中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-28

论文引用格式:Zhou Qiangqiang, Yu Jiacong, Xu Jiawei, Chen Yong, Huang Xin. Panoramic Visual Saliency Detection: A Survey of Principles, Methods and Applications[J/OL]. Journal of Image and Graphics, XXXX:1-28. DOI: 10.11834/jig.250560.(周强强, 喻嘉聪, 徐佳伟, 陈勇, 黄欣, 石艳娇, 张晴. 全景视觉显著性检测:原理、方法与应用综述[J/OL]. 中国图象图形学报,XXXX:1-28. DOI: 10.11834/jig.250560.)[DOI: 10.11834/jig.250560]

全景视觉显著性检测:原理、方法与应用综述

周强强[1],喻嘉聪[1],徐佳伟[1],陈勇[1],黄欣[1],石艳娇[2],张晴[2]

1. 江西师范大学人工智能学院,南昌　330000; 2. 上海应用技术大学智能技术学部,上海　201418

摘　要: 随着虚拟现实(Virtual Reality,VR)技术的迅速普及,VR360°全景图像(Omnidirectional Image,ODI)与全景视频(Omnidirectional Video, ODV)在娱乐、教育、医疗等领域展现出巨大的应用潜力。然而,由于全景内容具有球面畸变、视角分布不均以及实时交互等特有挑战,传统显著性检测方法难以有效应对VR全景场景的复杂性。本文综述了当前VR360°全景图像/视频下显著性检测的研究进展,从传统机器学习方法到基于深度学习(如CNN、Transformer和LSTM架构)的方法进行了全面回顾。文章首先介绍了VR360°内容的成像原理与几何特性;随后,重点讨论了传统方法与深度学习技术在全景显著性检测中的应用与局限,特别是在全景环境中融合了多模态信息的相关研究;此外,综述还对现有数据集、评测指标及现有方法的性能进行了系统整理。本文还考察了全景显著性检测在图像质量评估和视频质量评估等领域中的实际应用案例,以揭示其在技术优化和用户体验提升方面的潜在价值。最后,文章展望了未来在VR360°全景显著性检测领域的前沿方向,旨在为后续研究和技术落地提供理论支持和实践指南。本文提及的算法、数据集已汇总至https://github.com/jiacongyu/PVSD。

关键词: 虚拟现实;全景显著性预测;全景显著性目标检测;深度学习;视觉注意机制;沉浸式体验

Panoramic Visual Saliency Detection: A Survey of Principles, Methods and Applications

Zhou Qiangqiang[1], Yu Jiacong[1], Xu Jiawei[1], Chen Yong[1], Huang Xin[1]

1. School of Artificial Intelligence, Jiangxi Normal University, Nanchang 330000, China; 2. Department of Intelligent Technology, Shanghai University of Technology, Shanghai 201418, China

Abstract: With the rapid advancement and widespread adoption of virtual reality (VR) technology, 360° omnidirectional images (ODI) and videos (ODV) have emerged as transformative tools across diverse domains such as entertainment, education, healthcare, gaming, and immersive training simulations, offering unparalleled panoramic experiences by capturing a full 360° field of view that empowers users to freely navigate and interact with content in a highly engaging manner, redefining user interaction by simulating real-world environments and providing a sense of presence that conventional 2D media cannot replicate. However, the inherent challenges of ODI and ODV, including spherical distortions arising from various projection techniques, uneven viewpoint distributions due to user-controlled perspectives, and the pressing need for real-time processing to maintain seamless immersion, pose significant obstacles for traditional saliency detection methods,

收稿日期:2025-11-06;**修回日期**:2026-03-05

*通信作者:周强强,男,讲师,主要研究方向为计算机视觉、模式识别。E-mail:qiang@jxnu.edu.cn

基金项目:国家自然科学基金地区基金项目(62262030,62262029);江西省自然科学基金面上项目(20232BAB202021)

Supported by: National Natural Science Foundation of China (Grant No. 62262030, No. 62262029); Jiangxi Provincial Natural Science Foundation 20232BAB202021

which are typically designed for planar 2D content and often struggle to handle the spatial and temporal complexities unique to VR environments, making saliency detection—aimed at identifying regions of interest that attract human attention—even more critical in VR settings where users can explore content from any angle, necessitating models that can predict attention across an entire spherical field. This survey provides an in-depth and comprehensive review of the latest developments in saliency detection for VR 360° content, spanning a wide spectrum of approaches from traditional machine learning techniques rooted in handcrafted features to cutting-edge deep learning methodologies, including convolutional neural networks (CNNs), Transformers for capturing global dependencies, and long short-term memory (LSTM) architectures tailored for temporal dynamics in video sequences, beginning with an outline of the fundamental principles underpinning VR 360° content, delving into intricate imaging mechanisms, geometric properties of spherical data representations, and distinct characteristics—such as boundary discontinuities and projection-induced artifacts—that set it apart from conventional 2D images and videos, thereby necessitating specialized approaches for effective saliency modeling. The survey systematically categorizes existing methods into key thematic areas: adaptations of traditional feature-based techniques relying on low-level cues like color, contrast, and texture; advanced deep learning architectures designed for robust feature extraction under spherical constraints; multi-projection domain modeling addressing challenges of different projection formats such as equirectangular, cubic, and spherical projections; and multimodal fusion strategies integrating diverse data sources including visual, auditory, and depth information to enhance prediction accuracy, noting that traditional methods, while computationally efficient and straightforward, often fall short in addressing spherical distortions and spatial discontinuities, as seen in early approaches like superpixel-based segmentation struggling with boundary effects and color dictionary sparse representation lacking adaptability to dynamic VR content, whereas deep learning methods have revolutionized the field by enabling end-to-end learning and delivering superior performance in handling dynamic viewpoints and multimodal data, with notable examples including CNN-based models like SalGCN leveraging spherical graph convolutions to mitigate projection distortions, Transformer-based frameworks such as SalViT360 excelling in capturing long-range dependencies across 360° videos by modeling global context, and LSTM architectures like HiBayes-LSTM enhancing temporal modeling for scanpath prediction by integrating user behavior data for higher precision. A critical component of this review is the detailed examination of available datasets and evaluation metrics pivotal for benchmarking and advancing research, summarizing prominent datasets for ODI and ODV saliency detection such as Salient360! with extensive head and eye-tracking annotations, HTRO focusing on diverse VR scenarios, and PAVS10K, a large-scale dataset for panoramic video saliency, highlighting their scale, annotation types like fixation maps and scanpaths, and applicability to tasks like saliency prediction (SP) and salient object detection (SOD), alongside an in-depth analysis of evaluation metrics including Pearson Correlation Coefficient (CC) for saliency map similarity, Normalized Scanpath Saliency (NSS) for gaze alignment, Kullback-Leibler Divergence (KLD) for divergence measurement, precision-recall-based F-measure, error metrics like Mean Absolute Error (MAE), and structural similarity metrics such as S-measure and E-measure, discussing their effectiveness in VR contexts and the need for adaptations to address 360°-specific challenges like distortion at poles in equirectangular projections, with comparative analyses revealing that deep learning approaches, particularly those incorporating multimodal fusion and advanced projections, consistently outperform traditional methods, achieving higher scores on key metrics like Area Under Curve (AUC) and CC. Furthermore, we explore a wide array of practical applications of saliency detection in VR across critical areas such as image and video quality assessment, adaptive compression for bandwidth optimization, and virtual cinematography for automated content creation, where saliency-guided models like SG360BIQA enhance quality prediction accuracy by prioritizing user-attended regions to align objective metrics with subjective experience, techniques like RoSal360 optimize bitrate allocation by assigning higher resources to salient areas based on saliency maps for improved transmission efficiency in resource-constrained VR systems, and attention-driven deep reinforcement learning models simulate human gaze patterns for smoother, intuitive VR experiences by guiding camera movements in virtual cinematography, underscoring the transformative potential of saliency detection in optimizing VR systems and elevating user immersion across use cases from immersive storytelling to teleconferencing. Despite remarkable progress, numerous challenges persist, including handling geometric distortions inherent in spherical representations, ensuring real-time processing to meet stringent latency requirements of VR applications, and addressing the scarcity of unified multimodal datasets integrating

visual, auditory, and interaction data for holistic modeling, prompting future research directions like expanding multi-modal integration with depth cues and real-time user interaction data for context-aware models, developing lightweight and unsupervised learning frameworks to reduce computational overhead while maintaining accuracy, and establishing large-scale, standardized benchmarks for reproducible research and innovation, while emerging techniques such as transfer learning for cross-domain adaptation, contrastive learning for feature discrimination, and data augmentation tailored to spherical geometries could enhance model generalizability and tackle data limitations in this nascent field. In summary, this survey contributes a thorough, detailed analysis of saliency detection for VR 360° content, bridging the gap between traditional heuristic-based methodologies and modern deep learning approaches leveraging computational power and data-driven insights, offering comprehensive insights into their practical deployment in real-world VR applications, critically addressing current limitations like dataset scarcity and computational constraints, and exploring emerging trends such as multimodal fusion and unsupervised learning to pave the way for future advancements, aiming to drive VR technology toward more efficient, immersive, and user-centric applications that transform how we interact with digital environments and shape the future of immersive media and human-computer interaction. The algorithms and datasets mentioned in this article have been summarized athttps://github. com/jiacongyu/PVSD.

Key words: Virtual reality; Panoramic saliency prediction; Panoramic salient object detection; Deep learning; Visual attention mechanism; Immersive experience

0 引　言

虚拟现实(VR)技术是一种通过360°全景图像构建沉浸式虚拟环境,随着其迅猛发展,全景内容在医疗、教育和娱乐等领域应用日益广泛。与传统平面图像不同,VR360°全景环境的内容以球面投影形式呈现,为用户提供了沉浸式体验,但其高分辨率和动态视角的特性也带来了存储与传输的挑战。

显著性检测(Saliency Detection,SD)是模拟人类视觉注意机制的计算机视觉技术,旨在识别图像中最重要的区域。该技术包括显著性预测(Saliency Prediction,SP)和显著性目标检测(Salient Object Detection,SOD)两个主要方向:前者预测注视点分布,后者检测和分割显著物体。此外,扫视路径预测(Scanpath Prediction)作为显著性预测的延伸,预测人眼观看图像时的时序注视轨迹。如图1所示,在全景环境(第1行)和2D场景(第2行)中,两种方法呈现出不同特征:显著性预测生成连续概率分布图指示人眼关注的区域,而显著性目标检测输出二值化分割掩码来提取完整的显著物体轮廓。

VR360°全景环境中的显著性检测面临着与传统平面图像截然不同的挑战。该任务需融合用户动态视角行为和沉浸式体验特性,其空间连续性和视角变化特点要求专用建模方法,已成为优化VR内容制作和提升用户体验的关键问题。尽管该领域发展迅速,但仍存在全景畸变处理困难、视角依赖性建模不足等问题,限制了检测方法的准确性和鲁棒性。现有综述主要关注早期发展和2D环境下的显著性检测:Wang等人(2021)从学习范式等角度分析了2D环境下深度学习显著性检测方法;金海燕等人(2022)总结了显著性目标检测的研究现状和应用领域;李婉蓉等人(2022)从方法和数据集等角度分析了显著性检测发展现状;丁颖等人(2019)回顾了VR全景显著性检测的投影域适配方法;Buzzelli等人(2020)分析了全景环境显著性估计方法、数据集与评价指标。然而,这些综述缺乏对近年来全景显著性检测最新进展的全面总结,特别是深度学习方法和数据集的发展。

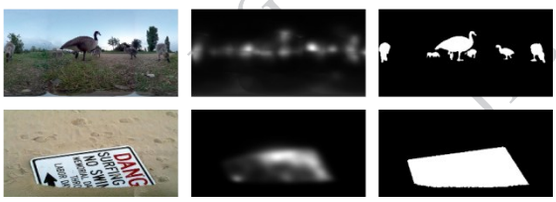

图1　2D场景与全景环境下的显著性预测和显著性目标检测

Fig. 1　Saliency Prediction and Salient Object Detection in 2D Scenes and Panoramic Environments

本综述系统梳理了2017-2025年发表于国内外权威期刊与会议的全景显著性检测方法的全景显著性检测方法,按传统手工特征改进、基于卷积神经网络/Transformer/LSTM架构、多投影域建模、视听多模

态融合四条主线回顾技术演进脉络;综合评估各类方法在应对球面畸变、用户视角变化等难题时的优势与局限性;归纳多个全景数据集的规模与适用场景,总结显著性检测在质量评估、压缩传输、虚拟摄影等应用中的部署效果;最后从多模态信息融合、轻量化无监督学习、构建统一的大规模基准数据集等方向展望未来发展趋势。

总之,本文的主要贡献如下:

1)本研究对传统方法在VR 360°全景图像/视频显著性检测中的应用进行了系统回顾。传统方法虽然具有计算复杂度相对较低、实现简单的特点,但在处理全景图像的球面几何畸变、空间不连续性方面存在显著局限性,难以准确捕获全景环境中的显著性特征。这一分析为后续深度学习方法的发展提供了重要的技术对比基础和改进方向。

2)本研究详细梳理了2017至2025年间深度学习在该领域的进展,并按卷积神经网络(Convolutional Neural Network,CNN)、Transformer、LSTM等架构,以及球面和立方体等多投影域建模进行分类。重点评估了这些方法在应对全景畸变、动态视角依赖和多模态融合方面的优劣,为设计更高效、鲁棒的端到端模型提供了重要参考。 3)本研究归纳了多个全景图像/视频数据集的规模及适用场景,并对比了不同算法在各类评价指标上的表现。通过分析算法优劣的影响因素,为未来模型优化和性能评估提供了指导。 4)本研究指出了当前研究中的核心挑战,包括全景畸变处理、多模态信息整合不足及实时性需求。同时,提出了未来研究方向,如扩展多模态融合、以及构建统一的大规模基准数据集等方向。这些方向旨在推动领域创新,促进虚拟现实技术的应用落地并提升用户体验。

