, \textit{if}\left[v_s\left(C_p^t\right)\le v_s\left(C_q^t\right)\right]\cap\left[P_b\left(C_p^t\right)\le P_b\left(C_q^t\right)\right]\\ \varnothing\ , \textit{ortherwise}\end{cases} \tag{7-3}$$

where $P_b\left(C_p^t\right)$ is the mean background probability of the cluster C_p^t by using the method in [29].

Static-based Saliency Reconstruction. The final background set is obtained by combining all background candidates as $\Phi_{CB}^t=\Phi_{SB}^t\cup\Phi_{SC}^t\cup\Phi_{SU}^t$. Then, three types of features considering the color components, spatial location, and texture distribution are used to describe each superpixel. The color features in different color spaces are the intuitive representation of the superpixel, which is denoted as $\boldsymbol{c}=[R, G, B, L, a, b, H, S, V]$. The position coordinates benefit for depicting the spatial

relationship of the superpixel, which is represented as $\boldsymbol{p}=[x,y]$. The texton histogram $\boldsymbol{t}$ describes the local texture information of the superpixel [131]. All these hand-crafted features are firstly normalized to [0,1], and then concatenated into a feature vector to represent the superpixel r_k^t, which is denoted as $\boldsymbol{x}_k^t=\left[\boldsymbol{c}_k^t,\boldsymbol{p}_k^t,\boldsymbol{t}_k^t\right]^T$. The background dictionary $\mathbf{D}_B^t$ is constructed by the feature representations of the stacking background seeds in Φ_{CB}^t.

Based on the assumption that reconstruction error should be different for foreground and background through a sparse reconstruction model, the image saliency can be measured by the reconstruction error [27]. Each superpixel r_k^t is encoded by:

$$\alpha_k^{t^*}=\arg\min_{\alpha_k^t}\left\|\boldsymbol{x}_k^t-\mathbf{D}_B^t\cdot\alpha_k^t\right\|_2^2+\lambda\cdot\left\|\alpha_k^t\right\|_1 \tag{7-4}$$

where $\alpha_k^{t^*}$ is the optimal sparse coefficient for superpixel r_k^t, $\mathbf{D}_B^t$ denotes the background dictionary for frame F^t, $\boldsymbol{x}_k^t$ is the feature representation of superpixel r_k^t, λ is set to 0.01 as suggested in [27], $\|\cdot\|_1$ and $\|\cdot\|_2$ indicate the ℓ_1-norm and ℓ_2-norm functions, respectively.

For the sparse reconstruction with a background dictionary, the salient region will have a large reconstruction error, while the reconstruction error of the background region should be small. Thus, the saliency of superpixel r_k^t can be measured by the reconstruction error ε_k^t:

$$S_s\left(r_k^t\right)=\varepsilon_k^t=\left\|\boldsymbol{x}_k^t-\mathbf{D}_B^t\cdot\alpha_k^{t^*}\right\|_2^2 \tag{7-5}$$

where $S_s\left(r_k^t\right)$ denotes the static saliency value of superpixel r_k^t via the reconstruction error ε_k^t.

2) Motion-based saliency prior

Moving target attracts more attention in visual perception, thus, a motion-based saliency prior is introduced to represent the salient object from the perspective of motion space. An example of the optical flow data is shown in Fig. 7-2, where the spatial distribution of moving object is more concentrated than the background regions in the optical flow data. In addition, the moving object is often different from the background regions in terms of the magnitude of optical flow (MOF), which is consistent with the uniqueness cue in the color space. Based on these observations, the color-related cues is extended to the motion field and determine the background seeds for dictionary construction.

Motion Compactness Cue. The intuition is that, in the whole video sequences,

the spatial location distribution of moving object is more concentrated and compact in the optical flow field, whereas the background is distributed over the entire image. Therefore, a "motion compactness" cue is introduced to describe the distribution of the optical flow data and determine the background candidates. Similar to the color-based spatial variance, the motion-based spatial variance is defined as:

$$v_m\left(r_k^t\right)=\frac{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t \cdot \left\| \boldsymbol{p}_l^t - \tilde{\boldsymbol{\mu}}_k^t \right\|_2}{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t} \tag{7-6}$$

where $m_{kl}^t=\exp\left(-\left\|\boldsymbol{v}_k^t-\boldsymbol{v}_l^t\right\|_2/\sigma^2\right)$ is the motion similarity between two superpixels, $\boldsymbol{v}_k^t$ denotes the optical flow vector of superpixel r_k^t, n_l^t represents the number of pixels that belong to superpixel r_l^t, $\boldsymbol{p}_l^t=\left[x_l^t,y_l^t\right]$ is the centroid coordinates of superpixel r_l^t, and σ^2 is a constant parameter. $\tilde{\boldsymbol{\mu}}_k^t=[\mu x_k^t,\mu y_k^t]$ represents the spatial mean in the optical flow field, which is defined as:

$$\begin{cases} \mu x_k^t=\dfrac{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t \cdot x_l^t}{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t} \\ \mu y_k^t=\dfrac{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t \cdot y_l^t}{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t} \end{cases} \tag{7-7}$$

where a larger $v_m\left(r_k^t\right)$ indicates that the distribution of superpixel r_k^t in optical flow field is more dispersed, the background probability of the superpixel is greater. Then, the top Q_1 superpixels with larger motion-based spatial variance are composed to the background candidate set Φ_{MC}^t.

(a) (b) (c)

Fig. 7-2 Optical flow data of different video frames. (a) RGB image. (b) Optical flow map. (c) The MOF data.

Motion Uniqueness Cue. In general, the moving target exhibits different motion appearance compared with the background regions in the MOF data. Therefore, a "motion uniqueness" cue in the MOF field is defined by calculating the global contrast of each superpixel:

$$u_m\left(r_k^t\right)=\sum_{k=1,k\neq l}^{N^t}\left|M_f\left(r_k^t\right)-M_f\left(r_l^t\right)\right|\cdot e^{-\frac{E_d\left(\boldsymbol{p}_k^t,\boldsymbol{p}_l^t\right)}{\sigma^2}} \tag{7-8}$$

where $u_m\left(r_k^t\right)$ is the motion-based uniqueness measure of superpixel r_k^t, $M_f\left(r_k^t\right)$ denotes the MOF value of the superpixel r_k^t, and $E_d(\cdot)$ is the Euclidean distance function between two superpixels, which emphasizes the effect of closer superpixels. The smaller the uniqueness value is, the greater background probability of the superpixel achieves. Thus, top Q_1 superpixels with smaller motion-based uniqueness value are selected to build the background candidate set Φ_{MU}^t.

Motion-based Saliency Reconstruction. The final motion-related background set is determined by combining two background candidate sets, as $\Phi_{MB}^t=\Phi_{MC}^t\cup\Phi_{MU}^t$. For the motion-based sparse reconstruction, the motion feature is necessarily introduced to represent the motion cue. Furthermore, in order to guarantee the robustness of the feature representation, the basic color components are also embedded into the feature pool. Each superpixel is represented as a 12-dimensional feature vector $\boldsymbol{x}_k^t=\left[\boldsymbol{c}_k^t,\boldsymbol{m}_k^t\right]^T$, where $\boldsymbol{c}$ is the 9-dimensional color feature, and $\boldsymbol{m}$ denotes the 3-dimensional motion feature involving the components and magnitude of optical flow data. Then, the feature representation of each motion-related background seed is used to construct the background dictionary for frame F^t as $\mathbf{D}_B^t$. At last, as same as the static-based saliency reconstruction in Eqs. (7-4)~(7-5), the motion saliency of each superpixel is represented by the reconstruction error, which is denoted as $S_m\left(r_k^t\right)$.

3) Single-frame saliency map

The static saliency and motion saliency aim to discover the salient object from different feature domains. These two saliency maps are integrated to produce the single-frame saliency map as:

$$S_r\left(r_k^t\right)=S_s\left(r_k^t\right)\cdot S_m\left(r_k^t\right). \tag{7-9}$$

7.2.2 Inter-frame Saliency Propagation

The sequential relationship across the time axis is crucial to video saliency

detection. The salient object in an individual frame should be further discriminated by using the inter-frame information. Considering the high consistency and smoothness of the salient object in appearances and views between two adjacent frames, the previous frame can be employed to build a foreground dictionary and reconstruct the current frame in a forward way. Likewise, the current frame can be reconstructed by the next frame in a backward propagation manner. Therefore, a spatiotemporal saliency model is established via sparse propagation with a forward-backward strategy to smooth the salient object and suppress the background.