本文的其余部分结构如图2所示。本文第1章首先介绍了全景环境下显著性检测的基本理论与背景;第2章深入阐述了全景环境中的传统显著性检测方法;第3章重点介绍了基于深度神经网络的全景显著性检测,涵盖结合传统方法、基于卷积神经网络、Transformer和LSTM等架构的模型;并系统梳理了从递归序列建模向全局时空联合表征演进的时序建模范式转变;第4章专门探讨了基于投影变换的全景显著性检测方法,分析了不同投影域下模型的特点与性能;第5章分析了数据集与评价指标,对全景图像和全景视频的显著性检测数据集进行了详细描述,并对比了各方法性能;第6章探讨了全景显著性检测在相关领域中的应用;第7章剖析了当前研究中的潜在难题并提出了未来研究趋势;第8章对全文内容进行了综合归纳与总结。

1 基本理论与背景概述

1.1 全景图像与全景视频概述

VR技术能够为用户提供沉浸式虚拟现实体验。全景图像/视频作为VR内容的重要组成部分,主要通过鱼眼镜头或全景拼接技术获取。鱼眼镜头采用广角光学设计,能捕捉180°以上的超广角画面,形成球面全景图像,但在球面投影过程中,建筑等物体沿径向产生拉伸变形,边缘呈现桶形畸变(如图3所示)。全景拼接技术则通过多相机协同拍摄,利用图像配准和融合技术将多视角图像拼接成360°全景画面。这种球面成像方式赋予了全景内容独特的几何特点。首先,球面投影导致全景图像呈现明显的球面畸变,边缘区域产生严重的拉伸失真。其次,全景画面的视野范围大幅拓展,水平视场角可达360°,垂直视场角约180°,远超人眼视域。这些几何特性对后续的图像/视频处理和显著性检测构成新的挑战。

与传统2D图像和视频不同,VR沉浸式体验下用户的观看行为发生了显著变化。如图4所示,全景内容通过球面投影映射到用户视角,实现了沉浸式观看体验。图4还展示了VR全景图像常用的投影形式,包括等矩形投影(Equirectangular Projection, ERP)和立方体投影(Cube Map Projection, CMP)格式,体现了用户交互中的动态观察模式。

在观看全景图像时,用户可自由转动头部观察全景环境中的任意位置,且头部运动与眼球运动高度耦合。与仅关注注视点位置的2D显著性检测不同,360°全景内容的显著性检测需结合用户头部运动轨迹和注视点分布特征。用户头部运动数据反映其在全景空间中的视觉注意力分布,为显著性检测提供重要线索;观看时的注视点位置也能反映视觉兴趣点,可作为显著性检测的重要依据。因此,如何挖掘VR用户行为数据中的注意力信息并将其有效集成到显著性检测模型中,是该领域研究的重要方向。

此外,在VR全景环境中,用户观看图像和视频

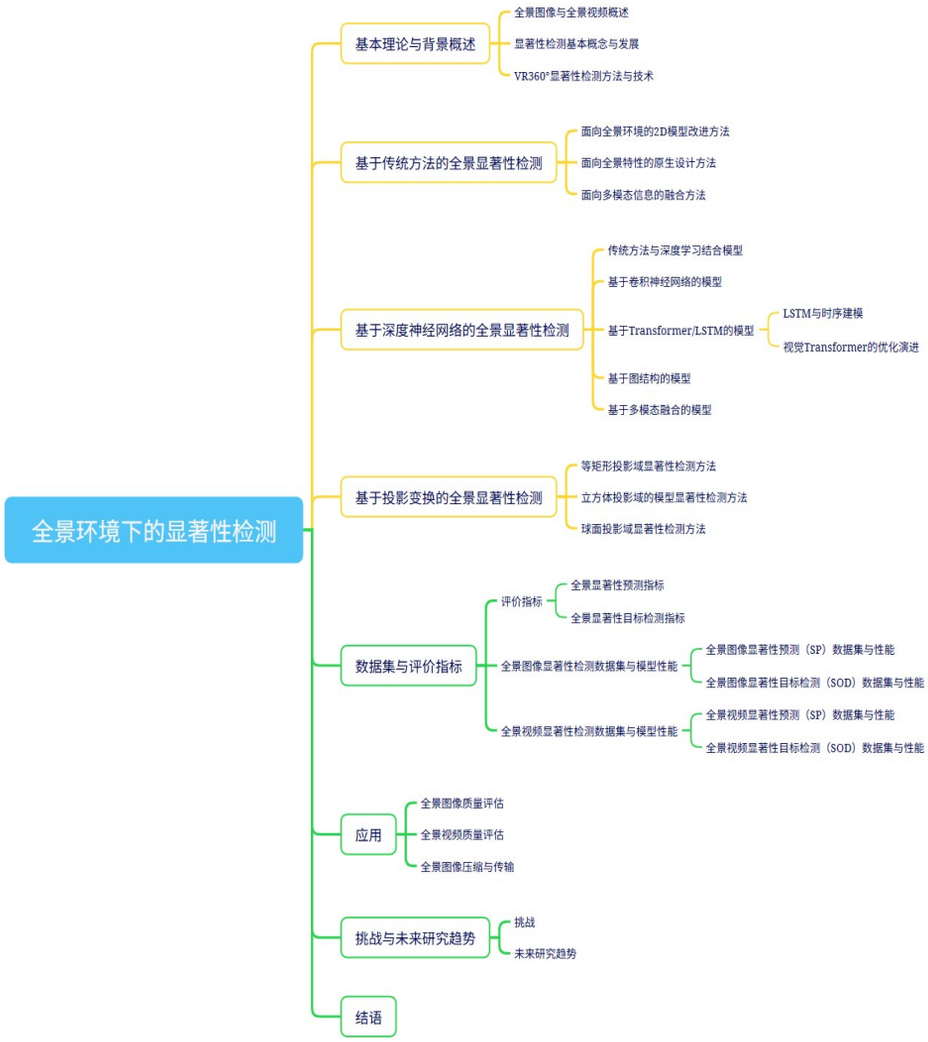

图2　本文架构

Fig. 2　The Structure of This Article

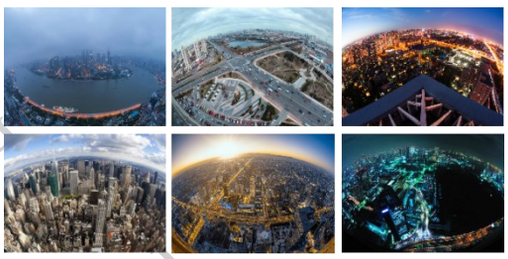

图3　鱼眼镜头下的城市

Fig. 3　The city viewed through a fisheye lens

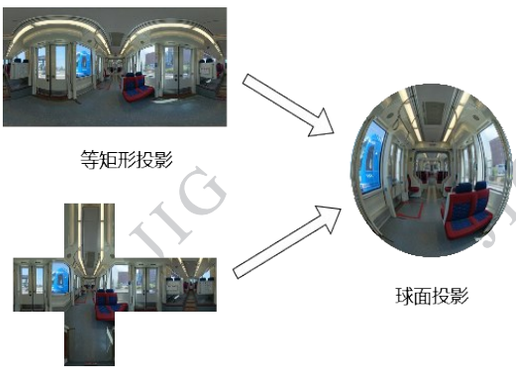

图4　全景环境下的常见投影方式

Fig. 4　Common Projection Methods in a Panoramic Environment

时，头部长时间倾斜会引发生理不适，使其更倾向于将注意力集中在水平中线附近的赤道区域，形成典型的"赤道偏倚"视觉偏好模式，反映了用户在交互过程中对内容分布的系统性偏移。

1.2　显著性检测基本概念与发展

正如前文所述，显著性检测（SD）是指从图像或

视频中自动识别和提取吸引人类视觉注意力区域的计算过程,其理论基础源于人类视觉系统的选择性注意力机制。Koch和Ullman(1987)首次提出视觉注意力的神经网络理论模型和显著性图概念,为该领域奠定了理论基础。Itti等人(2002)进一步拓展并实现了可操作的计算模型,通过提取低级视觉特征和中心-周边差异分析构建了完整的显著性检测框架。Harel等人(2006)提出的基于图的视觉显著性(Graph-based visual saliency,GBVS)算法,通过马尔可夫链在特征图上模拟激活扩散过程,显著提高了检测性能与生物学合理性。

随着研究深入,越来越多学者转向纯计算显著性检测模型的研究。如:Zhu等人(2014)提出基于鲁棒背景检测的显著性优化算法,创新性地引入边界连通性特征来刻画背景区域与图像边界的拓扑关系,在保持计算高效性的同时提高了显著目标检测的精度。

深度学习兴起后,显著性检测领域从基于手工设计特征的传统方法转向端到端的数据驱动模型,显著提升了检测性能。Liu等人(2024)提出的增强型视觉显著性变换器(Visual Saliency Transformer++,VST++)结合纯Transformer架构与选择-融合注意力机制,引入深度位置编码技术融合RGB与深度信息,为多模态显著性检测任务提供了高效且具有泛化能力的解决方案。

尽管深度学习模型在传统2D图像显著性检测中表现出色,但直接应用于VR 360°全景图像时仍面临诸多挑战:球面畸变导致的几何失真、视角依赖性引起的显著区域空间不连续性,以及用户自由视角选择削弱的中心偏好效应等,使得2D方法难以有效捕捉全景的全局上下文和动态显著性分布。以全景图像SOD任务为例,如表1所示,将传统2D图像SOD模型和专为全景图像设计的SOD模型在360-SOD数据集上进行对比,其中2D图像模型直接将等距柱状投影(ERP)图像作为输入。实验结果表明,全景模型在各项指标上均明显优于2D模型,这主要归因于全景环境的360°连续性、空间完整性和深度感知特性使得上下文信息在显著性判断中发挥更重要作用,同时用户观看方向的自主选择进一步强化了显著性分布对内容特征的依赖。

表1　在360-SOD数据集上2D与全景SOD方法的性能比较

Table 1　Performance Comparison of 2D and Panoramic SOD Methods on the 360-SOD Dataset

方法分类	年份	算法	FM ↑	SM ↑	Em ↑	MAE ↓
2D图像SOD模型	2022	PGNet(Xie等,2022)	0.673	0.780	0.823	0.026
		MSCNet(Lin等,2022)	0.656	0.735	0.834	0.046
	2023	BSCGNet(Feng等,2023)	0.675	0.784	0.831	0.024
全景图像SOD模型	2022	CSMANet(Zhang等,2022)	0.833	0.873	0.924	0.016
	2023	MIDP-Net(Dai等,2023)	0.7799	0.831	0.9101	0.022
	2024	ACoNet(陈晓雷等,2024)	0.7815	0.8493	0.9043	0.0181

注:所有方法均在相同的360-SOD测试集上进行评估。

1.3　VR360°显著性检测方法与技术

VR 360°显著性检测方法主要分为基于传统特征的方法和基于深度学习的方法。传统特征方法通过对经典显著性检测算法进行几何适应性修正和球面特征提取优化,以克服全景图像的投影失真和空间连续性问题;深度学习方法则依赖神经网络的自动特征学习能力,实现更高精度的端到端显著性检测。这些技术不断演进以适应全景环境的球面几何结构、视角畸变等特性。表2归纳了VR全景环境下显著性检测算法的分类,涵盖了代表性方法的作者、关键技术与适用场景及显著性类型。

2　基于传统方法的全景显著性检测

在全景虚拟现实(VR)领域,基于传统特征的显著性检测方法主要依赖手工设计的特征提取技术。这类算法通过整合色彩对比度、空间频率和纹理等低层视觉线索,采用特征融合与分割策略定位视觉场景中的显著区域。相比深度学习方法,传统显著性检测算法具有计算复杂度低、理论基础清晰且无

表2　VR全景环境下的显著性检测(SD)算法分类

Table 2　Classification of Salient Detection (SD) Algorithms in VR Panoramic Environments

方法		适用场景	关键技术	检测任务
分类	作者			
传统方法	Fang等人(2018)	ODI	超像素分割、360°边界测量	SP
	苏群等人(2018)	ODI	多角度分割、稠密稀疏重建	SOD
	Cokelek等人(2021)	ODV	空间音频融合、音频频率显著性	SP
深度学习方法	Lv等人(2020)	ODI	图显著性预测网络、球面冠插值	SP
	Wu等人(2022)	ODI	畸变容忍、特征增强样本自适应融合	SOD
	Guo等人(2024)	ODV	对比学习、声学分离	SOD

需大规模标注数据等优势，在特定应用场景中仍具实用价值。本节将系统探讨三类主要的传统全景显著性检测范式：面向全景环境的2D模型改进方法、面向全景特性的原生设计方法，以及面向多模态信息的融合方法。

2.1　面向全景环境的2D模型改进方法

传统2D显著性检测模型在计算机视觉领域的深厚积累为全景显著性检测提供了重要基础。面向全景环境的2D模型改进方法主要围绕几何畸变适配和投影策略优化两个核心挑战展开。在几何畸变适配方面，Fang等人(2018)通过超像素分割和特征对比度重构将传统2D显著性框架扩展至全景域。随后，苏群等人(2018)采用立方体投影策略，将全景图分割成多个角度区域并投影至立方体表面，以减轻经纬投影在两极产生的畸变。这两种方法虽然采用了不同的适配策略，但都旨在克服全景环境下的几何失真问题。在投影策略优化方面，Lebreton等人(2018)提出了更为系统化的解决方案。该工作包含三种计算框架(BMS360、GBVS360和ProSal)，其中ProSal框架通过引入150°视场激活策略和自适应赤道先验，使现有2D模型无需深度修改即可适用于全景环境，实现了2D模型向球面域的通用化迁移。这类方法的核心优势在于充分利用了成熟2D技术的先验知识，但在处理全景图像的空间连续性和尺度不一致性方面仍存在固有局限。

2.2　面向全景特性的原生设计方法

面向全景特性的原生设计方法摒弃了对传统2D模型的依赖，转而从360°图像的内在特性出发构建专门的显著性检测算法。这类方法主要围绕视觉感知机制模拟和几何结构适配两个技术路径展开。在视觉感知机制方面，基于稀疏表示的方法通过过完备颜色字典和加权视觉敏锐度函数来模拟人类颜色感知，有效捕捉底层图像结构(Ling等，2018)。在几何结构适配方面，基于区域生长的方法将空间密度模式与眼注视预测相结合，通过对象提案融合应对全景环境下的复杂背景挑战(Zhu等，2017)。相比于2D模型扩展方法，原生设计方法的核心优势在于能够直接处理全景图像的球面特性，但缺乏大规模数据集的支撑，准确性仍有局限。