For the inter-frame sparsity-based saliency propagation, the directly adjacent frames are most relevant to the current frame, which benefits for capturing the common attributes of the salient objects. The proposed forward-backward propagation strategy is a heuristic method for inter-frame relationship abstraction in a progressive manner. The forward saliency and backward saliency are progressively correlated, where the forward saliency result is embedded into the feature pool to construct the dictionary and conduct the backward propagation. Through the bidirectional propagation processes, the exploitation of inter-frame relationship becomes more comprehensive and accurate.

1) Forward propagation

In the forward propagation, the current frame is reconstructed by a foreground dictionary derived from the previous frame, and the video is sequentially processed from the first frame to the last frame.

First, top Q_2 superpixels with larger single-frame saliency values in frame F^{t-1} are selected as the foreground seeds in the forward pass. Then, using the spatiotemporal features, each superpixel is represented as $\boldsymbol{x}_k^t = \left[\boldsymbol{c}_k^t, \boldsymbol{p}_k^t, \boldsymbol{t}_k^t, \boldsymbol{m}_k^t, S_r\left(r_k^t\right) \right]^T$, where $\boldsymbol{c}$ represents the 9-dimensional color feature, $\boldsymbol{p}$ is the 2-dimensional spatial coordinates, $\boldsymbol{t}$ is the 15-dimensional texton histogram, $\boldsymbol{m}$ denotes the 3-dimensional motion feature vector, and S_r is the single-frame saliency value. The feature representations of all foreground seeds from frame F^{t-1} are stacked to construct the forward foreground dictionary for frame F^t, which is denoted as $\mathbf{D}_F^{t-1}$.

Each superpixel in the current frame F^t is reconstructed by the forward foreground dictionary $\mathbf{D}_F^{t-1}$ through the sparse framework, and the reconstruction error $\overline{\varepsilon_k^t}$ is calculated to measure the forward saliency of superpixel r_k^t. Since the foreground dictionary is used for sparse reconstruction, the reconstruction error of the foreground

regions should be small, and the background regions have a large reconstruction error. In other words, the superpixel with smaller reconstruction error should be assigned to a greater saliency value, and vice versa. Therefore, following [135], the forward saliency of superpixel r_k^t is measured by an exponential function of reconstruction error:

$$S_f\left(r_k^t\right)=\exp\left(-\bar{\varepsilon}_k^t/\sigma^2\right)=\exp\left(-\left\|\boldsymbol{x}_k^t-\mathbf{D}_F^{t-1}\cdot\bar{\alpha}_k^{t*}\right\|_2^2/\sigma^2\right) \tag{7-10}$$

where $S_f\left(r_k^t\right)$ is the saliency value in the forward pass, $\bar{\alpha}_k^{t*}$ denotes the optimal sparse coefficient obtained by solving Eq. (7-4) with the forward foreground dictionary $\mathbf{D}_F^{t-1}$, and $\sigma^2=0.1$ represents a weighted parameter.

2) Backward propagation

The forward propagation captures the pre-order inter-frame relationship. Similarly, a backward pass is further carried out, which processes the video from the last frame to the first frame in a post-order way. The backward pass is the same as the forward pass, except for the foreground dictionary construction.

In the backward propagation, the single-frame saliency and forward saliency are combined to determine the foreground seeds. First, top $Q_2/2$ superpixels with larger saliency values in the single-frame and forward saliency models are selected, respectively. Then, the union of these superpixels are determined as the final foreground seeds in the backward pass. Different from the forward pass, the forward saliency S_f is added into the feature pool, which is denoted as $\boldsymbol{x}_k^t=\left[\boldsymbol{c}_k^t,\boldsymbol{p}_k^t,\boldsymbol{t}_k^t,\boldsymbol{m}_k^t,S_r\left(r_k^t\right),S_f\left(r_k^t\right)\right]^T$. Finally, the backward reconstruction error $\tilde{\varepsilon}_k^t$ is used to define the backward saliency:

$$S_b\left(r_k^t\right)=\exp\left(-\tilde{\varepsilon}_k^t/\sigma^2\right)=\exp\left(-\left\|\boldsymbol{x}_k^t-\mathbf{D}_F^{t+1}\cdot\tilde{\alpha}_k^{t*}\right\|_2^2/\sigma^2\right) \tag{7-11}$$

where $\tilde{\alpha}_k^{t*}$ denotes the optimal sparse coefficient obtained by solving Eq. (7-4) with the backward foreground dictionary $\mathbf{D}_F^{t+1}$.

7.2.3 Global Optimization

In order to achieve superior and globally consistent saliency map, an efficient optimization model with an energy function that consists of four complementary terms is proposed.

Unary Data Term. This term encourages the similarity between the final saliency map and initial saliency map, which is defined as:

$$E_u = \sum_k \left(\widehat{s_k^t} - s_k^t \right)^2 \tag{7-12}$$

where $\widehat{s_k^t}$ represents the final optimized saliency value of superpixel r_k^t, and $s_k^t = S_r\left(r_k^t\right) + S_f\left(r_k^t\right) + S_b\left(r_k^t\right)$ is the initial saliency value by combining three obtained saliency maps.

Spatiotemporal Smooth Term. This term favors that all the similar and spatiotemporally adjacent superpixels across the whole video should be assigned to consistent saliency scores, which is calculated by:

$$E_s = \sum_{(k,l)\in\Omega_{st}} \omega_{kl} \cdot \left(\widehat{s_k^t} - \widehat{s_l^t} \right)^2 \tag{7-13}$$

where ω_{kl} is the Lab color similarity between superpixels r_k and r_l, and $\Omega_{st} = \Omega_s \cup \Omega_t$ is the spatiotemporal adjacent set. Ω_s is the spatially adjacent set in one frame:

$$\Omega_s = \left\{ \left(r_m^t, r_n^t \right) \mid r_m^t \text{ and } r_n^t \text{ are spatially adjacent in } F^t \right\}. \tag{7-14}$$

Following the settings in [95], the temporally adjacent set Ω_t is represented as:

$$\Omega_t = \left\{ \left(r_m^t, r_n^{t'} \right) \mid \left\| \boldsymbol{p}_m^t - \boldsymbol{p}_n^{t'} \right\| \le 800 \;\&\; \left| t - t' \right| = 1 \right\} \tag{7-15}$$

where $\boldsymbol{p}_m^t$ is the spatial coordinates of superpixel r_m^t, and t' denotes the frame index.

Spatial Incompatibility Term. Inspired by the related work [138], the distributions of the salient and background regions should have high probabilities at mutually exclusive domains. Thus, the spatial incompatibility term enforces that the same region should not have high foreground and background probabilities simultaneously, which is represented as:

$$E_i = \sum_{(k,l)\in\Omega_s} \omega_{kl} \cdot \widehat{s_k^t} \cdot \widehat{s_l^t} \tag{7-16}$$

When a highly probable salient region is surrounded by unlikely background neighbors, the spatial incompatibility energy is reduced. Therefore, for a low spatial incompatibility energy, the foreground and the background should form their own dominant regions.

Global Consistency Term. The salient objects in video should be salient with distinct motion patterns in each individual frame, and appear in most of the frames. Therefore, the salient objects should be consistently highlighted throughout the whole video sequences. However, most of the existing methods process the video frame by frame and ignore the global property across the whole video sequence. In this way, the saliency result only guarantees the local consistency rather than global consistency. In

this work, the global consistency term is proposed to constrain the consistency from the global perspective, which imposes the appearance of salient object approximate to a global video foreground model and is described as:

$$E_g = \sum_k \kappa_k \cdot \widehat{s_k^t}^2 \tag{7-17}$$

where $\kappa_k = \chi^2\left(\boldsymbol{h}_k, \boldsymbol{h}_{vs}\right)$ is the chi-square distance of Lab color histograms between the superpixel and video foreground model. The top 10 superpixels with larger initial saliency value in each frame are extracted as the foreground samples to represent the foreground distribution of the whole video.

To sum up, the energy function is defined as follows:

$$E = \eta_1 \cdot E_u + \eta_2 \cdot E_s + \eta_3 \cdot E_i + \eta_4 \cdot E_g \tag{7-18}$$

where η_i is the weighting parameter for balancing the relative influence of different components. Following [95], the weighting parameter η_1 for unary data term is set to 0.5 to constrain the updating change not to be large, and other weighting parameters are set to 1 with equal contribution.