2.3　面向多模态信息的融合方法

鉴于全景环境包含的海量信息超越了单一视觉特征的处理能力，多模态融合方法成为提升显著性检测精度的关键途径。在全景图像领域，该类方法的核心在于利用深度线索补偿全景投影造成的空间几何失真。例如Battisti等人(2019)通过单目深度估计与低层视觉特征的协同建模，并引入加权策略来校正全景图像固有的赤道偏倚问题，为静态全景场景的显著性检测提供了有效的多模态解决方案。

全景视频的多模态显著性检测则聚焦于视听信息的时空协同建模，形成了两种主要的技术路径：基于神经生物学的原型对象建模和基于信号处理的空间声学定位。前者以Ramenahalli等人(2020)的AVSM模型为代表，通过整合多维视听特征构建原型对象，利用中心-周边机制实现对动态场景的统一处理；后者如Cokelek等人(2021)的频率分析方法，通过Mel频率倒谱系数处理立体声音频通道，生成空间声学显著性图谱并与视觉模型后融合。这些方法共同推动了全景视频显著性检测从单一感官向多感官认知建模的范式转变。

3　基于深度神经网络的全景显著性

检测

传统全景显著性检测方法在处理投影失真和边界模糊等特有问题时仍存在精度局限，难以有效应对复杂动态场景。近年来，基于深度神经网络的显著性检测方法凭借其强大的特征学习和建模能力，显著提升了检测精度和鲁棒性，已成为当前研究热点。如图5所示，该领域经历了从几何校正与投影适配（2017-2019）、多投影融合与几何感知建模（2019-2022）、轻量化与多模态融合（2022-2024）到全局统一与端到端优化（2024-至今）四个发展阶段，逐步实现了从显式畸变补偿到跨投影协同建模、从静态单帧分析到动态时空联合表征、从单一视觉模态到多感知融合的技术跨越。本章节将从网络结构及多模态融合等维度对深度学习方法进行分类分析，探讨各类模型的技术特点与应用优势。

图5　当前代表性方法发展脉络和时间表

Fig. 5　The development history and timeline of current representative methods

3.1　传统方法与深度学习结合模型

早期全景图像显著性检测研究主要采用"特征互补和投影校正"的融合范式，通过传统视觉特征与深度学习的协同建模解决投影失真问题。该类方法可分为两个技术路径：统计偏差校正和多特征融合。统计偏差校正方面，以Sitzmann等人（2018）基于头部运动统计的VR显著性预测框架为代表，通过赤道偏倚优化处理等矩形投影的固有偏差；在多特征融合方面，Mazumdar等人（2019）的COSE模型结合YOLO目标检测与全局色彩特征实现内容导向的显著性预测。这类方法的核心在于利用传统先验知识补偿深度模型对全景几何特性的认知不足，通过领域知识注入提升复杂场景下的预测精度。

与图像方法相比，全景视频显著性检测更强调时序连续性建模与运动预测能力，但在深度学习早期阶段，研究重心仍延续了几何校正与空间特征提取的技术路径，尚未充分释放时序信息的潜力。代表性工作包括基于单帧深度网络的扩展方法和多投影融合策略两类。如图6所示，Nguyen等人（2018）提出的PanoSalNet模型采用VGGNet迁移学习结合中心偏差修正机制，通过多阶段处理流程实现全景显著性预测。模型接收全景视频的单帧图像作为输入，首先通过卷积操作逐层提取多尺度特征并生成特征图；其次利用最大池化在局部区域内进行归一化以增强特征对比度；再通过反卷积层将低分辨率特征图上采样为高分辨率显著性图；最后应用先验滤波器，通过中心偏差修正机制降低等矩投影图像四角的显著性值，消除投影失真的影响。这一设计体现了空间特征提取与几何修正的有机结合。Tliba等人（2021）则采用双投影策略，同时处理ERP和CMP格式并通过加权概率分布融合预测结果。相比全景图像方法，视频显著性检测更强调时序连续性和运动预测能力，通过引入领域先验增强深度模型对全景环境的时空建模能力，为后续端到端深度学习方法的发展奠定了理论基础。

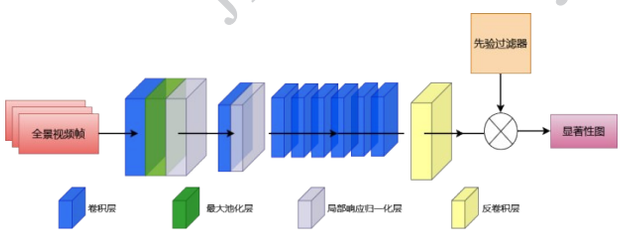

图6　PanoSalNet模型（Nguyen等，2018）

Fig. 6　PanoSalNet Model (Nguyen et al., 2018)

3.2　基于卷积神经网络的模型

CNN作为特征提取与融合的核心工具，已在传统2D显著性检测领域积累了丰富的应用经验。这些方法主要通过多尺度特征聚合、注意力机制和编码-解码架构实现对显著区域的精准预测。例如，HVPNet（Xu等，2025a）在多阶段框架中动态整合跨模态特征。这类工作验证了CNN在处理图像与视频显著性任务中的有效性，为其在全景环境下的应用提供了重要参考。

在全景视频领域，研究者们重点探索了时空信息融合策略。代表性工作如STDMMF-Net(Chen等，2023b)采用双流架构，结合CNN从视频帧提取的空间特征与从光流中提取的运动特征，再通过ILA和BMA模块融合多模态信息。这类方法的核心在于利用CNN的空间学习优势与光流的运动捕捉能力，实现全景视频中的鲁棒显著性目标检测。与视频任务相比，全景图像显著性检测的关键挑战在于处理投影几何畸变和多视角信息整合。主流解决方案可归纳为三类技术路径：几何感知卷积、多视角特征融合和自适应特征校正。几何感知方法通过球面卷积(Sui等，2023)等技术替代传统卷积操作，直接在网络层面补偿投影畸变。其中，如图7所示，Huang等人(2023)提出的轻量化LDNet模型采用ResNet-18为骨干网络，构建了多级扭曲感知处理流程：输入的全景图像首先经过常规卷积层初步提取特征，随后进入LD-ResNet-18骨干网络，将后两层普通卷积替换为扭曲感知深度可分离卷积(DDSConv)实现抗畸变特征提取；然后依次经过扭曲感知通道增强(DCE)模块优化通道相关性、密集调制结构(含DDSConv、上采样与跳跃连接)实现多尺度融合、以及扭曲感知自相关(DSC)模块捕获全局-局部依赖关系；此后再借助DCE模块以密集连接方式逐层整合特征，最终经1×1卷积与上采样生成显著性图；在视角变化失真修正与模型轻量化之间取得了有效平衡。多视角融合策略则采用多视口采样技术，通过整合不同视场的局部特征实现全局显著性建模(Chao等，2020b)；自适应校正方法包括利用可学习赤道偏置层适应注视点分布的方法(Yamanaka等，2023)，以及采用多层特征交互模块和两阶段解码策略优化显著性预测的方法(Dai等，2023)，这些设计在保持2D CNN预训练优势的同时，有效适配了全景图像的空间分布特性。总体而言，这些技术进展体现了CNN特征提取能力在全景环境下的有效延伸，通过针对性的几何修正和特征融合策略有效应对了全景成像的几何约束问题。

3.3 基于Transformer/LSTM的模型

3.3.1 LSTM与时序建模

全景显著性检测的时序建模经历了从递归序列建模向全局时空联合表征的范式转变。早期工作通常将2D显著性图序列输入LSTM类网络处理时序信息，但这类方法因递归或顺序计算架构的固有限制，难以高效学习长时依赖关系。为突破这一瓶颈，后续研究沿多条技术路线进行了探索。在优化策略层面，Xu等人(2018a)将LSTM与深度强化学习相结合，通过模拟头部运动的连续性约束来优化奖励函数；Qiao等人(2020)提出MT-DNN模型，采用ConvLSTM处理CNN提取的空间特征，借助多任务联合优化增强时序学习能力。在特征表示层面，Dahou等人(2021)提出ATsal模型，在全景视频显著性预测中首次应用注意力机制，采用双流结构分别处理全局和局部信息；Zhang等人(2023b)的CAV-Net采用条件变分自编码器框架，通过多模态特征融合有效应对主观随机性。这些进展虽在一定程度上改善了时序学习性能，但本质上仍依赖递归机制，难以实现真正的全局时空联合表征。Wan等人(2024)针对特征对齐问题，提出了融合轴向注意力的球面卷积LSTM网络，通过球面卷积处理投影失真、轴向注意力模块捕捉全局时空依赖和双流架构融合双向时序信息，体现了通过多技术融合来增强递归模型表征能力的思路。然而，上述各类方法均未能建立跨帧的全局时空关联，缺乏对长视频序列进行高效统一建模的能力，这一共性瓶颈成为了推动范式转变的关键因素。

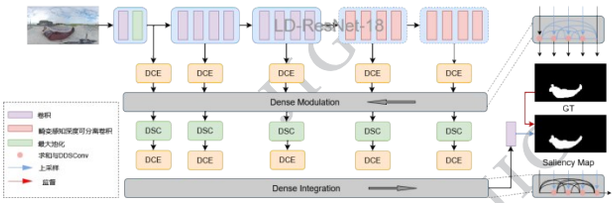

图7 LDNet模型(Huang等，2023)

Fig. 7 LDNet Model (Huang et al. , 2023)

3.3.2 视觉Transformer的优化演进

为克服这些技术瓶颈，Transformer架构逐渐成为主流选择：Yun等人(2022)的PAVER模型首次将视觉变换器引入全景视频编码，通过变形卷积分割策略和自注意力机制实现远距离依赖建模；Cokelek等人(2023)提出的SalViT360模型是首个将切线图像应用于全方位显著性预测的方案，该模型通过Transformer的球面几何感知时空自注意力机制聚合全局和时序信息，采用分阶段注意力策略有效降低计算复杂度。这一技术演进清晰地体现了全景视频显著性检测从局部时序建模向全局时空建模的根本性转变。

全景图像显著性检测面临的核心技术挑战集中在几何失真处理和多尺度目标检测两个维度，Transformer架构在全局特征提取和自适应注意力机制方面的优势为解决这些问题提供了有效途径。在失真感知技术方面，Zhao等人(2023)的DATFormer整合了失真映射模块和基于Transformer的失真自适应注意力机制，通过多头自注意力动态调整多尺度特征权重以减轻赤道区域的鱼眼效应；陈晓雷等人(2023)的URMNet采用鲁棒视觉变换模块提取多尺度特征，结合多头注意力选择性融合空间和通道信息。在多视角建模方面，如图8所示，Wu等人(2022)的Sample Adaptive View Transformer(SAVT)模型展现了独特的技术创新：它以ResNet-50为编码器，通过Res1至Res5五个阶段逐级提取多尺度特征，并利用通道适配模块(Ad)调整特征维度；提取的特征经特征融合模块(FFM)整合后进入核心SAVT模块，其包含三个并行的View Transformer(VT)分支——水平变换分支、垂直变换分支与缩放变换分支，借助莫比乌斯变换分别学习不同水平视角、垂直视角及多尺度下的特征表示，有效缓解了全景图像中的边缘不连续和尺度变化问题；随后，Sample Adaptive Fusion(SAF)模块根据样本特征动态调整三个VT分支的输出权重并完成信息融合；整合多视图信息生成精确的显著性映射。与传统图像切割策略不同，SAVT保持了完整的全景视图完整性，采用多分支协同策略处理尺度变化，显著提升了对可变规模目标的检测能力。此外，Zhu等人(2025a)的IMRE模型通过模拟神经心理学中的回顾性记忆过程，利用Vision Transformer(ViT)作为源编码器从模糊上下文中提炼语义特征，借鉴扩散模型的去噪机制实现迭代式显著性推理，体现了该领域从纯技术驱动向生物启发式建模的发展趋势。

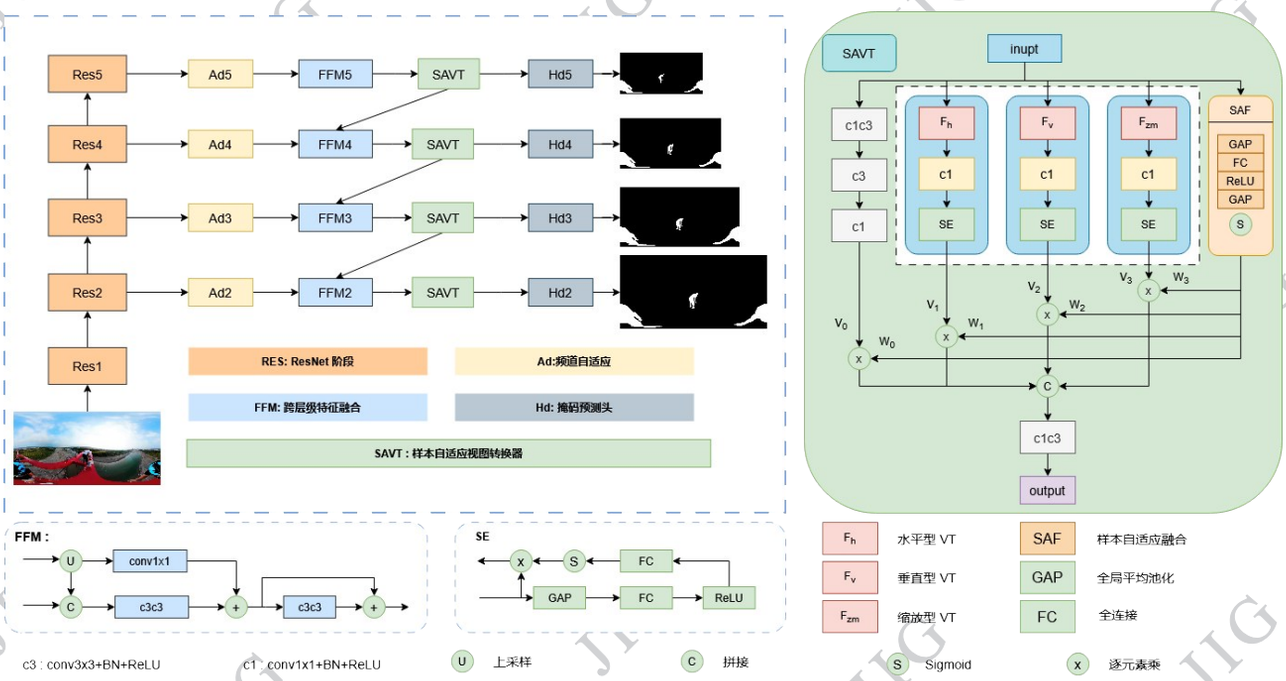

图8　SAVT显著性检测模型(Wu等，2022)