Let $\boldsymbol{s} = [s_k]_{N_A \times 1}$, and $\hat{\boldsymbol{s}} = \left[\widehat{s_k}\right]_{N_A \times 1}$, where $N_A = \sum_{i=1}^{N} N^i$ is the total number of superpixels in the whole video. The energy function can be rewritten as the following matrix form:

$$\mathbf{E} = \eta_1 \cdot \left(\hat{\boldsymbol{s}} - \boldsymbol{s}\right)^T \cdot \left(\hat{\boldsymbol{s}} - \boldsymbol{s}\right) + \eta_2 \cdot \hat{\boldsymbol{s}}^T \cdot \left(\mathbf{D}_{st} - \mathbf{W}_{st}\right) \cdot \hat{\boldsymbol{s}} + \eta_3 \widehat{\boldsymbol{s}}^T \cdot \mathbf{W}_s \cdot \hat{\boldsymbol{s}} + \eta_4 \cdot \hat{\boldsymbol{s}}^T \cdot \mathbf{K} \cdot \hat{\boldsymbol{s}} \tag{7-19}$$

where $\mathbf{W}_{st} = \left[\omega_{kl}\right]_{N_A \times N_A}^{(k,l) \in \Omega_{st}}$ is the spatiotemporal color similarity matrix, $\mathbf{D}_{st} = diag\left(d_1, d_2, \cdots, d_{N_A}\right)$ denotes the degree matrix, $d_i = \sum_{j=1,(i,j) \in \Omega_{st}}^{N_A} \omega_{ij}$, $\mathbf{W}_s = \left[\omega_{kl}\right]_{N_A \times N_A}^{(k,l) \in \Omega_s}$ is the spatial color similarity matrix, and $\mathbf{K} = diag\left(\kappa_1, \kappa_2, \cdots, \kappa_{N_A}\right)$ is the difference matrix between the superpixels and global foreground model.

Combining these four quadratic function terms, the energy function is a convex function, which can be solved by setting its derivative with respect to $\hat{\boldsymbol{s}}$ to be 0. The transformation formula is represented as:

$$\eta_1 \cdot \left(\hat{\boldsymbol{s}} - \boldsymbol{s}\right) + \eta_2 \cdot \left(\mathbf{D}_{\mathrm{st}} - \mathbf{W}_{\mathrm{st}}\right) \cdot \hat{\boldsymbol{s}} + \eta_3 \cdot \mathbf{W}_s \cdot \hat{\boldsymbol{s}} + \eta_4 \cdot \mathbf{K} \cdot \hat{\boldsymbol{s}} = 0 \tag{7-20}$$

Then, the solution is obtained by:

$$\hat{\boldsymbol{s}} = \left[\eta_1 \cdot \mathbf{I} + \eta_2 \cdot \left(\mathbf{D}_{st} - \mathbf{W}_{st}\right) + \eta_3 \cdot \mathbf{W}_s + \eta_4 \cdot \mathbf{K}\right]^{-1} \cdot \left(\eta_1 \cdot \boldsymbol{s}\right) \tag{7-21}$$

where $\mathbf{I}$ is an identity matrix with the size of $N_A \times N_A$.

7.3 Experimental Results

7.3.1 Experimental Settings

The proposed approach is evaluated on the SegTrackV1 dataset [115], DAVIS dataset [117], and ViSal dataset [95]. For quantitative evaluation, three criteria including the PR curve, F-measure, and MAE score are used. In experiments, the number of superpixels for each frame is set to 500, and the number of seeds are set to $(Q_1, Q_2) = (250, 50)$. The proposed method is implemented by MATLAB 2014a on a Quad Core 3.7GHz workstation with 16GB RAM. The proposed method takes an average of 17.03 seconds to process one frame with a resolution of 854×480, in which the optical flow calculation costs 65% of the runtime, the single-frame saliency calculation takes 10% of the runtime, the inter-frame saliency costs 22% of the runtime, and the global optimization occupies 3% of runtime. In the future, a faster optical flow method with parallel technique can be used to reduce this cost further.

7.3.2 Comparison with State-of-the-art Methods

The proposed method is compared with 15 state-of-the-art methods, including 6 static image saliency methods for each frame (*i.e.*, HS [26], DSR [27], BSCA [30], RRWR [31], HDCT [32], and DCLC [33]), 2 co-saliency detection methods for each video (*i.e.*, CCS [79] and SCS [77]), and 7 video saliency detection methods (*i.e.*, SP [94], CVS [95], RWRV [96], SG [97], SGSP [101], STBP [99], and VFCN [105]), where VFCN is a deep learning based video saliency detection method. All the compared methods are implemented by the source codes or released results provided by the authors. The qualitative comparison of different methods on three datasets are illustrated in Fig. 7-3, and the quantitative evaluation results are reported in Fig. 7-4 and Table 7-1.

Visual results of different methods are shown in Fig. 7-3. For the image saliency model (*e.g.*, DSR and RRWR), it is difficult to extract the salient object completely and accurately from a complex scene due to the lack of motion perception and inter-frame constraint. For example, in the Flamingo video, two birds are both detected as the salient objects by the DSR and RRWR methods. In fact, only the front one is the unique salient object in the whole video. In other words, it is insufficient to directly use the

static saliency model to detect the salient object in video. In the Dog video, the salient object and the background have the similar color appearance, which leads to some backgrounds are wrongly detected as foregrounds by RRWR method. In the Lucia video, the bench is relatively static compared to the moving human, and should not be detected as the salient object in the video. However, the image saliency models fail to effectively suppress these regions without considering the motion constraints. In the Parachute video, some backgrounds are wrongly highlighted by the image saliency models due to the strong luminance. For the co-saliency detection model, benefiting from the introduction of inter-image correspondence, some backgrounds are effectively suppressed, such as the trees and lawns in the Lucia video. However, some foregrounds are missed through the CCS model, such as the salient objects in the Parachute video. Moreover, for the co-saliency model, it is difficult to distinguish motion related salient object from all foreground objects, such as the Flamingo video. Without introducing the motion cue, the back of the bird is wrongly retained by CCS method.

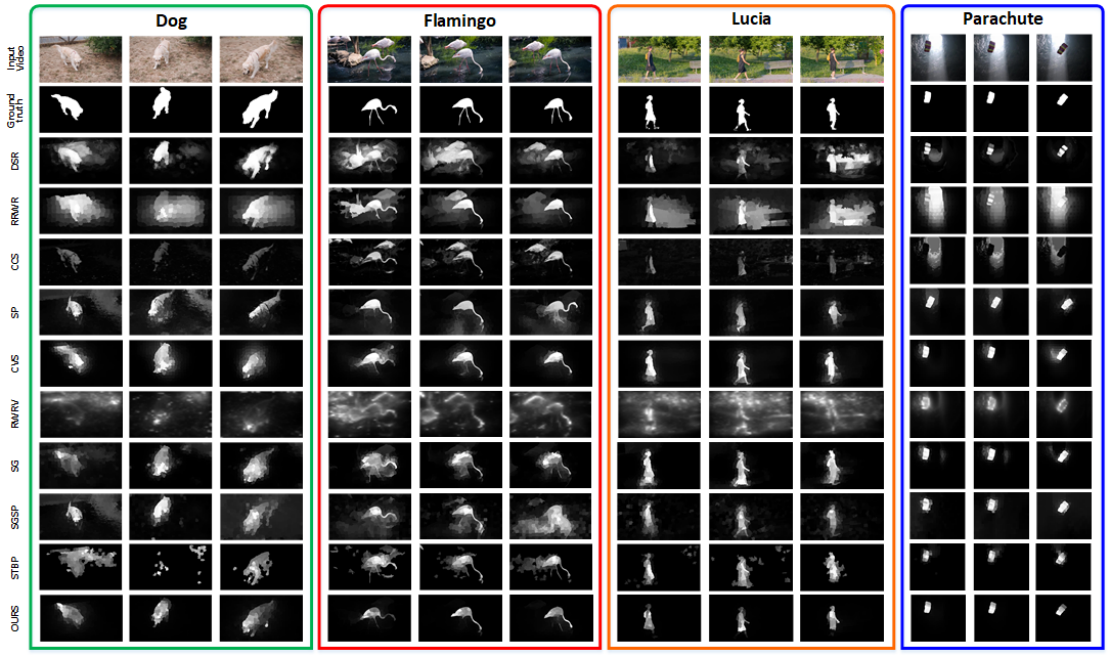

Fig. 7-3 Visual examples of different saliency detection methods.

By contrast, the video saliency detection methods produce better results. The proposed method achieves the best and most consistent performance compared with other methods. The salient objects are accurately and completely detected from some challenging videos, such as Flamingo video. Note that, other video saliency detection models either cannot exactly locate the salient object (*e.g.*, RWRV) or cannot effectively

suppress the background regions (*e.g.*, SG and SGSP). For example, in the Dog video, the salient object is not accurately and completely detected by the RWRV and STBP methods. In addition, some video saliency models fail to discover the salient object accurately from the clustered backgrounds, such as the SG and STBP models in the Flamingo video. The SGSP method induces many false positives in the background regions, and cannot locate the front bird perfectly. In the Lucia video, compared with other video saliency methods, superior performance in shape preserving and pinpointing is achieved through the proposed method.

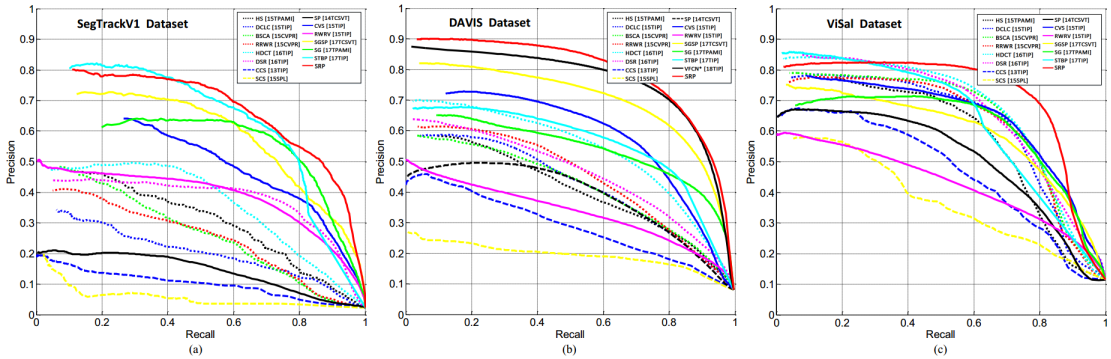

Fig. 7-4 PR curves of different methods on three datasets. (a) SegTrackV1 dataset. (b) DAVIS dataset. (c) ViSal dataset.

The PR curves are shown in Fig. 7-4. As visible, the proposed method achieves the highest precision of the whole PR curves on these three datasets with remarkable performance gain. In particular, on the DAVIS dataset, the proposed SRP method achieves better performance than the deep learning based video saliency method (*i.e.*, VFCN). The F-measure and MAE scores are reported in Table 7-1. From the table, it can be seen that the proposed method obtains the highest F-measure on these three datasets and the minimum MAE score on the ViSal dataset. The proposed method achieves the second and third places in term of MAE score on the DAVIS and SegTrackV1 datasets, respectively. In addition, the performance gains of the proposed method against others are more remarkable. Compared with the second best method in terms of F-measure, the percentage gain of the proposed method reaches 3.75% on the SegTrackV1 dataset, 2.19% on the DAVIS dataset, and 6.67% on the ViSal dataset. Moreover, the proposed unsupervised method is superior to the deep learning based VFCN method, and the percentage gain of F-measure achieves 2.19% on the DAVIS dataset. All the quantitative measures demonstrate the effectiveness of the proposed method.

Table 7-1 Quantitative comparisons with different methods on three datasets.

	SegTrackV1 Dataset		DAVIS Dataset		ViSal Dataset	
	F-measure	MAE	F-measure	MAE	F-measure	MAE
DCLC	0.2755	0.1496	0.4783	0.1350	0.6700	0.1265
DSR	0.4445	0.1305	0.4972	0.1303	0.6923	0.1061
RRWR	0.3267	0.1963	0.5089	0.1693	0.6707	0.1690
HS	0.3821	0.3142	0.4523	0.2505	0.6442	0.2019
BSCA	0.3579	0.2366	0.4680	0.1957	0.6949	0.1703
HDCT	0.4681	0.1268	0.5664	0.1346	0.7047	0.1282
CCS	0.1486	0.1437	0.3476	0.1510	0.5317	0.1427
SCS	0.1137	0.2664	0.2307	0.2567	0.4384	0.2523
SP	0.2159	0.1195	0.4616	0.1430	0.5723	0.1510
CVS	0.5370	0.1085	0.6212	0.1004	0.6676	0.1139
RWRV	0.4458	0.1511	0.3776	0.2001	0.4662	0.1903
SG	0.6218	0.0810	0.5553	0.1034	0.6640	0.1129
SGSP	0.6275	0.1258	0.6911	0.1374	0.6226	0.1772
STBP	0.6583	0.0342	0.5848	0.1015	0.6815	0.0987
VFCN	-	-	0.7488	0.0588	-	-
ours	0.6830	0.0949	0.7652	0.0688	0.7517	0.0924

7.3.3 Module Analysis

Each main component (*i.e.*, single-frame saliency reconstruction integrating the static saliency and motion saliency, inter-frame saliency with forward and backward propagations, and global optimization) is comprehensively evaluate on the DAVIS dataset, and the quantitative comparison results are presented in Fig. 7-5 and Table 7-2.

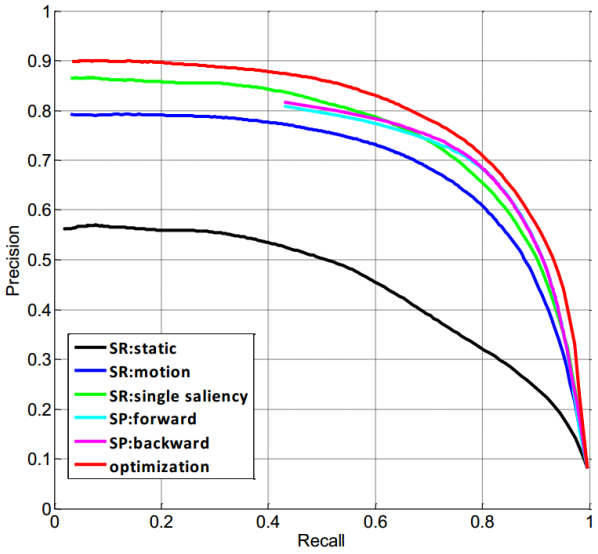

Fig. 7-5 PR curves of different modules of the proposed method on the DAVIS dataset.

Compared with the static saliency result, the motion saliency model achieves the higher precision of the P-R curves, and the F-measure is increased by 19.42%, which shows the effectiveness of the motion information in video saliency detection. Through the multiplying combination, the F-measure and MAE score of the single-frame saliency model reach 0.7358 and 0.0807, which is better than the other existing video saliency models. To fully capture the inter-frame relationship, the sparsity-based propagation with forward-backward strategy is proposed. As can be seen, the performance is further improved through the saliency propagation model, and the F-measure reaches 0.7381 after the backward propagation. Considering the spatiotemporal smoothness and global consistency, an optimization model is designed to improve the saliency map, and the output is regarded as the final video saliency result. From Fig. 7-5, the optimized result achieves the highest precision of the PR curves, which is marked by the red line. The same conclusion can be drawn from the F-measure reported in Table 7-2, which demonstrates the rationality and effectiveness of the optimization model. On the whole, the performance is gradually improved through the different modules in the proposed method.

Table 7-2 F-measures of different modules on the DAVIS dataset. SR: single-frame saliency reconstruction that integrates the static and motion saliencies. SP: inter-frame saliency propagation.

Modules		F-measure	MAE
SR	Static Saliency	0.5029	0.1206
	Motion Saliency	0.6971	0.0807
	Single Saliency	0.7358	0.0712
SP	Forward Propagation	0.7318	0.0924
	Backward Propagation	0.7381	0.0793
Global Optimization		0.7652	0.0688

7.3.4 Parameter Discussion

The influence of different seed numbers is comprehensively discussed, the tendency chart of F-measure on the DAVIS dataset is shown in Fig. 7-6. Generally, the salient regions in each frame are much smaller than the background regions. To explore the single-frame saliency, some background seeds are selected, and the number is denoted as Q_1. More background seeds can be chosen to construct a more complete background dictionary. For the inter-frame saliency propagation, the foreground seeds

are determined to propagate the sequential relationship across the time axis in a forward-backward way. The number of foreground seeds is denoted as Q_2. In order to avoid the introduction of interference, the number of foreground seeds should not be too large. In all the experiments, the ratio of Q_1 to Q_2 is fixed as 5:1. Selecting 100 or 120 background seeds for each frame is too small to completely reconstruct the single-frame saliency and will degenerate the performance. As the seed number increases, the performance becomes better, and the performance reaches optimum when (Q_1, Q_2) is set to (250,50). Subsequently, the performance begins to drop. The main reason is that too many seeds will introduce some false seed regions and decrease the reconstruction and propagation accuracy. As above, the performance is not highly sensitive to the parameter (Q_1, Q_2), and it is set to (250,50) in all experiments.