Fig. 8　SAVT Saliency Detection Model (Wu et al., 2022)

3.4　基于图结构的模型

图结构方法通过显式建模节点间复杂的空间关系，有效处理全景图像中的球面几何变形和不规则空间连接，相比传统网格结构具有天然优势。此类方法将全景内容分解为离散节点，利用图卷积操作学习节点间相互作用，实现对显著区域的精准定位。

全景图像显著性检测中的图结构方法主要围绕多尺度空间建模和预测机制创新两个核心问题展开。在多尺度建模方面，Zhang等人(2024)采用多尺度超像素分割与球面斐波那契采样构建层次化图结构，通过图卷积提取多尺度显著性特征；在预测机制创新方面，Zhu等人(2025b)的ScanDTM模型首次将扩散模型引入扫视路径预测任务，采用双图卷积网络融合确定性图结构与概率性扩散过程，有效提

升了模型的泛化能力和预测精度。

相比静态图像，全景视频显著性检测面临着更为复杂的挑战，需要在图结构框架下同时处理时序动态变化与球面几何约束的耦合问题。Yang等人(2024b)提出的360Spred模型采用3D可分离图卷积网络处理时空特征，通过图信号构造模块将ERP视频转换为球面图信号，利用时空分离卷积与注意力机制融合实现动态显著目标的精确跟踪。如上文所述，图结构方法通过灵活的节点关系建模和多样化的特征融合策略，为全景显著性检测从静态图像向动态视频的技术演进提供了有效的解决方案。

3.5 基于多模态融合的模型

多模态信息融合已成为提升复杂场景显著性检测性能的重要技术路径。在传统2D显著性检测领域，通过整合RGB、深度、听觉等多源信息，模型显著提升了对复杂环境的适应能力。近期代表性工作包括：Xu等人(2025b)提出的SOMA-Net模型通过双阶段稀疏语义增强和正交互注意力融合增强跨模态特征互补性；Yi等人(2025)的GL-DMNet模型利用位置-通道互融机制和级联变换器解码器实现全局-局部多尺度特征表达。这些多模态方法在显著性检测中的成功实践，为全景显著性检测领域的多模态技术发展提供了重要的理论基础。然而，当多模态融合扩展至全景视频领域时，时序动态与跨模态交互的耦合带来了新的挑战，推动了面向全景视频的多模态时序建模研究。

全景视频多模态显著性检测的时序建模形成了独立于视觉单模态的技术演进脉络。与之前所述的"由递归到全局"的范式不同，多模态时序融合的核心挑战在于跨模态时序对齐与异构信息动态权重分配。早期方法主要采用后期融合策略：Li等人(2023b)利用Transformer全局建模优势设计标签引导蒸馏机制，通过音频-视觉特征的后端拼接增强对应关系，但未能实现真正的时序交互。中期方法转向特征级融合与几何感知：如图9所示，Guo等人(2024)针对实例级全景音频-视觉显著性检测和排序任务，设计了以ResNet-50为视觉编码器的统一框架，构建多阶段融合处理流程：输入视频帧与音频轨道经ResNet提取多级视觉特征，同时由音频编码器提取音频特征；视觉特征进入扭曲感知像素解码器(DPD)，通过双投影点采样(ERP投影与CMP投影)和特征聚合动态调整采样位置，生成畸变感知视觉特征图；随后，音频-视觉空间激活模块(AV-SAM)将音频特征投影至视觉特征空间进行元素求和，实现初步跨模态融合；与此同时，音频ProtoNet将音频特征转换为多个隐含子空间的音频原型；二者共同送入音频-视觉实例对齐模块(AV-IAM)生成实例级融合特征图；最终经时空对象解码器，利用掩码自注意力与交叉注意力机制，通过分类头判定显著目标、排序头输出显著性顺序、掩码头生成分割掩码，首次实现了实例级处理与音频-视觉对齐的对比学习机制。最新方法实现了端到端查询基时序交互：Wan等人(2025)的CASP模型创新性地设计可学习音频查询，通过动态捕捉跨模态依赖并施加自适应一致性损失，显式增强时序连贯性，其查询机制天然适配球面视频的时空特性；Zhu等人(2025c)的OmniAVS模型基于ImageBind多模态基础模型，通过分层音频-视觉融合模块分别处理语义特征(W通道)与方向特征(X、Y、Z通道)，实现了音频空间线索与视觉时序特征的深度耦合；Cokelek等人(2025)的SalViT360-AV模型则在球面Transformer架构中集成时序适配器，将多模态融合与球面几何适应统一于端到端框架。这一演进体现了多模态时序建模从特征级联到语义交互再到几何-时序联合优化的深化过程。

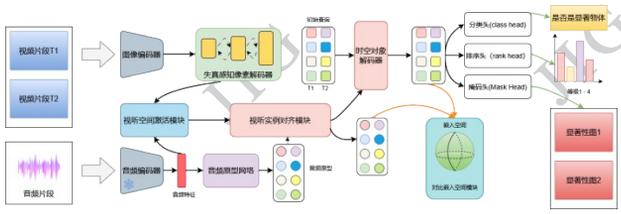

图9 全景音频-视觉显著性检测和排序框架(Guo等，2024)

Fig. 9 Panoramic Audio-Visual Salience Detection and Ranking Framework (Guo et al., 2024)

相比全景视频的快速发展，全景图像多模态显著性检测仍处于起步阶段。Chen等人(2023a)的MFFPANet模型通过RGB与物体级语义融合，采用动态互补特征融合和渐进式多尺度聚合处理静态场景，但其融合策略相对简单。多模态方法通过引入音频空间线索和语义信息，有效突破了单一视觉模态在复杂全景场景中的感知局限，代表了显著性检测从单模态向多感知融合的重要发展方向。然而，该领域仍面临模态对齐精度不足、计算复杂度高等关键挑战。未来研究需要在保证融合效果的前提下优化模型效率，探索全景环境下的深度、热红外等更

多模态信息的集成策略，并建立更完善的多模态全景数据集以支撑模型训练与评估。

4 基于投影变换的全景显著性检测

全景内容在获取和表示过程中存在复杂的几何失真和边界模糊问题，对传统显著性检测算带来了诸多挑战。为应对这些特殊视觉特性，基于投影变换的全景显著性检测方法成为重要研究方向。该类方法通过将全景图像和视频映射至等矩形投影、立方体投影、球面投影等不同几何投影域，利用各投影域的几何特性优化模型性能，有效缓解球面几何失真问题。不同投影变换方式在处理空间分布和时序变化方面各具优势，下文将详细分析三种主要投影变换模型的应用特点与技术优势。

4.1 等矩形投影域显著性检测方法

等矩形投影（ERP）作为将球面全景内容映射到矩形平面的标准格式，在全景显著性检测中占据主导地位。ERP格式具有良好的数据兼容性，VR全景图像与视频通常以此格式存储，为神经网络处理提供了统一的输入表示。然而，ERP投影在获得矩形便利性的同时，不可避免地引入了几何畸变和边界不连续性等固有问题：极地区域的过度拉伸、图像左右边界的连接性丢失等，这些特性对传统显著性检测算法构成根本性挑战。因此，如何有效补偿ERP几何失真并保持全景内容的空间连续性，成为该投影域显著性检测的核心技术问题。

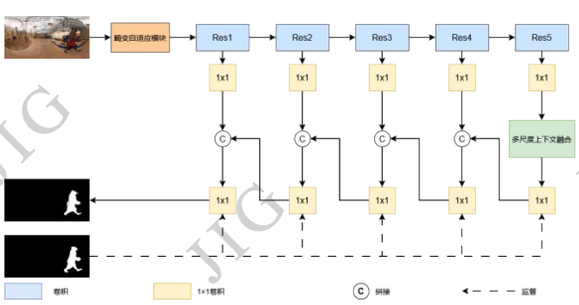

图10 DDS模型架构（Li等，2019）

Fig. 10 Architecture of the DDS model (Li et al., 2019)

针对ERP投影固有缺陷，全景图像显著性检测方法呈现出从单一补偿到系统性适应的明显演进趋势。如图10所示，早期代表性工作DDS（Li等，2019）开创性地引入畸变自适应模块，通过深监督机制逐步校正ERP几何失真，确立了显式畸变补偿的基本范式：输入图像首先经过畸变自适应模块处理，被分割为多个局部图像块，并针对每个图像块应用自适应卷积核进行几何校正；校正后的特征输入ResNet-50骨干网络，通过多个残差模块逐级提取层次化特征；最高层特征进入多尺度上下文集成模块，同时将粗层次显著性特征与细层次特征连接融合，生成精细化的显著性图；网络训练过程中采用深度监督机制，对各侧输出层计算损失以优化参数；最终输出校正后的显著性预测图。Ma等人（2020）进一步提出多阶段处理框架，将ERP域的粗定位与无畸变补丁上的精细化检测相结合，有效分解了投影复杂性；而Xu等人（2021a）的SalGAIL另辟蹊径，利用ERP格式的统一表示优势，将人类头部轨迹数据与ERP图像无缝整合，通过生成对抗模仿学习捕捉前中心偏差（FCB）等ERP域特有的注视模式，开辟了基于用户行为的显著性建模路径。近期研究更加注重ERP域的内在几何结构建模：ACoNet（陈晓雷等，2024）通过相邻细节融合捕捉ERP边界连续性，CPNet（Wen等，2025）采用分割-拼接策略系统性处理投影不连续问题；而DPNet（陈晓雷等，2025）则创新性地引入经纬度空间权重，通过位置感知模块显式建模ERP的球面坐标特性。这些方法反映出ERP域显著性检测已从被动适应投影失真转向主动利用投影结构，通过深入理解ERP几何特性实现更精准的显著性建模。

整体而言，当前ERP域显著性检测研究呈现出显著的方法多样化和技术成熟化趋势。从技术路径来看，该领域已形成从几何补偿、结构建模到行为融合的多元化发展格局，不同方法在处理ERP固有缺陷方面各有优势。然而，现有方法仍主要聚焦于单帧图像处理，对于ERP域的时序建模和轻量化部署需求关注不足。未来发展应重点关注轻量化几何建模和多模态融合机制，实现从几何补偿向结构利用的根本转变。

4.2 立方体投影域的模型显著性检测方法

立方体投影通过将360°全景内容映射至六个立方体面，有效缓解了等矩形投影在极地区域的严重畸变问题，为全景显著性检测提供了更均匀的空间表示。相比ERP的全局连续性优势，CMP在局部几何保真度和边缘区域处理方面表现突出，但也引入了面间接缝伪影等新挑战。基于两种投影方式的互补特性，CMP-ERP双投影融合逐渐成为该领域的主

流技术路径，通过局部精细化和全局一致性的协同建模策略，有效平衡了几何失真补偿与计算效率的技术权衡。

在全景视频显著性检测中，立方体投影域的应用主要聚焦于时空特征的联合建模。Cheng等人(2018)的开创性工作建立了立方体投影与ConvLSTM结合的基本框架，如图11所示，通过立方体填充(Cube Padding)机制在卷积和池化过程中实现面间信息融合，结合时序一致性优化缓解标注数据稀缺的问题，作者首次将弱监督学习引入了全景视频显著性检测任务：模型采用双分支级联结构，静态分支将ERP格式视频帧转换为立方体映射，通过采用立方体填充策略的CNN提取显著性特征并与全连接层权重相乘，经后处理生成静态显著性图；时序分支接收静态分支的显著性特征图，输入ConvLSTM模块聚合时序信息得到特征图，再次经后处理转换回ERP格式，并通过优化当前帧与前一帧显著性图的时序一致性损失，生成最终的时序显著性预测结果。然而，这类方法仍主要依赖传统的LSTM架构，对于长程时序依赖和跨面时空关联的建模能力有限，且缺乏对视频内容动态变化的自适应机制。

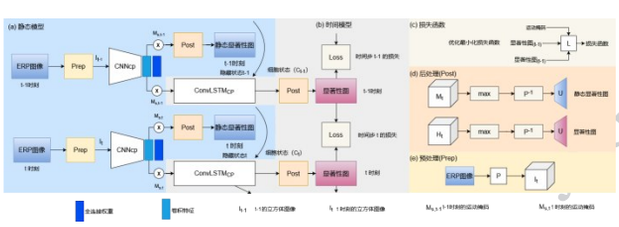

图11　弱监督显著性预测模型(Cheng等，2018)

Fig. 11　Weakly Supervised Saliency Prediction Model(Cheng et al.，2018)

全景图像显著性检测在立方体投影域呈现出从双分支融合到多投影协同优化的清晰演进路径。早期代表性工作以Dai等人(2020)的研究为典型，确立了CMP-ERP双分支并行处理的基本范式，通过全局-局部特征融合有效结合了两种投影的优势。中期发展以注意力机制为核心，其中Huang等人(2020)提出的FANet通过投影特征自适应模块融合多投影特征，而Zhang等人(2022)的CSMA-Net则引入通道-空间互注意力机制，显著提升了跨投影信息整合能力。近期研究呈现出两个主要发展方向：一是以Cong等人(2023)的MPFR-Net为代表的多投影集成策略，通过多视角几何建模进一步提升检测精度；二是以Zou等人(2023)的EPSNet为代表的自监督学习范式，通过代理任务辅助训练优化特征学习过程。最新的360Mamba框架由song等人(2025)提出，创新性地引入状态空间模型，通过几何感知的全局扫描策略和自适应融合机制，在保持计算效率的同时实现了对长程空间依赖的有效建模，代表了该领域向高效序列建模的技术转向。