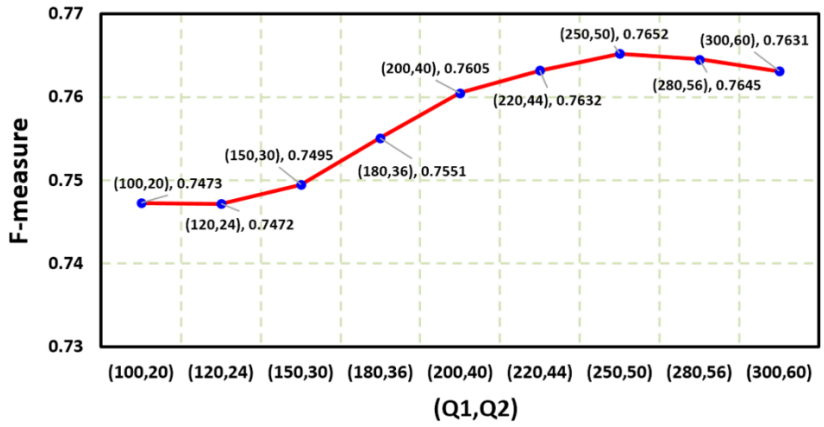

Fig. 7-6 F-measure of different (Q_1, Q_2) on the DAVIS dataset.

7.4 Summary

This chapter proposed a sparsity-based video saliency detection algorithm, which integrates a saliency reconstruction model, a saliency propagation model, and a global optimization model. Saliency reconstruction and propagation models leveraged on the novel motion priors to discover the salient objects. In addition, their sparse representations not only allowed them to extract the salient object from individual frames efficiently, but also captured the inter-frame correspondence along the time axis in a progressive way. Moreover, the performance was further improved by the global optimization model. The comprehensive analysis have demonstrated that the proposed method outperforms the state-of-the-art models.

Chapter 8 Conclusion and Future Work

8.1 Conclusion

In this thesis, the comprehensive information including the depth cue, inter-image correspondence, and spatiotemporal constraint are explicitly explored, the corresponding RGBD saliency model, co-saliency model, and video saliency model are innovatively designed, and the superior performances on large-scale benchmark datasets are achieved.

First, deep exploiting the depth information, a novel saliency detection method for stereoscopic images is presented based on depth confidence analysis and multiple cues fusion. Considering the influence of different qualities of depth map, a depth confidence measure is designed based on the depth distribution, which aims to reduce the negative effect of poor depth map on saliency detection. In addition, the compactness prior in color space is extended to depth domain, and a stereoscopic compactness saliency model is proposed by integrating color and depth information. In order to improve the robustness of the model, the foreground saliency based on multiple cues contrast with depth-refined foreground seeds selection scheme is combined with the compactness saliency to generate the final result.

Then, three co-saliency detection models for RGBD images are further investigated.

(1) The first model to address the co-saliency detection for RGBD images is proposed based on multi-constraint feature matching and cross label propagation. In this work, the depth information is worked as a useful complement of color feature, the similarity matching at the superpixel and image levels is designed to capture the inter-image corresponding relationship, and the Cross Label Propagation (CLP) scheme is proposed to optimize the intra and inter saliencies in a cross way and generate the final co-saliency map.

(2) Utilizing the existing single saliency result as the initialization, an iterative RGBD co-saliency detection framework is designed in a refinement-cycle manner. In

this work, a novel depth descriptor, named depth shape prior (DSP), is proposed to exploit the shape attributes from the depth map and enhance the identification of co-salient objects from RGBD images. In addition, the inter-image correspondence is modeled as a superpixel-level common probability function among multiple images in the deletion scheme, and the iterative updating strategy is introduced to improve the homogeneity and consistency in a cycle way. The proposed framework can effectively exploit any existing 2D saliency model to work well in RGBD co-saliency scenarios.

(3) In order to achieve a win-win situation of the accuracy and efficiency, a co-saliency detection method for RGBD images is presented based on hierarchical sparsity reconstruction and energy function refinement. In this work, the hierarchical sparsity framework is used to capture the corresponding relationship among multiple images, where the global foreground dictionary is built to reconstruct each image and capture the global inter-image correspondence, and a set of foreground dictionaries constructed by other images are utilized to reconstruct the current image and obtain multiple pairwise inter saliency maps from the local perspective. In addition, an energy function refinement model, including the unary data term, spatial smooth term, and holistic consistency term, is designed to improve the intra-image smoothness and inter-image consistency. Compared with the first two RGBD co-saliency models, sparsity representation is firstly used to capture the inter-image correspondence, which improve the performance and guarantee the efficiency. Moreover, the co-saliency detection optimization is first formulated as a global energy function optimization problem, which considers the holistic consistency among different images in the group.

Third, taking the spatial prior, motion cue, and temporal constraint into account, a video saliency detection model based on sparse reconstruction and propagation is proposed. The sparsity-based saliency reconstruction model is utilized to generate single-frame saliency map by making the best use of the static and motion priors. Then, an efficient sparsity-based saliency propagation model is used to capture the correspondence among different frames and produce the inter-frame saliency map, where the salient object is sequentially reconstructed by the forward and backward dictionaries. Finally, in order to attain the global and temporal consistency of the salient object in the whole video, a global optimization model is presented, which integrates unary data term, spatiotemporal smooth term, spatial incompatibility term, and global consistency term.

The extensive comparisons and comprehensive discussions on the corresponding benchmark datasets are conducted, which demonstrate that the proposed algorithms perform favorably against state-of-the-art methods both qualitatively and quantitatively.

8.2 Challenges

In the last decades, a plenty of saliency detection methods have been proposed to obtain the remarkable progresses and performance improvements. However, there still exist many issues that are not well resolved and needed to be further investigated.

For RGBD saliency detection, how to capture the accurate and effective depth representation to assist in saliency detection is a challenge. Taking the depth information as an additional feature to supplement color feature is an intuitive and explicit way, but it ignores the potential attributes in the depth map, such as shape and contour. By contrast, depth measure based method aims at exploiting these implicit information to refine the saliency result. For example, the depth shape can be used to highlight the salient object and suppress the background, and the depth boundary can be utilized to refine the object boundary and obtain sharper saliency result. In addition, the whole object usually has high consistency in the depth map. Therefore, the depth information can be used to improve the consistency and smoothness of the acquired saliency map. Generally, depth measure based methods can achieve a better performance. However, how to effectively exploit the depth information to enhance the identification of salient object has not yet reached a consensus. On the whole, combining the explicit and implicit depth information to obtain a more comprehensive depth representation is a meaningful attempt for RGBD saliency detection.

For co-saliency detection, how to explore inter-image correspondence among multiple images to constrain the common properties of salient object is a challenge. Inter-image corresponding relationship plays an essential role in determining the common object from all the salient objects, which can be formulated as a clustering process, a matching process, a propagation process, or a learning process. However, these methods may either be noise-sensitive or time-consuming. The accuracy of corresponding relationship is directly related to the performance of the algorithm. Thus, capturing the accurate inter-image correspondence is an urgent problem to be addressed. At present, there have been some attempts to detect co-salient object using deep

learning network. However, these methods often simply cascade the features produced from the single image and re-learn, rather than designing a specific inter-image network to learn the effective inter-image correspondence.

For video saliency detection, how to combine more information and constraints, such as motion cue, inter-frame correspondence, and spatiotemporal consistency, is a challenge. Motion cue plays more important role in discovering the salient object from the clustered and complex scene. The inter-frame correspondence represents the relationship among different frames, which is used to capture the common attribute of salient objects from the whole video. The spatiotemporal consistency constrains the smoothness and homogeneity of salient objects from the spatiotemporal domain. The main contributions of the existing methods are often concentrated in these three aspects. In addition, the video saliency detection algorithm based on deep learning is still immature, and only a few methods have been proposed, which is a relatively underexplored area. However, it is a challenging task to learn the comprehensive features including intra-frame, inter-frame, and motion through a deep network under the limited training samples.

8.3 Future Work

In the future, some research directions and emphases of saliency detection can be focused on:

(1) **New attempts in learning based saliency detection methods, such as small samples training, weakly supervised learning, and cross-domain learning.** Limited by the labelled training data, more work, such as designing a special network, can be explored in the future to achieve high-precision detection with small training samples. In addition, weakly supervised salient object detection method is a good choice to address the insufficient pixel-level saliency annotations. Furthermore, the cross-domain learning is another direction that needs to be addressed for learning based RGBD saliency detection method.

(2) **Extending the saliency detection task in different data sources, such as light filed image, RGBD video, and remote sensing image.** In the light filed image, the focusness prior, multi-view information, and depth cue should be considered jointly. For the RGBD video, the depth constraint should be introduced to assist in the

spatiotemporal saliency. In the remote sensing image, due to the high angle shot photographed, some small targets and shadows are included. Thus, how to suppress the interference effectively and highlight the salient object accurately should be further investigated in the future.