4.3　球面投影域显著性检测方法

球面域能够在VR环境下的无失真表示中维持全景图像和全景视频的原始几何完整性，避免传统投影方式带来的变形问题。在全景视频显著性检测中主要呈现出从基础球面卷积到多模块协同优化的技术演进。以Zhang等人(2018)提出的球面卷积神经网络为代表，该类方法的核心创新在于球面卷积算子设计和时序建模机制两个方面。如图12所示，在球面卷积设计上，模型在球面冠上定义卷积核，通过沿球面旋转实现高效卷积运算，并根据全景图上各区域的空间位置对内核进行重新采样与旋转以适应投影畸变；球面冠状核在不同球面位置保持相同形状，投影至等矩形全景图后虽产生形变，但通过重新采样技术实现参数共享机制，模型还引入了球面U-Net网络，结合时间连贯性约束进行序列显著性检测，捕捉视频帧间的动态注意力模式，并采用球面均方误差损失函数优化训练过程。该框架作为首个专注于球面投影域的深度学习模型，为后续研究奠定了几何感知卷积的基础范式。在时序建模方面，球面U-Net结合时间连贯性实现序列显著性检测，后续研究围绕时序建模和反馈机制进行了深化。Li等人(2023a)的SPVP360模型将球面几何约束与用户交互反馈机制有机结合，如图13所示，构建了多分支特征处理流程：显著特征提取模块(ST-SPCNN)从单帧提取空间特征以避免投影畸变，同时从连续帧堆叠提取时间特征以建模动态时空关系，经多层球面卷积处理后通过卷积注意力模块(CBAM)动态优化特征表示；视场角预测模块(FoV)将稀疏用户反馈转换为聚合热图，并采用球面卷积门控循环单元(SP-ConvGRU)捕捉时间依赖性以整合历史信息；最终通过加权融合显著特征与FoV特征，实现全局-局部信息平衡与高效显著性预测。这一演进代表了球面域方法向时空联合优化和交互反馈方向的发展。相比于2D卷积扩展方法，球面投影域方法的优势在于参数共享机制和原生几何适配能力，但其计算复

杂度较高且对球面数据结构的依赖限制了其通用性。

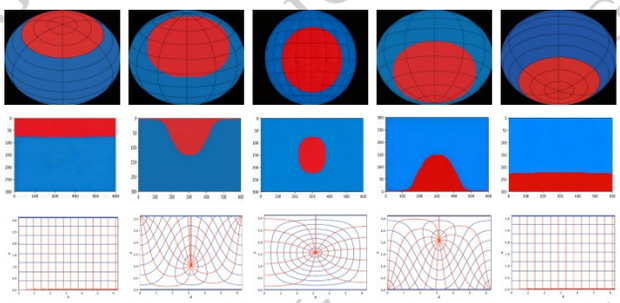

图 12　球面冠状核参数共享机制(Zhang等,2018)

Fig. 12　Shared Mechanism of Spherical Caudate Nucleus Parameters (Zhang et al., 2018)

球面投影域在全景图像显著性检测中形成了以几何畸变消除为核心、多样化技术路径并进的发展格局。主流技术路径可分为两类:一是以Monroy等人(2018)的SalNet360和Martin等人(2020)为代表的直接球面建模方法,通过将全景图像分解为无失真球面块或采用球面切平面表示,在网络架构层面根本性地消除投影失真,平均性能提升约20%;二是以Lv等人(2020)为代表的球面信号转换方法,通过测地线二十面体像素化将ERP图像转换为球形图信号处理。相比全景视频方法,球面域图像检测技术更注重保持静态几何一致性,体现了球面投影在不同任务场景下的适应性优化特征。

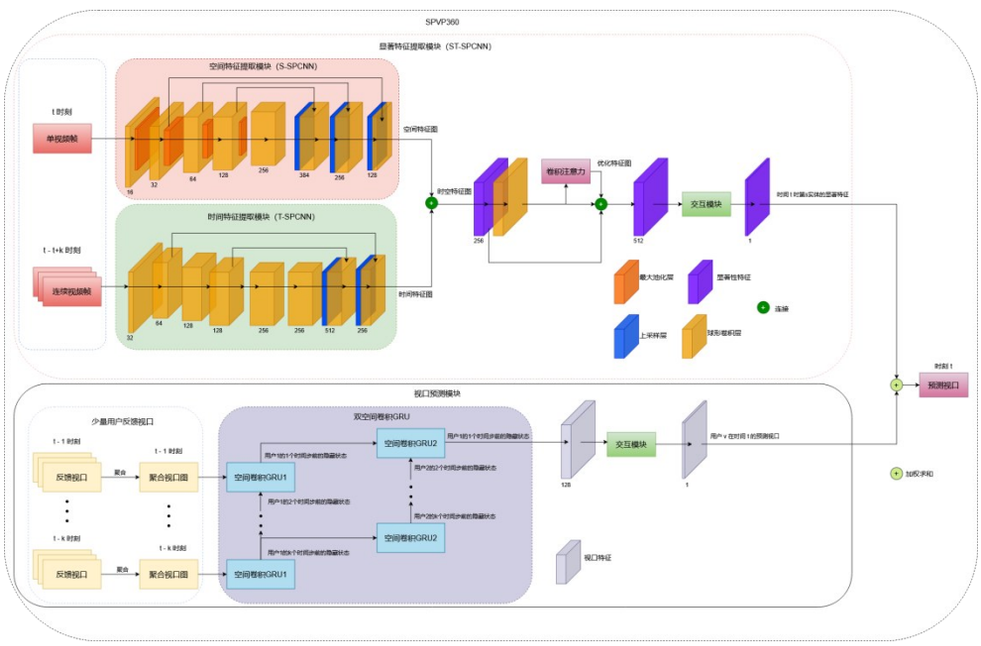

图 13　SPVP360模型(Li等,2023a)

Fig. 13　SPVP360 Model (Li et al., 2023a)

5　数据集与评价指标

5.1　评价指标

5.1.1　全景显著性预测指标

全景显著性预测任务需要专门的定量评估体系。然而,正如Djilali等人(2021)指出,传统2D评价指标在360°环境下存在适应性局限,所以也有学者对它们的计算方式提出了修改。本节将阐述全景显著性预测的客观评价指标。

(1)CC

皮尔逊相关系数(Pearson Correlation Coefficient,CC)(Bravais,1844)是用于衡量预测值$\boldsymbol{M}$(算法预测的显著图)和真实值$\boldsymbol{T}$(基准显著图)之间线性相关程度的指标。随着两者相关强度增加,CC值随之升高,最高可达到1,从而有效评估显著性检测算法的吻合程度。CC的计算公式如下:

$$CC=\frac{\mathrm{cov}(M,T)}{\sigma(M)\times\sigma(T)} \tag{1}$$

其中,$\boldsymbol{M}$表示预测的显著图,$\boldsymbol{T}$表示真实密度图,$\mathrm{cov}(\boldsymbol{M},\boldsymbol{T})$表示预测的显著图和真实密度图之间的协方差,$\sigma(\boldsymbol{M})$表示模型预测的显著图的标准差,$\sigma(\boldsymbol{T})$表示真实密度图的标准差。

（2） NSS

标准化扫视路径显著度（Normalized Scanpath Saliency，NSS）（Peters 等，2005）用于评估预测显著度图与注视点的吻合程度。计算归一化后的显著性图（均值为0、方差为1）在注视点上的平均值，NSS值可以表示估计值与真实值的对应关系：值为0表明随机对应，高值表示高度对应，低值表示反对应关系。NSS的计算公式如下：

$$NSS = \frac{1}{N}\sum_{i=1}^{N}\frac{S(x_i,y_i)-\mu_s}{\sigma_s} \tag{2}$$

其中，$\boldsymbol{S}(x_i, y_i)$是显著性图 $\boldsymbol{S}$ 在第 i 个真实注视点 (x_i, y_i)处的像素值；μ_s是显著性图 $\boldsymbol{S}$ 的均值，σ_s是显著性图 $\boldsymbol{S}$ 的标准差，N是真实注视点的总数。

（3） KLD

KLD （KL-Divergence）（Kullback 和 Leibler，1951）用于衡量两个概率分布之间的差异。在显著性检测中，预测的显著图和基准显著图会被视为概率密度函数，表示像素注意概率的连续分布，因此KLD值越小，表明两者越接近，算法性能越好。KLD的计算公式如下：

$$KLD = \frac{1}{2}\sum_{i=0}^{255}\left(P_i \log\frac{P_i}{Q_i} + Q_i \log\frac{P_i}{Q_i}\right) \tag{3}$$

其中，$\boldsymbol{P}_i$和$\boldsymbol{Q}_i$分别是归一化后的预测显著图与基准显著图在i处的值，i的取值范围是0到255，表示像素灰度值的范围。

正如Djilali等人（2021）所说，传统KLD基于分类分布假设，需要将输入归一化为概率分布，而在360°全景场景中因为观众可同时注视多个区域，而归一化过程会将能量集中在平均注视点上，导致模型预测的显著图与实际注视行为不匹配。因此作者设计了基于伯努利假设的KLD新定义，以适应360°全景场景：

$$KLD(q \mathbin{/\!/} p) = \sum_{i=1}^{N} q_i \log\frac{q_i}{p_i} \tag{4}$$

其中，q_i 表示真实分布中第 i 个像素被注视的概率，p_i 表示预测分布中第 i 个像素被注视的概率，N表示注视点分布图中的像素数量。

（4） SIM

相似度度量（SIMilarity measure，SIM）（Judd等，2012），将预测显著度图与标注图视为归一化概率分布进行比较，其值越大越能反映模型的预测效果。SIM的计算公式如下：

$$SIM(S,G) = \sum_i \min(S_i, G_i) \tag{5}$$

其中，$\boldsymbol{S}$表示预测的显著度图，$\boldsymbol{G}$表示真实标注的显著度图，$\boldsymbol{S}_i$表示预测显著度图在第 i 个像素位置的显著度值，$\boldsymbol{G}_i$表示真实显著度图在第 i 个像素位置的显著度值。

（5）ROC曲线和ROC曲线下面积（AUC）

在显著性检测领域，显著图可被看作像素点是否属于注视点的二值分类器。真阳性率（*TPR*）指正确辨识注视点的概率，假阳性率（*FPR*）则指错误辨识背景点的概率。通过调节阈值，绘制TPR与*FPR*的关系曲线，即*ROC*曲线，此曲线清晰展现模型对显著点的检测能力，而AUC（Judd等，2012）作为该曲线下的面积，其值越趋近于1，则表明算法的性能越优秀。*FPR*与*TPR*的计算方式如下：

$$TRP = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

其中，*TPR* 表示模型正确识别显著点的概率，*FPR* 表示模型错误识别显著点的概率，*TP* 表示正确识别显著点的像素数量，*FN* 表示未被识别为显著点的实际显著点像素数量，*FP* 表示错误识别为显著点的背景像素数量，*TN* 表示正确识别为背景的像素数量。为了在不同数据分布条件下规范化显著性评估，学者们还提出了多种AUC指标的改进变体，主要包括AUC-Judd（Judd 等，2012）和 AUC-Borji（Borji 等，2012）等。

5.1.2 全景显著性目标检测指标

（1）PR曲线

精确率-召回率（PR）曲线采用阈值扫描的方式获取显著性检测结果在精确率-召回率二维空间中的性能轨迹，从而提供比单一指标更为全面的算法评估视角，以召回率为横轴，准确率为纵轴，可以绘制准确率-召回率曲线（PR曲线），曲线位置越靠近右上方，说明算法性能越好。

（2） F-measure

F-measure（Achanta 等，2009）是精确率（PR）和召回率（RE）的加权调和平均值，用于评价预测显著性图的准确率和完整性。F-measure值越大，表示显著性检测算法的效果越好。

（3） MAE

平均绝对误差（MAE）（Perazzi 等，2012）可直接

基于预测结果进行计算，而无需引入任何阈值设定，MAE值越小，表示显著性检测算法的效果越好。

（4） S-measure

S-measure（Fan等，2017）是一种综合性的结构相似性度量指标，用于评估预测显著性图与真实标注在空间结构和强度分布上的一致性。该指标的取值范围为［0，1］，数值越高，表示算法性能越佳，显著性图与真值图越相似。

（5） E-measure

增强对齐度量（E-measure）（Fan等，2018）通过整合局部像素值与图像整体平均值于单一框架内，以获取图像级统计数据和像素级匹配细节，值越高越好。

5.2 全景图像显著性检测数据集与模型性能

随着民用360°摄像机的普及，360°全景图像和视频的获取变得日益便利，表3概述了用于全景图像显著性检测的若干数据集，全景图像显著性预测（SP）数据集通常记录观察者的头部和眼部运动轨迹，图像来源于真实拍摄，分辨率从4096×2048到24，028×12，014不等；相比之下，显著性目标检测（SOD）数据集（如360-SSOD）包含更多场景，较少记录运动轨迹，分辨率范围从1024×1024至8K不等。其中Salient360！挑战赛数据集应用最为广泛，本节将介绍各代表性的全景图像显著性检测数据集的特点。

表3 全景图像显著性检测（SD）数据集

Table 3 Panoramic Image Saliency Detection（SD）Dataset

数据集	场景数量（个）	记录类型	数据类型	分辨率（像素）	检测任务
SaliencyVR	22	头部、眼部运动	计算机生成	/	SP
Salient!360	98	头部、眼部运动	真实拍摄	5376×2688 ~ 18332×9166	SP
AOI	600	头部、眼部运动	真实拍摄	4,000 × 2,000 ~ 24,028 × 12,014	SP
HTRO	1080	头部运动	真实拍摄	8,000 × 4,000	SP
360-SOD	500	无	真实拍摄	1024 × 1024	SOD
360-SSOD	1,105	无	真实拍摄	1024 × 1024	SOD
F-360iSOD	107	眼部运动	真实拍摄	2048 × 1024	SOD
ODI-SOD	6,263	无	真实拍摄	2K ~ 8K	SOD

5.2.1 全景图像显著性预测（SP）数据集与性能

Saliency VR数据集（Sitzmann等，2018）记录了参与者在观看22张立体全景图像时的头部和眼部运动数据。实验涵盖多种观看条件，包括不同的用户姿势（站立/坐姿）和显示设备（头戴式/桌面监视器）。所有实验均采用眼动追踪技术采集数据。

表4总结了不同显著性预测模型在Saliency VR数据集上的性能表现。基于KLD等评估指标，Zhu等人（2025a）的模型表现最佳，该模型通过模拟人类回顾记忆机制，实现对先前信息的高效回想和推理，显著提升了预测准确性。Chao等人（2020b）位列第二，其Multi-FoV策略和自适应加权损失能够精确捕获360°图像的多视角细节，有效匹配预测与真实分布。Xu等人（2021a）排名第三，通过GAIL机制在CC指标上展现良好鲁棒性。此外，Sitzmann等人（2018）的框架整合了头部运动统计和投影优化，为VR环境显著性预测提供了基础方案。Sui等人（2023）将深度马尔可夫模型与球面卷积神经网络相结合，为后续研究奠定了重要基础。综合分析表明，结合类脑认知机制模拟和多视角信息融合策略是提升全景图像显著性预测模型性能的重要发展方向。

Salient360！数据集（Rai等，2017）是为ICME 2017同名竞赛开发的。该数据集经过持续迭代，最新版本包含用于评估预测扫视路径及显著图的工具包，涵盖头部和眼部运动数据。该数据集收录了98个VR全景图像，融入了环境配置和前景元素数量等多种因素，在数据多样性和规模上显著优于早期数据集。

表5汇总了不同显著性预测模型在Salient360！数据集上的实验结果，模型性能在过去几年间持续提升。综合性能表现优异的模型包括Zhu等人（2025a）、Martin等人（2020）、Zhang等人（2024）、陈

表4　不同显著性预测(SP)模型在Saliency VR数据集上的性能对比

Table 4　Performance Comparison of Different Saliency Prediction (SP) Models on the Saliency VR Dataset

年份	算法	KLD ↓	CC ↑	NSS ↑	AUC ↑	模型特点
2018	Sitzmann等人(2018)	/	0.49	/	/	传统方法+深度学习
2020	Chen等人(2020)	/	0.51	/	/	ERP+CMP投影
	Chao等人(2020b)	0.433	0.507	1.551	**0.83**	ERP投影
2021	Xu等人(2021a)	0.454	**0.603**	1.042	0.772	ERP投影
2023	Sui等人(2023)	0.609	0.521	0.958	0.767	ERP投影
2025	Zhu等人(2025a)	**0.424**	0.569	**1.893**	0.827	Transformer架构

注：加粗字体表示每类指标最优值，下划线表示次优值。

表5　不同显著性预测(SP)模型在Salient360！数据集上的性能对比

Table 5　Performance Comparison of Different Saliency Prediction (SP) Models on the Salient360！ Dataset

年份	算法	KLD ↓	CC ↑	NSS ↑	AUC ↑	模型特点
2020	Lv等人(2020)	0.428	0.589	0.945	0.736	球面投影
	Chen等人(2020)	0.402	0.661	0.975	0.746	ERP+CMP投影
	Martin等人(2020)	0.4411	0.7212	**3.0309**	**0.9624**	球面投影
	Chao等人(2020b)	0.363	0.662	0.978	0.747	ERP投影
	Dai等人(2020)	0.317	0.686	0.983	0.748	ERP+CMP投影
2021	Xu等人(2021a)	0.366	0.757	0.893	0.708	ERP投影
	Yang等人(2021)	/	0.824	1.753	0.845	LSTM
2023	Zhang等人(2023a)	/	0.913	2.285	0.871	ERP+CMP投影
	Sui等人(2023)	0.384	0.66	1.042	0.777	ERP投影
	Yamanaka等人(2023)	0.2464	0.7891	1.1751	0.7623	ERP投影
	Zou等人(2023)	**0.1125**	0.7141	0.8642	0.7607	ERP+CMP投影
	陈等人(2023)	0.5834	0.6683	2.9874	0.9336	Transformer架构
2024	Zhang等人(2024)	/	**0.9206**	2.1782	0.8751	球面投影和图结构
2025	Zhu等人(2025a)	0.342	0.903	1.938	0.925	Transformer架构

注：加粗字体表示每类指标最优值，下划线表示次优值。

晓雷等人(2023)与Zhang等人(2023a)。其中，Zhu等人(2025a)的模型性能表现仍然突出；Martin等人(2020)将卷积核参数化到球面切平面，在球面域进行360°感知卷积，有效缓解了投影失真；Zhang等人(2024)采用多尺度图结构显式编码球面拓扑关系，实现显著性特征的精准聚合；陈晓雷等人(2023)利用Transformer的全局建模能力和多注意力融合机制，捕获全景图像的多尺度长程依赖；Zhang等人(2023a)通过立方体投影与自适应赤道偏差感知模块联合建模全景几何和视觉先验，成功对齐人眼注视分布。投影策略角度看，ERP+CMP双投影方案在NSS和AUC等指标上超越单一ERP投影，体现了融合多视角几何信息的优势；球面投影从早期性能落后到2023-2024年逼近或达到SOTA水平，反映了球面卷积与等变网络等技术的显著改进。基于Transformer架构的模型也在多项指标上展现出优异性能，表明自注意力机制在处理全景图像的全局依赖关系方面具有巨大潜力。这些趋势表明，显著性预测在全景场景中正沿着深层网络架构与适配投影表示相结合的路径发展，为后续研究奠定了坚实基础。

AOI数据集(Xu等，2021)包含30名参与者对600张全景图像的头部和眼动数据，其大规模特性

为深度学习显著性预测模型训练提供了充足样本。该研究通过系统分析揭示了人类观看全景图像时的头部运动模式，包括前中心偏好（FCB）和观测者间幅度相似性。

HTRO数据集（Yang等，2021）记录了20名受试者在头戴式显示器中对1，080张8K等距柱状全景图像进行自由观看时产生的21，600条完整头部轨迹。该数据集在规模和内容多样性方面显著优于同类数据。作者利用凸包、方向分布量化等统计工具对数据进行系统分析，得出以下发现：全景静态图像诱发的可视区域比例高于全景视频，且头部移动以左右水平方向为主，并显示出时间连续性。这些结果为全景显著性建模和轨迹预测研究提供了坚实的行为基准。

5.2.2　全景图像显著性目标检测（SOD）数据集与性能

360-SOD数据集（Li等，2019）是首个360°全景图像显著性目标检测数据集，包含500张高分辨率等角矩形全景图像，涵盖复杂室内外场景，采用人工标注实现像素级显著性标识。该数据集通过系统分析揭示了人类观看全景图像的关键特征：投射失真、大尺度场景和小显著目标。

表6　不同显著性目标检测（SOD）模型在360-SOD数据集上的性能对比

Table 6　Performance Comparison of Different Salient Object Detection (SOD) Models on the 360-SOD Dataset

年份	算法	FM ↑	SM ↑	Em ↑	MAE ↓	模型特点
2022	CSMANet（Zhang等，2022）	**0.833**	0.873	0.924	0.016	ERP+CMP投影
2023	MIDP-Net（Dai等，2023）	0.7799	0.831	0.9101	0.022	CNN
	MPFRNet（Cong等，2023）	0.682	0.788	0.815	0.019	ERP+CMP投影
	DTAFormer（Zhao等，2023）	0.7742	0.8493	0.9071	0.0174	Transformer架构
	LDNet（Huang等，2023）	0.617	0.768	0.858	0.029	CNN
	MFFPANet（Chen等，2023a）	0.682	0.788	0.815	0.019	多模态
2024	ACoNet（陈晓雷等，2024）	0.7815	0.8493	0.9043	0.018	ERP投影
2025	CPNet（Wen等，2025）	0.8	0.862	0.925	0.018	ERP投影
	DPNet（陈晓雷等，2025）	0.7884	0.8502	0.9103	0.019	ERP投影
	360Mamba（song等，2025）	0.829	**0.886**	**0.929**	**0.015**	ERP+CMP投影

注：加粗字体表示每类指标最优值，下划线表示次优值。

表6汇总了不同显著性目标检测模型在360-SOD数据集上的实验结果。综合表现最优的模型包括CSMANet、CPNet与360Mamba。CSMANet采用ER与立方体双分支并行输入，借助通道-空间互相关注模块对齐两种投影的互补线索，兼顾全局语义和局部精细几何；CPNet采用分割-拼接策略保持物体连续性，用双向感知与边缘增强模块提炼多尺度上下文，获得锐利边界和一致显著图；360Mamba采用Mamba状态空间与ERP/CMP双投影融合，借助全局引导Mamba块对齐多尺度球面特征，兼顾长程上下文建模与几何畸变补偿。值得注意的是，LDNet虽整体表现中等，但其设计侧重于全景图像显著性目标检测的轻量化。从表6可见，显著性目标检测模型性能随年代推移稳步提升，但与显著性预测不同，采用ERP+CMP双投影与单一ERP投影的模型表现差距较小，主要因为目标级检测更注重语义判别，高层语义特征在深层网络中被自然吸收。Transformer和Mamba架构虽出现较少，但显示出显著潜力。未来研究应重点探索Transformer和Mamba架构在该领域的应用，以进一步提升全景显著性目标检测的性能。

相比360-SOD数据集，360-SSOD涵盖1，105张高分辨率全景图像，显著提升了数据规模。其标注语义分布更均衡，涉及10大类（如人物、车辆等），保留了360°图像的多样性特征。采用人工像素级显著性标注，包含复杂室内外场景中的显著目标。

表7列出了360-SSOD数据集上各显著性检测网络的实验结果。CSMANet（Zhang等，2022）仍性能突出，MS-SOD（Ma等，2020）和MFFPANet（Chen等，2023a）也表现良好。虽在360-SOD数据集上表现一

表7　不同显著性目标检测(SOD)模型在360-SSOD数据集上的性能对比

Table 7　Performance Comparison of Different Salient Object Detection (SOD) Models on the 360-SSOD Dataset

年份	算法	FM ↑	SM ↑	Em ↑	MAE ↓	模型特点
2020	FANet(Huang等,2020)	0.423	0.657	0.706	0.052	ERP+CMP投影
	MS-SOD(Ma等,2020)	**0.82**	**0.902**	0.92	**0.015**	CNN
2022	CSMANet(Zhang等,2022)	0.661	0.784	0.859	0.028	ERP+CMP投影
2023	DTAFormer(Zhao等,2023)	0.644	0.7698	0.8322	0.0261	Transformer架构
	LDNet(Huang等,2023)	0.557	0.727	0.84	0.035	CNN
	MFFPANet(Chen等,2023a)	0.813	0.885	**0.925**	0.016	多模态
2024	ACoNet(陈晓雷等,2024)	0.6564	0.7796	0.8632	0.0288	ERP投影
2025	CPNet(Wen等,2025)	0.474	0.666	0.723	0.052	ERP投影

注:加粗字体表示每类指标最优值,下划线表示次优值。

般,但在语义均衡的360-SSOD数据集中,二者充分释放了多阶段语义排名与OLS互补优势,性能显著提升。MS-SOD采用分阶段框架,先通过对象级语义显著性排名实现对畸变的鲁棒定位,再在无畸变补丁上进行精细化检测,兼顾语义先验与几何校正;MFFPANet通过动态互补特征融合(DCFF)将物体级语义信息转化为空间-通道动态权重,自适应增强RGB特征,借助渐进式多尺度特征聚合(PMFA)逐级融合上下文,提升复杂场景下的显著目标完整性和边缘精度。二者分别从任务分解与多模态互补角度,为全景显著性目标检测提供了可借鉴的范例。

F-360iSOD(Zhang等,2020)是首个为360°全景图像提供实例级语义标注的数据集,包含107张高分辨率等角矩形全景图,涵盖复杂室内外场景。该数据集标注了1,165个显著物体,覆盖72个类别,为全景图像分析研究提供了丰富资源。

表8　不同显著性目标检测(SOD)模型在F-360iSOD数据集上的性能对比

Table 8　Performance Comparison of Different Salient Object Detection (SOD) Models on the F-360iSOD Dataset

年份	算法	FM ↑	SM ↑	Em ↑	MAE ↓	模型特点
2019	DDS(Li等,2019)	0.325	0.612	0.7	0.057	ERP投影
2020	FANet(Huang等,2020)	0.381	0.587	0.747	0.061	ERP+CMP投影
2023	MIDP-Net(Dai等,2023)	0.4099	0.6657	0.7705	0.0463	CNN
	MPFRNet(Cong等,2023)	0.813	**0.885**	**0.925**	**0.016**	ERP+CMP投影
	DTAFormer(Zhao等,2023)	**0.8322**	0.7664	0.7295	0.0361	Transformer架构
2025	CPNet(Wen等,2025)	0.382	0.651	0.729	0.051	ERP投影

注:加粗字体表示每类指标最优值,下划线表示次优值。

表8展示了不同显著性目标检测模型在F-360iSOD数据集上的运行结果。DTAFormer性能表现突出,失真感知机制在F-360iSOD数据集的实例级语义标注中得以充分验证,更好地适应了复杂的全景场景中的显著物体分割;MIDP-Net和MPFRNet也表现出色。尽管二者在360-SOD数据集上表现一般,但在图像畸变较轻的F-360iSOD数据集中,充分发挥了边缘-显著性联合细化与多投影互补融合的优势,性能显著提升。其中,MIDP-Net通过多层特征交互(MLFI)聚合跨尺度语义,借助两阶段解码器(TPD)联合边缘与显著性线索,实现精确分割;MPFRNet并行输入等距柱状图与四种立方体展开图,经动态加权融合(DWF)自适应整合投影互补特征,优化综合指标。两者分别从CNN内部结构优化与多投影互补融合角度,为全景显著性目标检测提供了技术路径。