Bibliography

[1] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185-207, 2013.

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

[3] H. Fu, D. Xu, and S. Lin, “Object-based multiple foreground segmentation in RGBD video,” *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1418-1427, 2017.

[4] Z. Tao, H. Liu, H. Fu, and Y. Fu, “Image cosegmentation via saliencyguided constraint clustering with cosine similarity,” in *Proc. AAAI*, 2017, pp. 4285-4291.

[5] J. Sun, X. Liu, W. Wan, J. Li, D. Zhao, and H. Zhang, “Database saliency for fast image retrieval,” *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 359-369, 2015.

[6] Y. Gao, M. Shi, D. Tao, and C. Xu, “Video hashing based on appearance and attention features fusion via DBN,” *Neurocomputing*, vol. 213, pp. 84-94, 2016.

[7] J. Lei, M. Wu, C. Zhang, F. Wu, N. Ling, and C. Hou, “Depth preserving stereo image retargeting based on pixel fusion,” *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1442-1453, 2017.

[8] Y. Fang, K. Zeng, Z. Wang, W. Lin, Z. Fang, and C.-W. Lin, “Objective quality assessment for image retargeting based on structural similarity,” *IEEE J. Emerg. Sel. Topic Circuits Syst.*, vol. 4, no. 1, pp. 95-105, 2014.

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

[10] J. Lei, C. Zhang, Y. Fang, Z. Gu, N. Ling, and C. Hou, “Depth sensation enhancement for multiple virtual view rendering,” *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 457-469, 2015.

[11] C. Li, J. Guo, R. Cong, Y. Pang, and B. Wang, “Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior,” *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664-5677, 2016.

[12] J. Lei, L. Li, H. Yue, F. Wu, N. Ling, and C. Hou, “Depth map superresolution considering view synthesis quality,” *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1732-1745, 2017.

[13] J. Lei, J. Duan, F. Wu, N. Ling, and C. Hou, “Fast mode decision based on grayscale similarity and inter-view correlation for depth map coding in 3D-HEVC,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 706-718, 2018.

[14] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian, "Saliency-aware nonparametric foreground annotation based on weakly labeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1253-1265, 2016.

[15] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098-1110, 2016.

[16] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang, and B. Chen, "No-reference quality assessment of deblocked images," *Neurocomputing*, vol. 177, pp. 572-584, 2016.

[17] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Trans. Vis. Comput. Graph*, vol. 23, no. 8, pp. 2014-2027, 2017.

[18] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level CNN: Saliency-aware 3-D CNN with LSTM for video action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 510-514, 2017.

[19] H. Jacob, F. Padua, A. Lacerda, and A. Pereira, "Video summarization approach based on the emulation of bottom-up mechanisms of visual attention," *J. Intell. Information Syst.*, vol. 49, no. 2, pp. 193-211, 2017.

[20] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. ECCV*, 2012, pp. 101-115.

[21] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318-1334, 2013.

[22] S. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor system description, issues and solutions," in *Proc. CVPRW*, 2004, pp. 35-45.

[23] "Stereo camera," http://en.wikipedia.org/wiki/Stereocamera.

[24] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of cosaliency detection algorithms: Fundamentals, applications, and challenges," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, pp. 1-31, 2018.

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

[26] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 717-729, 2016.

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

[28] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by UFO: Uniqueness, focusness and objectness," in *Proc. ICCV*, 2013, pp. 1976-1983.

[29] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust

background detection," in *Proc. CVPR*, 2014, pp. 2814-2821.

[30] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. CVPR*, 2015, pp. 110-119.

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

[32] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 9-23, 2015.

[33] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu, "Salient region detection via integrating diffusion-based compactness and local contrast," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3308-3320, 2015.

[34] H. Peng, B. Li, H. Ling, W. Hua, W. Xiong, and S. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818-832, 2017.

[35] J. Lei, B. Wang, Y. Fang, W. Lin, P. L. Callet, N. Ling, and C. Hou, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783-1795, 2016.

[36] Z. Wang, D. Xiang, S. Hou, and F. Wu, "Background-driven salient object detection," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 750-762, 2017.

[37] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, 2013, pp. 3166- 3173.

[38] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. CVPR*, 2014, pp. 280-287.

[39] R. Achanta, S. Hemami, F. Estrada, and S. Ssstrunk, "Frequency-tuned salient region detection," in *Proc. CVPR*, 2009, pp. 1597-1604.

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

[41] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. CVPR*, 2013, pp. 2083-2090.

[42] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251-268, 2017.

[43] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135-1149, 2015.

[44] S. He, R. W. Lau, W. Liu, Z. Huang, and Q. Yang, "SuperCNN: A superpixelwise convolutional neural network for salient object detection," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 330-344, 2015.

[45] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. CVPR*, 2016, pp. 660-668.

[46] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. CVPR*, 2016, pp. 478-487.

[47] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. CVPR*, 2016, pp. 678-686.

[48] J. Zhang, Y. Dai, and F. Porikli, "Deep salient object detection by integrating multi-level cues," in *Proc. WACV*, 2017, pp. 1-10.

[49] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. CVPR*, 2017, pp. 5300-5309.

[50] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. ICCV*, 2017, pp. 212-221.

[51] J. Zhang, B. Li, Y. Dai, F. Porikli, and M. He, "Integrated deep and shallow networks for salient object detection," in *Proc. ICIP*, 2017, pp. 271-276.

[52] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. CVPR*, 2015, pp. 1265-1274.

[53] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. CVPR*, 2015, pp. 5455-5463.

[54] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proc. CVPR*, 2017, pp. 3796-3806.

[55] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *Proc. CVPR*, 2018, pp. 9029-9038.

[56] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. CVPR*, 2012, pp. 454-461.

[57] J. Lei, H. Zhang, L. You, C. Hou, and L. Wang, "Evaluation and modeling of depth feature incorporated visual attention for salient object segmentation," *Neurocomputing*, vol. 120, pp. 24-33, 2013.

[58] Y. Fang, J. Wang, M. Narwaria, P. L. Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2625-2636, 2014.

[59] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A

benchmark and algorithms," in *Proc. ECCV*, 2014, pp. 92-109.

[60] H. Song, Z. Liu, H. Du, G. Sun, O. L. Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204-4216, 2017.

[61] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274-2285, 2017.

[62] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171-3183, 2018.

[63] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. CVPR*, 2018, pp. 3051-3060.

[64] J. Guo, T. Ren, J. Bei, and Y. Zhu, "Salient object detection in RGBD image based on saliency fusion and propagation," in *Proc. ICIMCS*, 2015, pp. 1-5.

[65] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Process.: Image Commun.*, vol. 38, pp. 115-126, 2015.

[66] J. Guo, T. Ren, and J. Bei, "Salient object detection in RGB-D image via saliency evolution," in *Proc. ICME*, 2016, pp. 1-6.

[67] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. CVPR*, 2016, pp. 2343-2350.

[68] H. Song, Z. Liu, H. Du, and G. Sun, "Depth-aware saliency detection using discriminative saliency fusion," in *Proc. ICASSP*, 2016, pp. 1626-1630.

[69] H. Sheng, X. Liu, and S. Zhang, "Saliency analysis based on depth contrast increased," in *Proc. ICASSP*, 2016, pp. 1347-1351.

[70] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 663-667, 2017.

[71] H. Li and K. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365-3375, 2011.

[72] K. Chang, T. Liu, and S. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. CVPR*, 2011, pp. 2129-2136.

[73] Z. Tan, L. Wan, W. Feng, and C.-M. Pun, "Image co-saliency detection by propagating superpixel affinities," in *Proc. ICASSP*, 2013, pp. 2114-2118.

[74] H. Li, F. Meng, and K. Ngan, "Co-salient object detection from multiple images,"

IEEE Trans. Multimedia, vol. 15, no. 8, pp. 1869-1909, 2013.

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

[76] Z. Liu, W. Zou, L. Li, L. Shen, and O. L. Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 88-92, 2014.

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

[78] Y. Zhang, L. Li, R. Cong, X. Guo, H. Xu, and J. Zhang, "Co-saliency detection via hierarchical consistency measure," in *Proc. ICME*, 2018, pp. 1-6.

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

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

[81] R. Huang, W. Feng, and J. Sun, "Saliency and co-saliency detection by low-rank multiscale fusion," in *Proc. ICME*, 2015, pp. 1-6.

[82] C. Ge, K. Fu, F. Liu, L. Bai, and J. Yang, "Co-saliency detection via inter and intra saliency propagation," *Signal Process.: Image Commun.*, vol. 44, pp. 69-83, 2016.