Wu等人(2022)提出的ODI-SOD数据集是首个大规模360°全景图像显著性检测数据集,包含6,

263张2K分辨率等角矩形全景图像，涵盖多样室内外场景，并提供像素级物体显著性标注。

5.3 全景视频显著性检测数据集与模型性能

与传统2D显著性检测数据集相比，360°视频提供了更丰富场景和分散注意力分布，识别显著对象面临更大挑战。因此，研究者建立了专门的全景视频显著性检测数据集（如表9所示），以展现该领域的最新发展。

表9 全景视频显著性检测（SD）数据集

Table 9 Saliency Detection (SD) Dataset for Panoramic Videos

数据集	场景数量（个）	数据类型	分辨率（像素）	记录类型	检测任务
Salient360!V2（David等，2018）	19	真实拍摄	4 K	头部、眼部运动	SP
VR-scene（Xu等，2018b）	208	真实拍摄	4 K	头部、眼部运动	SP
360saliency（Zhang等，2018）	104	真实拍摄	150×300	眼部运动	SP
Wild-360（Cheng等，2018）	85	真实拍摄	1920×960～3840×1920	头部、眼部运动	SP
AVP-360（Chao等，2020a）	15	真实拍摄	4 K	头部运动	SP
PVS-HM（Xu等，2018a）	76	真实拍摄	3K～8K	头部、眼部运动	SP
YT360-EyeTracking（Cokelek等，2025）	81	真实拍摄	3840×1920	头部、眼部运动	SP
SHD360（Zhang等，2021）	41	真实拍摄	4 K	/	SOD
PAVS10K（Zhang等，2023b）	67	真实拍摄	4 K	眼部运动	SOD

5.3.1 全景视频显著性预测（SP）数据集与性能

大部分全景视频显著性预测（SP）数据集记录了观察者头部和眼部运动轨迹，部分数据集仅记录头部或眼部运动。这些数据集的视频来源包括真实拍摄和计算机合成，分辨率从1920×1080到8K不等。然而，多数数据集的视频数量相对有限，原因之一是全景视频采集及注视标注的高难度。

David等人（2018）构建的Salient360！V2数据集包含19个4K分辨率的360°全景视频，涵盖旅游、运动等内容；提供48名用户的头部运动轨迹、内容记忆，用于分析不同背景用户的观看行为差异。

Xu等人（2018b）构建的VR-scene数据集包含208个4K分辨率的360°动态视频，涵盖室内场景、户外活动、纪录片等多种内容。每个视频由至少31名参与者观看，使用VR头显眼动仪记录注视点数据。

Zhang等人（2018）构建的360saliency数据集包含104个360°全景视频，涵盖篮球、滑板等多种内容类型，每个视频由超过20名被试观看并采集眼动轨迹数据。

Cheng等人（2018）构建的Wild-360数据集包含85个360°全景视频共55000帧，其中60个用于训练、25个用于测试。这些视频来源于YouTube，涵盖自然、野生动物等多种内容类型。作者选取了具有挑战性的视频，包含多个显著目标分布于不同视角；对于测试集，收集了30名被试的注视轨迹数据，聚合生成逐帧显著性热图作为地面真值。

Chao等人（2020a）构建的AVP-360数据集包含15个全景视频（其中3个用于训练、12个用于测试），涵盖对话、音乐和环境三种内容类型，并提供静音、单声道和环境立体声三种音频模态。该数据集收集了45名被试（每种音频模态15名）的视点中心轨迹数据，考察音频和视觉信息对视觉注意力的影响，通过音频能量图（AEM）分析音频源位置与注意力的关系。

Xu等人（2018a）构建的PVS-HM数据集包含76个分辨率为3K至8K的全景视频序列，涵盖计算机动画、驾驶和风景等多种内容类型。每个视频序列时长为10至80秒（平均26.9秒），由58名参与者观看，使用HTC Vive头显记录头动数据，通过aGlass设备捕获眼动数据。

Cokelek等人（2025）构建的YT360-EyeTracking数据集包含81个分辨率为3840×1920像素的全景视频序列，涵盖音乐、纪录片等多种内容类型。每个视频序列时长为30秒，帧率为24至30fps。该数据集设置了三种音频条件（静音、单声道和环绕声），以系

统性地研究音频对全景视频显著性的影响，并记录了参与者的头部运动和眼动数据，为全景音视频显著性预测研究提供支持。

表10　全景视频显著性预测(SP)模型运行结果

Table 10　Results of the Saliency Prediction (SP) Model for Panoramic Video Operation

年份	算法	KLD ↓	CC ↑	NSS ↑	AUC ↑	数据集	模型特点
2021	Dahou等人（2021）	6.327	0.363	2.58	0.881	Salient360!V2	多模态
		5.561	0.363	2.185	0.862	VR-scene	
2021	Tliba等人（2021）	7.965	0.308	1.885	0.815	Salient360!V2	传统方法+深度学习
2022	Yun等人(2022)	/	0.616	/	0.923	Wild-360	Transformer架构
2023	Li等人(2023)	/	0.62	1.17	0.66	360saliency	球面投影
		/	0.62	1.11	0.65	VR-scene	
	Cokelek等人（2023）	5.744	0.586	2.63	/	VR-scene	Transformer架构
		1.841	0.626	2.191	/	PVS-HM	
2024	Zhu等人(2023)	/	0.768	2.69	0.922	AVP-360	多模态
	Yang等人（2024b）	/	0.787	3.838	0.774	PVS-HM	球面投影
		/	0.661	4.529	0.937	360saliency	
2025	Cokelek等人（2025）	5.334	0.599	2.821	/	VR-scene	多模态
		1.841	0.626	2.191	/	PVS-HMEM	
		8.341	0.512	2.449	/	YT360-EyeTracking	
	Zhu等人（2025c）	8.112	0.6271	3.9625	0.935	360saliency	

如表10所示，近年来全景视频显著性预测模型的性能整体呈现提升趋势。采用多模态方法和球面投影技术的研究逐渐增加。尽管全景视频显著性预测(SP)领域的数据集丰富多样，但这增加了不同模型间性能比较的难度。在性能指标上，Qiao等人(2020)、Yang等人(2024b)和Cokelek等人(2025)的模型表现出色。其中，Qiao等人(2020)通过ConvLSTM融合CNN提取的空间特征，显式建模视口位置分布偏置与对象吸引力，构建多任务学习框架；Yang等人(2024b)利用3D可分离图卷积网络在球面投影域内直接捕获光流与时空特征，通过注意力机制融合后重建ERP显著性图，避免重复投影失真；Cokelek等人(2025)凭借球面几何感知的时空注意力机制及音频-视觉融合适配器，有效地捕捉多模态交互和全局上下文信息。

5.3.2　全景视频显著性目标检测(SOD)数据集与性能

如表9下部分所示，目前全景视频显著性目标检测领域的数据集数量稀少，均包含基于真实拍摄的4K分辨率360°视频，其中PAVS10K数据集额外包含音频通道，支持视听联合建模。

Zhang等人(2021)提出的SHD360数据集含37,403帧4K分辨率全景视频，涵盖41个真实生活场景，包括室内活动和户外运动。该数据集为6,268个关键帧提供六级层次标注，包括超类、子类、相关属性、边界框、像素级对象和实例级掩码，标注了16,238个显著人类实例。

Zhang等人(2023)提出的PAVS10K是首个大规模全景视频显著性目标检测数据集，包含67个4K分辨率全景视频共62,455帧。该数据集提供每一帧的对象级和实例级像素级显著性掩码标注，并附带丰富的视频特征标注，包括视频类别、属性及多名被试的头部运动数据。

表11展示了全景视频显著性目标检测模型的运行结果。从表中可以看出，多模态方法——尤其是视听联合方法，已成为该领域的主流。其中，Guo

表11　全景视频显著性目标检测(SOD)模型运行结果

Table 11　Operating Results of the Panoramic Video Salient Object Detection (SOD) Model

年份	算法	FM ↑	SM ↑	Em ↑	MAE ↓	数据集	模型特点
2023	Chen等人(2023b)	0.7153	0.8179	0.8773	0.0323	SHD360	CNN
		0.3156	0.6371	0.649	0.029	PAVS10K	
	Zhang等人(2023b)	0.414	0.674	**0.747**	0.027	PAVS10K	多模态
	Li等人(2023b)	0.404	0.678	0.732	**0.026**	PAVS10K	多模态
2024	Guo等人(2024)	**0.432**	**0.699**	0.74	0.033	PAVS10K	多模态

注:加粗字体表示每类指标最优值,下划线表示次优值。

等人(2024)的模型表现最优,其失真感知像素解码器能有效抑制全景畸变,顺序式音频-视觉融合模块实现了声音的精准映射,将视听线索有效整合。

6　应　用

全景图像质量评估。全景图像质量评估通过建立客观的数学模型,量化失真对视觉感知质量的影响,为优化图像处理算法和评价显示设备性能提供重要支撑。在VR环境中,显著性检测技术通过模拟人类视觉注意力机制增强评估模型的精度。Qiu和Shao(2021)的显著性引导模型(SG360BIQA)利用显著性预测网络生成显著性图,将其作为注意力权重指导关键区域的特征提取。Yan等人(2025)的语义描述模型(IQCaption360)采用显著性检测进行自适应特征聚合,通过多任务学习将显著性引导的特征应用于失真预测和质量描述。这些方法通过显著性检测解决了传统方法忽视视觉重要区域的问题,推动全景图像质量评估向差异化方向发展。

全景视频质量评估。全景视频质量评估需要兼顾空间域质量和时间域的运动信息与帧间依赖关系。传统评估方法对所有区域采用均匀度量,忽视了人眼视觉显著性的差异,导致评估结果与用户主观感受不符。针对这一问题,研究者们将显著性检测引入全景视频质量评估,以量化用户的实际注视区域并赋予差异化评估权重。Yang等人(2024a)采用显著性预测模块引导视口提取,通过注意力机制建模视口质量变化,使评估更贴近人类视觉焦点。这类方法表明,显著性检测作为关键技术,通过量化用户注视偏好和视觉优先级,显著改进了全景视频质量评估的主观一致性。

全景图像压缩与传输。全景图像压缩与传输需在带宽受限的条件下平衡编码效率与视觉保真度。传统均匀分配码率的方法难以优先保护用户实际注视的高显著性区域,导致主观质量下降和带宽浪费。为此,研究者将显著性检测技术引入码率分配策略,以实现差异化的视觉质量保护。如:Wang等人(2023)提出的RoSal360模型则以显著性权重为核心,结合强化学习实时优化传输策略,在带宽约束和时延限制下动态调整各区域的编码质量。上面的方法表明,显著性检测通过量化用户的视觉优先级,能够在有限的带宽条件下实现码率与保真度的最优平衡,提高了全景媒体传输的效率和用户体验。

全景显著性检测在360°视频虚拟摄影中也有重要作用。传统虚拟摄影方法忽视人类注视偏好,导致窄视口(NFOV)选取偏离真实观看轨迹。Wang等人(2020)基于显著性图设计了深度强化学习模型,将全景显著性作为状态输入、嵌入奖励函数,使智能体以显著区域为导向实现平滑的虚拟摄影。这些应用表明,全景显著性检测通过量化人眼注视偏好,为虚拟摄影、质量评估等多个全景内容处理领域提供了关键支持,极大提高了用户体验与系统性能。

7　挑战与未来研究趋势

尽管VR360°全景图像和视频显著性检测领域通过深度学习方法已实现了实现了诸多技术突破,在显著性预测上取得了实质性进展,但仍存在技术瓶颈和实际挑战。未来潜力巨大,需要进一步创新。本节将分析关键问题并探讨研究方向,以推进显著性检测技术在VR中的发展。

7.1　挑战

(1)**全景数据中的畸变问题**,这种畸变主要由球面投影技术引起,导致图像边缘和曲面区域失真,从

而影响特征提取的准确性。此外，不同视角下的特征提取难题进一步加剧了这一问题，原因在于用户头部运动和视点变化会动态改变显著性分布，难以模拟VR用户的注意机制。

(2)**实时性要求**。现有的深度学习框架因复杂的网络结构和大量参数导致高计算开销，难以实时处理全景数据流。模型需采用高效计算和轻量化设计，以满足VR环境的动态交互和低延迟需求。

(3)**数据集和模型的不足**。全景视频显著性检测模型的数量远少于图像模型，尤其缺乏针对显著性目标检测的专用框架，这限制了研究者在动态视频场景中的创新和应用扩展；同时，现有的全景视频显著性预测数据集虽丰富多样，但缺乏一个大型、统一的基准数据集，导致模型训练的泛化性和评估标准的统一性难以保障。

7.2 未来研究趋势

(1)**扩展多模态信息整合。**多模态融合在传统2D图像显著性检测领域取得了显著成果。然而，在全景环境下，现有的研究主要依赖视觉和听觉模态，通过引入深度信息(如3D深度图)和用户热度数据(如注意力热力图)，可以提供更丰富的语境，提升模型对动态场景的理解和预测准确性。这些不同的融合方式有助于模拟真实VR用户行为，克服单一模态的局限性，并增强显著性检测的鲁棒性。

(2)**数据集以及综合评价指标。**需开发多模态(如视觉和听觉整合)和大型统一的全景视频显著性预测基准数据集，并设计相应的标注体系，以促进模型训练和评估的一致性。还要制定适应不同场景的多维度指标，综合考虑实时性能、用户沉浸感，从而更全面衡量模型有效性。

迁移学习在计算机视觉任务中已被证明是一种有效的技术，特别是在数据稀缺或标注成本高昂的场景下具备显著优势。然而，在全景显著性检测领域，现有研究较少探索跨域或跨任务的迁移学习应用。未来的研究可以聚焦于从传统2D显著性检测模型或相关VR任务(如视口预测)中迁移预训练知识到全景环境，以减少对大规模标注全景数据集的依赖。

(3)**轻量化和无监督/弱监督模型。**开发针对球面数据特性的专用注意力模块，可有效处理全景图像的几何畸变和动态场景。同时，采用无监督或弱监督学习方法，能减少对大规模标注数据的依赖，降低训练门槛，并提升模型的泛化能力。

对比学习作为一种自监督学习范式，近年来在特征表示学习中展现出强大潜力。在全景显著性检测中，对比学习可用于挖掘球面图像中显著区域与非显著区域的潜在差异，通过构建正负样本对来增强模型对复杂场景中视觉重要性的判别能力。未来的研究方向可以探索如何设计适用于全景数据的对比损失函数，以及如何结合球面几何特性优化样本选择策略，从而提升模型在动态VR环境中的鲁棒性和适应性。

(4)数据增强。数据增强技术在提升模型训练效果和泛化能力方面发挥了重要作用，但在全景显著性检测中的应用仍显不足。鉴于全景图像的高分辨率和球面畸变特性，传统2D图像增强方法(如旋转、翻转)可能无法直接应用。未来的研究可以开发针对全景数据的新型增强策略，例如基于球面几何的视点模拟变换等，以模拟真实VR用户交互中的多样化视角和行为模式。这将有助于缓解数据稀缺问题，提升模型对复杂全景环境的适应能力。

8 结 语

本文对VR360°全景图像与视频显著性检测进行了全面综述，从基本理论与背景概述入手，逐层探讨了传统方法和深度学习方法的演进、数据集与评价指标的应用、实际领域的推广以及面临的挑战与未来趋势。系统分析了现有技术的优势与局限，强调了显著性检测在提升用户体验和系统优化方面的关键作用，为后续研究提供了理论框架和实践参考。未来，随着多模态融合和轻量化模型的深入探索，该领域有望进一步推动虚拟现实技术的创新与落地。

参考文献(References)

Achanta R, Hemami S, Estrada F and Susstrunk S. 2009. Frequency tuned salient region detection//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 1597-1604 [DOI: 10.1109/CVPR.2009.5206596]

Battisti F and Carli M. 2019. Depth-based saliency estimation for omnidirectional images. Electronic Imaging, 2019(10): 1-5 [DOI: 10.2352/ISSN.2470-1173.2019.11.IPAS-271]

Borji A, Sihite D N and Itti L. 2012. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study.