[83] R. Huang, W. Feng, and J. Sun, "Color feature reinforcement for cosaliency detection without single saliency residuals," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 569-573, 2017.

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

[85] L. Wei, S. Zhao, O. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *Proc. IJCAI*, 2017, pp. 3041-3047.

[86] D. Zhang, D. Meng, C. Lia, L. Jiang, Q. Zhao, and J. Han, "A selfpaced multiple-instance learning framework for co-saliency detection," in *Proc. ICCV*, 2015, pp. 594-602.

[87] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning based for co-saliency detection framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473-2483, 2018.

[88] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based RGBD image cosegmentation with mutex constraint," in *Proc. CVPR*, 2015, pp. 4428-4436.

[89] H. Song, Z. Liu, Y. Xie, L. Wu, and M. Huang, "RGBD co-saliency detection via bagging-based clustering," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1722-1726, 2016.

[90] Z. Ren, S. Gao, D. Rajan, L.-T. Chia, and Y. Huang, "Spatiotemporal saliency detection via sparse representation," in *Proc. ICME*, 2012, pp. 158-163.

[91] Y. Xue, X. Guo, and X. Cao, "Motion saliency detection using low-rank and sparse decomposition," in *Proc. ICASSP*, 2012, pp. 1485-1488.

[92] Y. Fang, W. Lin, Z. Chen, C. Tsai, and C. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27-38, 2014.

[93] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3910-3921, 2014.

[94] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522-1540, 2014.

[95] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185-4196, 2015.

[96] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2552-2564, 2015.

[97] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. CVPR*, 2015, pp. 3395-3402.

[98] W. Wang, J. Shen, R. Yang, and F. Porikli, "A unified spatiotemporal prior based on geodesic distance for video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20-33, 2018.

[99] T. Xi, W. Zhao, H. Wang, and W. Lin, "Salient object detection with spatiotemporal background priors for video," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3425-3436, 2017.

[100] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3156-3170, 2017.

[101] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527-2542, 2017.

[102] F. Guo, W. Wang, J. Shen, L. Shao, J. Yang, D. Tao, and Y. Y. Tang, "Video saliency detection using object proposals," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3159-3170, 2018.

[103] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient

object," in *Proc. CVPR*, 2007, pp. 1-8.

[104] T.-N. Le and A. Sugimoto, "Spatiotemporal utilization of deep features for video saliency detection," in *Proc. ICMEW*, 2017, pp. 465-470.

[105] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38-49, 2018.

[106] T.-N. Le and A. Sugimoto, "Deeply supervised 3D recurrent FCN for salient object detection in videos," in *Proc. BMVC*, 2017, pp. 1-13.

[107] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 349-364, 2018.

[108] "MSRA10K," http://mmcheng.net/gsal/.

[109] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors," in *Proc. CVPR*, 2017, pp. 4142-4150.

[110] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. ICCV*, 2005, pp. 1800-1807.

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

[112] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215-232, 2016.

[113] K. Li, J. Zhang, and W. Tao, "Unsupervised co-segmentation for indefinite number of common foreground objects," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1898-1909, 2016.

[114] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568-579, 2018.

[115] D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *Proc. BMVC*, 2010, pp. 1–11.

[116] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. ICCV*, 2013, pp. 2192–2199.

[117] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. CVPR*, 2016, pp. 724–732.

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

[119] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274-2282, 2012.

[120] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189-2202, 2012.

[121] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 169-176.

[122] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 17, pp. 971-987, 2002.

[123] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. ICIP*, 2014, pp. 1115-1119.

[124] X. Wei, Z. Tao, C. Zhang, and X. Cao, "Structured saliency fusion based on dempster-shafer theory," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1345-1349, 2015.

[125] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on markov absorption probabilities," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1639-1649, 2015.

[126] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Inner and inter label propagation: Salient object detection in the wild," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3176-3186, 2015.

[127] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. ACM-SIAM Symp. Discr. Algorithms*, 2007, pp. 1027-1035.

[128] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145-175, 2001.

[129] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv: 1409.1556*, pp. 1-14, 2014.

[130] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819-823, 2016.

[131] T. Leung and J. Malik, "Recognizing surface using three-dimensional textons," in *Proc. ICCV*, 1999, pp. 1010-1017.

[132] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037-2041, 2006.

[133] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R^3Net: Recurrent

residual refinement network for saliency detection," in *Proc. IJCAI*, 2018, pp. 684-690.

[134] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "An iterative co-saliency framework for RGBD images," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 233-246, 2019.

[135] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. CVPR*, 2015, pp. 5216-5223.

[136] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Dense and sparse labeling with multidimensional features for saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1130-1143, 2018.

[137] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500-513, 2011.

[138] W.-D. Jang, C. Lee, and C.-S. Kim, "Primary object segmentation in videos via alternate convex optimization of foreground and background distributions," in *Proc. CVPR*, 2016, pp. 696-704.

Research Achievements

List of Publications:

1. **Runmin Cong**, Jianjun Lei, Huazhu Fu, Weisi Lin, Qingming Huang, Xiaochun Cao, and Chunping Hou, "An iterative co-saliency framework for RGBD images," IEEE Transactions on Cybernetics, vol. 49, no. 1, pp. 233-246, 2019. (SCI, JCR Q1, IF=8.803)
2. **Runmin Cong,** Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," IEEE Transactions on Image Processing, vol. 27, no. 2, pp. 568-579, 2018. (SCI, JCR Q1, CCF-A, IF=5.071)
3. **Runmin Cong**, Jianjun Lei, Huazhu Fu, Fatih Porikli, Qingming Huang, and Chunping Hou, "Video saliency detection via sparsity-based reconstruction and propagation," IEEE Transactions on Image Processing, DOI: 10.1109/TIP.2019.2910377, 2019. (SCI, JCR Q1, CCF-A, IF=5.071)
4. **Runmin Cong**, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Nam Ling, "HSCS: Hierarchical sparsity based co-saliency detection for RGBD images," IEEE Transactions on Multimedia, DOI: 10.1109/TMM.2018.2884481, 2018. (SCI, JCR Q1, IF=3.977)
5. **Runmin Cong**, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang, "Review of visual saliency detection with comprehensive information," IEEE Transactions on Circuits and Systems for Video Technology, DOI: 10.1109/TCSVT.2018.2870832, 2018. (SCI, JCR Q1, IF=3.558)
6. **Runmin Cong**, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," IEEE Signal Processing Letters, vol. 23, no. 6, pp. 819-823, 2016. (SCI, JCR Q2, IF=2.831)
7. **Runmin Cong**, Ping Han, Chongyi Li, Jiaji He, and Zaiji Zhang, "Manmade target extraction based on multi-stage decision and its application for change detection in

polarimetric synthetic aperture radar image," Journal of Electronic Imaging, vol. 25, no. 5, pp. 1-13, 2016. (SCI, JCR Q4, IF=0.780)

8. Min Ni, Jianjun Lei, **Runmin Cong***, Kaifu Zheng, Bo Peng, and Xiaoting Fan, "Color-guided depth map super resolution using convolutional neural network," IEEE Access, vol. 2, pp. 26666-26672, 2017. (SCI, JCR Q1, IF=3.557, *corresponding author)
9. Ping Han, Binbin Han, Xiaoguang Lu, **Runmin Cong***, and Dandan Sun, "Unsupervised classification of PolSAR images based on multi-level feature extraction," International Journal of Remote Sensing, 2019. (SCI, JCR Q2, IF=1.782, *corresponding author)
10. **Runmin Cong**, Jianjun Lei, Huazhu Fu, Wenguan Wang, Qingming Huang, and Lijie Niu, "Research progress of video saliency detection," Journal of Software, vol. 29, no. 8, pp. 2527-2544, 2018. (EI, in Chinese)
11. Chongyi Li, Jichang Guo, **Runmin Cong**, Yanwei Pang, and Bo Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5664-5677, 2016. (SCI, JCR Q1, CCF-A, IF=5.071)
12. Chunle Guo, Chongyi Li, Jichang Guo, **Runmin Cong**, Huazhu Fu, and Ping Han, "Hierarchical features driven residual learning for depth map super-resolution," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2545-2557, 2019. (SCI, JCR Q1, CCF-A, IF=5.071)
13. Hua Li, Sam Kwong, Chuanbo Chen, Yuheng Jia, and **Runmin Cong**, "Superpixel segmentation based on square-wise asymmetric segmentation and structural approximation," IEEE Transactions on Multimedia, DOI: 10.1109/TMM.2019.2907047, 2019. (SCI, JCR Q1, IF=3.977)
14. Mengxin Han, **Runmin Cong**, Xinyu Li, Huazhu Fu, and Jianjun Lei, "Joint spatial-spectral hyperspectral image classification based on convolutional neural network," Pattern Recognition Letters, DOI: 10.1016/j.patrec.2018.10.003, 2018. (SCI, JCR Q2, IF=1.952)
15. Chongyi Li, Jichang Guo, Chunle Guo, **Runmin Cong**, and Jiachang Gong, "A hybrid method for underwater image correction," Pattern Recognition Letters, vol. 94, pp. 62-67, 2017. (SCI, JCR Q2, IF=1.952)

16. Chongyi Li, Jichang Guo, Bo Wang, **Runmin Cong**, Yan Zhang, and Jian Wang, "Single underwater image enhancement based on color cast removal and visibility restoration," Journal of Electronic Imaging, vol. 25, no. 3, pp. 1-16, 2016. (SCI, JCR Q4, IF=0.780)
17. Yonghua Zhang, Liang Li, **Runmin Cong**, Xiaojie Guo, Hui Xu, and Jiawan Zhang, "Co-saliency detection via hierarchical consistency measure," IEEE ICME, pp. 1-6, 2018. (CCF-B, **Best Student Paper Runner-up**)

Papers under review:

18. **Runmin Cong**, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong, "Going from RGB to RGBD saliency: A depth-guided transformation model," IEEE Transactions on Cybernetics, 2019. (SCI, JCR Q1, IF=8.803, **Minor Revisions**)
19. **Runmin Cong**, Hao Chen, Hongyuan Zhu, and Huazhu Fu, "Foreground detection and segmentation in RGB-D images," in *Paul Rosin, Yukun Lai, Yonghuai Liu, Ling Shao, RGB-D Image Analysis and Processing*, Springer, 2018. (**Book Chapter**)

List of China Patents:

1. **Runmin Cong**, Jianjun Lei, Chunping Hou, Chongyi Li, Xiaoxu He, and Jinhui Duan, "A saliency detection method for stereoscopic images," Patent No.: ZL 201610244589.9, Granted Date: 2018.08. (**Granted China Patent**)
2. Jianjun Lei, **Runmin Cong**, Chunping Hou, Jinhui Duan, and Dongyang Li, "A confidence measure for depth map," Patent No.: ZL 201610242241.6, Granted Date: 2018.08. (**Granted China Patent**)
3. Jianjun Lei, **Runmin Cong**, Chunping Hou, Jing Zhang, Xiaoting Fan, and Bo Peng, "A transformation model from RGB saliency to RGBD saliency," Application No.: 201910375809.5, Application Date: 2019.05 (Application China Patent).
4. Jianjun Lei, **Runmin Cong**, Zhe Zhang, Xinxin Zhu, Yuxin Song, and Yalong Jia, "A video saliency detection method," Application No.: 201910266112.4, Application Date: 2019.04 (Application China Patent).

5. Jianjun Lei, **Runmin Cong**, Chunping Hou, Chongyi Li, Liying Xu, and Zhe Zhang, "A cosaliency detection method for RGBD images," Application No.: 201810879724.6, Application Date: 2018.10. (Application China Patent)
6. Jianjun Lei, **Runmin Cong**, Chunping Hou, Sanyi Zhang, Yue Chen, and Yan Guo, "An iterative co-saliency detection method," Application No.: 201711064083.0, Application Date: 2017.11. (Application China Patent)
7. Jianjun Lei, **Runmin Cong**, Chunping Hou, Xinxin Li, Mengxin Han, and Xiaowei Luo, "A depth shape prior extraction method," Application No.: 201711065005.2, Application Date: 2017.11. (Application China Patent)
8. Jianjun Lei, **Runmin Cong**, Chunping Hou, Bo Peng, Xiaoting Fan, and Jing Zhang, "An inter-image saliency detection method," Application No.: 201710942099.0, Application Date: 2017.10. (Application China Patent)
9. Jianjun Lei, **Runmin Cong**, Chunping Hou, Jing Zhang, Xiaoting Fan, and Bo Peng, "A co-saliency detection method," Application No.: 201710942783.9, Application Date: 2017.10. (Application China Patent)

List of Participated Projects:

1. (**Principal Investigator**) "Research on visual saliency detection with comprehensive information", Outstanding Doctoral Dissertation Foundation of Tianjin University, 2018/11-2019/06.
2. (Principal Participator) "3D video coding and processing", The National Natural Science Foundation of China (Grant 61722112), 2018/01-2019/06.
3. (Principal Participator) "Real-time real 3D display technology for big data applications", The National Key R&D Program of China (Grant 2017YFB1002900), 2016/08-2019/06.
4. (Principal Participator) "Research on multiview video coding theory and technologies based on stereoscopic vision saliency", The National Natural Science Foundation of China (Grant 61271324), 2015/04-2016/12.

List of Honors & Awards

1. First Prize for Tianjin Scientific and Technological Progress Award, 2018
2. IEEE ICME Best Student Paper Award Runner-Up, 2018

3. National Scholarship for Ph.D Candidates, 2018
4. Special Award of Innovation Scholarship in Tianjin Municipality, 2018
5. Outstanding Doctoral Dissertation Foundation of Tianjin University, 2018
6. “National Model” Scholarship of Tianjin University, 2017
7. Academic Scholarship of Tianjin University, 2015-2018
8. Advanced Individual in Technological Innovation of Tianjin University, 2016-2018
9. Advanced Individual in International Exchange of Tianjin University, 2017/2018
10. Merit Student of Tianjin University, 2016/2018
11. Excellent Student Cadre of Tianjin University, 2017
12. Top Ten Excellent Youth in the School of EIE, Tianjin University, 2016

Acknowledgements

I would like to sincerely thank all the people who supported me through the time as Ph.D. student and who made this thesis possible in the first place.

First and foremost, I would like to gratefully thank my supervisors Prof. Qingming Huang and Prof. Jianjun Lei for their perpetual patience, constant encouragement, consistent support, and meticulous guidance. In the past four years, they not only gave me enough freedom to explore my research topics, but also gave me many constructive suggestions and discussions to help me get rid of all the difficulties in study and life. Without their help, I can not imagine the completion of this thesis. Besides, I am deeply grateful for invaluable advice and precious opportunity from my co-supervisors Prof. Weisi Lin and Prof. Sam Kwong, who hosted me as a visiting student at Nanyang Technological University (NTU) and City University of Hong Kong (CityU), respectively. I gratefully appreciate Dr. Huazhu Fu in Inception Institute of Artificial Intelligence, who helped me to learn and improve my writing word-for-word. I also thank Prof. Fatih Porikli in The Australian National University (ANU), Prof. Nam Ling in Santa Clara University (SCU), Prof. Xiaochun Cao in Institute of Information Engineering, Chinese Academy of Sciences (IIE CAS), Prof. Ming-Ming Cheng in Nankai University (NKU), Dr. Junhui Hou in CityU, for their invaluable discussions, constructive comments, and providing such a free environment between research groups.

The most enjoyable thing during my Ph.D study is the opportunity to work with so many talented research colleagues, including Chunping Hou, Jingyu Yang, Yonghong Hou, Sumei Li, Huanjing Yue, Changqing Zhang, Xiaojie Guo, Bo Peng, Xiaoting Fan, Jing Zhang, Min Ni, Kaifu Zheng, Ning Zhang, Zhe Zhang, Lijie Niu, Mengxin Han, Yue Chen, and Xiaowei Luo in TJU; Shiqi Wang, Liming Zhan, Yue Qian, Mengyuan Wu, Ran Wang, Xu Wang, Wei Gao, Yuheng Jia, Wenhui Wu, Xinqi Li, Xuelin Shen, Yi Chen, Jing Jin, Xuekai Wei, and Zhangkai Ni in CityU; Xinfeng Zhang, Leida Li, Shasha Mao, Ke Gu, Qiuping Jiang, Yupeng Cheng, and Sheng Yang in NTU, and Qianqian Xu, Hua Zhang, Wenqi Ren, Liang Yang, and Ling Du in IIE CAS. Moreover, special thanks also go to my close friends Chongyi Li and Sanyi Zhang

for being always there to discuss life choices and future plans. In addition, I would like to thank my hiking and rock climbing partners Mengyuan Wu, Ran Wang, and Yingjie Li, which expanded my experience and enriched my life.

Last but not least, I want to dedicate this thesis to my parents for their enduring love, support, and understanding. Because of you, I never feel alone and frustrated through difficult times. I have been doing my best, and I am keen to share the successes with you.