IEEE Transactions on Image Processing, 22(1): 55-69 [DOI: 10.1109/TIP.2012.2210727]

Bravais A. 1844. Analyse mathématique sur les probabilités des erreurs de situation d'un point. Paris: Imprimerie Royale

Buzzelli M. 2020. Recent advances in saliency estimation for omnidirectional images, image groups, and video sequences. Applied Sciences, 10(15): 5143 [DOI: 10.3390/app10155143]

Chao F Y, Ozcinar C, Wang C, Zerman E, Zhang L and Hamidouche W. 2020a. Audio-visual perception of omnidirectional video for virtual reality applications//Proceedings of the 2020 IEEE International Conference on Multimedia and Expo Workshops. London: IEEE: 1-6 [DOI: 10.1109/ICMEW46912.2020.9105956]

Chao F Y, Zhang L, Hamidouche W and Déforges O. 2020b. A multi-FoV viewport-based visual saliency model using adaptive weighting losses for 360° images. IEEE Transactions on Multimedia, 23: 1811-1826 [DOI: 10.1109/TMM.2020.3003642]

Chen G, Shao F, Chai X L, Jiang Q P and Ho Y S. 2023a. Multi-stage salient object detection in 360 omnidirectional image using complementary object-level semantic information. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(1): 776-789 [DOI: 10.1109/TETCI.2023.3259433]

Chen X L, Du Z L, Zhang X G and Wang X. 2025. Distortion-adaptive and position-aware network for salient object detection in 360° omnidirectional image. Journal of Image and Graphics, 30(8): 2758-2774 (陈晓雷, 杜泽龙, 张学功, 王兴. 2025. 畸变自适应与位置感知的360°全景图像显著目标检测网络. 中国图象图形学报, 30(8): 2758-2774) [DOI: 10.11834/jig.240592]

Chen X L, Wang X, Zhang X G and Du Z L. 2024. Adjacent-coordination network for 360° panoramic salient object detection. Journal of Electronics & Information Technology, 46(12): 1-10 (陈晓雷, 王兴, 张学功, 杜泽龙. 2024. 面向360度全景图像显著目标检测的相邻协调网络. 电子与信息学报, 46(12): 1-10) [DOI: 10.11999/JEIT240502]

Chen X L, Zhang P C, Du Z L and Ahmad I. 2023b. A spatial-temporal dual-mode mixed flow network for panoramic video salient object detection[EB/OL]. [2023-10-13]. https://arxiv.org/abs/2310.09016

Chen X L, Zhang P C, Lu Y B and Cao B N. 2023c. Salient object detection for panoramic images based on robust vision transformer and multi-attention. Journal of Electronics & Information Technology, 45(6): 1-9 (陈晓雷, 张鹏程, 卢禹冰, 曹宝宁. 2023. 基于鲁棒视觉变换和多注意力的全景图像显著性检测. 电子与信息学报, 45(6): 1-9) [DOI: 10.11999/JEIT220684]

Cheng H T, Chao C H, Dong J D, Wen H K, Liu T L and Sun M. 2018. Cube padding for weakly-supervised saliency prediction in 360 videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 1420-1429 [DOI: 10.1109/CVPR.2018.00154]

Cokelek M, Imamoglu N, Ozcinar C, Erdem E and Erdem A. 2021. Leveraging frequency based salient spatial sound localization to improve 360 video saliency prediction//Proceedings of the 2021 17th International Conference on Machine Vision and Applications. Tokyo: IEEE: 1-6 [DOI: 10.23919/MVA51890.2021.9511406]

Cokelek M, Imamoglu N, Ozcinar C, Erdem E and Erdem A. 2023. Spherical vision transformer for 360-degree video saliency prediction[EB/OL]. [2023-08-25]. https://arxiv.org/abs/2308.13004

Cokelek M, Ozsoy H, Imamoglu N, Ozcinar C, Ayhan I and Erdem E. 2025. Spherical vision transformers for audio-visual saliency prediction in 360° videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3): 1-15 [DOI: 10.1109/TPAMI.2025.3604091]

Cong R, Huang K, Lei J, Zhao Y, Huang Q and Kwong S. 2023. Multi-projection fusion and refinement network for salient object detection in 360 omnidirectional image. IEEE Transactions on Neural Networks and Learning Systems, 35(7): 9495-9507 [DOI: 10.1109/TNNLS.2022.3233883]

Dahou Y, Tliba M, McGuinness K and O'Connor N. 2021. ATSal: an attention based architecture for saliency prediction in 360° videos//Proceedings of the 2021 International Conference on Pattern Recognition. Milan: Springer: 433-448 [DOI: 10.1007/978-3-030-68796-0_22]

Dai F, Zhang Y Q, Ma Y K, Li H L and Zhao Q. 2020. Dilated convolutional neural networks for panoramic image saliency prediction//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE: 2558-2562 [DOI: 10.1109/ICASSP40776.2020.9053888]

Dai H W, Bao L X, Shen K Y, Zhou X F and Zhang J Y. 2023. 360° omnidirectional salient object detection with multi-scale interaction and densely-connected prediction//Proceedings of the 11th International Conference on Image and Graphics. Cham: Springer: 427-438 [DOI: 10.1007/978-3-031-46305-1_35]

David E J, Gutiérrez J, Coutrot A, Da Silva M P and Le Callet P. 2018. A dataset of head and eye movements for 360 videos//Proceedings of the 9th ACM Multimedia Systems Conference. Amsterdam: ACM: 432-437 [DOI: 10.1145/3204949.3208139]

Ding Y, Liu Y W, Liu J X, Liu K D, Wang L M and Xu Z. 2019. A survey on saliency detection for virtual reality panoramic images. Acta Electronica Sinica, 47(7): 1575-1583 (丁颖, 刘延伟, 刘金霞, 刘科栋, 王利明, 徐震. 2019. 虚拟现实全景图像显著性检测研究进展综述. 电子学报, 47(7): 1575-1583) [DOI: 10.3969/j.issn.0372-2112.2019.07.024]

Djilali Y A D, McGuinness K and O'Connor N E. 2021. Simple baselines can fool 360° saliency metrics//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE: 3743-3749 [DOI: 10.1109/ICCVW54120.2021.00333]

Fan D P, Cheng M M, Liu Y, Li T and Borji A. 2017. Structure-

measure: a new way to evaluate foreground maps//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4548-4557 [DOI: 10.1109/ICCV.2017.487]

Fan D P, Gong C, Cao Y, Ren B, Cheng M M and Borji A. 2018. Enhanced-alignment measure for binary foreground map evaluation//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press: 698-704

Fang Y M, Zhang X Q and Imamoglu N. 2018. A novel superpixel-based saliency detection model for 360-degree images. Signal Processing: Image Communication, 69: 1-7 [DOI: 10.1016/j. image. 2018. 07.009]

Feng D J, Chen H Y, Liu S N, Liao Z Y, Shen X Y, Xie Y K and Zhu J. 2023. Boundary-semantic collaborative guidance network with dual-stream feedback mechanism for salient object detection in optical remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 61: 4706317 [DOI: 10.1109/TGRS. 2023. 3332282]

Guo R, Niu D, Qu L, Qi Y, Shi J, Yue W, Xing B, Chen T and Ying X. 2024. Instance-level panoramic audio-visual saliency detection and ranking//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne: ACM: 1-9 [DOI: 10.1145/3664647.3681070]

Harel J, Koch C and Perona P. 2006. Graph-based visual saliency// Advances in Neural Information Processing Systems 19. Vancouver: MIT Press: 545-552

Huang M K, Li G Y, Liu Z and Zhu L C. 2023. Lightweight distortion-aware network for salient object detection in omnidirectional images. IEEE Transactions on Circuits and Systems for Video Technology, 33 (10): 6191-6197 [DOI: 10.1109/TCSVT. 2023. 3253685]

Huang M K, Liu Z, Li G Y, Zhou X F and Le Meur O. 2020. FANet: features adaptation network for 360° omnidirectional salient object detection. IEEE Signal Processing Letters, 27: 1819-1823 [DOI: 10.1109/LSP.2020.3028192]

Itti L, Koch C and Niebur E. 2002. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11): 1254-1259 [DOI: 10.1109/34.730558]

Jin H Y, Xiao Z L, Cai L and Wang B. 2022. Review on the theory and application of saliency object detection. Computer Technology and Development, 32(9): 1-7 (金海燕, 肖照林, 蔡磊, 王彬. 2022. 显著性目标检测理论与应用研究综述. 计算机技术与发展, 32(9): 1-7) [DOI: 10.3969/j.issn.1673-629X.2022.09.001]

Judd T, Durand F and Torralba A. 2012. A benchmark of computational models of saliency to predict human fixations. MIT-CSAIL-TR-2012-001. MIT

Koch C and Ullman S. 1987. Shifts in selective visual attention: towards the underlying neural circuitry//Vaina L M, ed. Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience. Dordrecht: Springer Netherlands: 115-141

Kullback S and Leibler R A. 1951. On information and sufficiency. Annals of Mathematical Statistics, 22(1): 79-86 [DOI: 10.1214/aoms/1177729694]

Lebreton P and Raake A. 2018. GBVS360, BMS360, ProSal: extending existing saliency prediction models from 2D to omnidirectional images. Signal Processing: Image Communication, 69: 69-78 [DOI: 10.1016/j.image.2018.03.006]

Li J, Han L, Zhang C, Li Q Y and Liu Z. 2023a. Spherical convolution empowered viewport prediction in 360 video multicast with limited FoV feedback. ACM Transactions on Multimedia Computing, Communications and Applications, 19 (1): 1-23 [DOI: 10.1145/3511603]

Li J, Su J M, Xia C Q and Tian Y H. 2019. Distortion-adaptive salient object detection in 360° omnidirectional images. IEEE Journal of Selected Topics in Signal Processing, 14(1): 38-48 [DOI: 10.1109/JSTSP.2019.2957982]

Li W R, Xu D, Shi J L and Huang S C. 2022. Survey of salient object detection: Methods, applications and trends. Application Research of Computers, 39(7): 1921-1929 (李婉蓉, 徐丹, 史金龙, 黄树成. 2022. 显著性物体检测研究综述: 方法, 应用和趋势. 计算机应用研究, 39(7): 1921-1929) [DOI: 10.19734/j.issn.1001-3695.2021.12.0645]

Li X, Cao H Y, Zhao S J, Li J L, Zhang L and Raj B. 2023b. Panoramic video salient object detection with ambisonic audio guidance//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI: 1424-1432 [DOI: 10.1609/aaai.v37i2.25227]

Lin Y H, Sun H, Liu N Z, Bian Y T, Cen J and Zhou H Y. 2022. A lightweight multi-scale context network for salient object detection in optical remote sensing images//Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Montreal, Canada: IEEE: 238-244 [DOI: 10.1109/ICPR56361.2022.9956350]

Ling J, Zhang K, Zhang Y, Yang D and Chen Z. 2018. A saliency prediction model on 360 degree images using color dictionary based sparse representation. Signal Processing: Image Communication, 69: 60-68 [DOI: 10.1016/j.image.2018.03.007]

Liu N, Luo Z, Zhang N and Han J. 2024. VST++: efficient and stronger visual saliency transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(11): 7300-7316 [DOI: 10.1109/TPAMI.2024.3388153]

Lv H, Yang Q, Li C, Dai W, Zou J and Xiong H. 2020. SalGCN: saliency prediction for 360-degree images based on spherical graph convolutional networks//Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM: 372-380 [DOI: 10.1145/3394171.3413733]

Ma G X, Li S, Chen C L Z, Hao A and Qin H. 2020. Stage-wise salient object detection in 360 omnidirectional image via object-level

semantical saliency ranking. IEEE Transactions on Visualization and Computer Graphics, 26(12): 3535-3545 [DOI: 10.1109/TVCG.2020.3023636]

Martin D, Serrano A and Masia B. 2020. Panoramic convolutions for 360 single-image saliency prediction//Proceedings of the CVPR Workshop on Computer Vision for Augmented and Virtual Reality. Seattle: IEEE: 1-10 [DOI: 10.1109/CVPRW50498.2020.00055]

Mazumdar P and Battisti F. 2019. A content-based approach for saliency estimation in 360 images//Proceedings of the 2019 IEEE International Conference on Image Processing. Taipei: IEEE: 3197-3201 [DOI: 10.1109/ICIP.2019.8803296]

Monroy R, Lutz S, Chalasani T and Smolic A. 2018. SalNet360: saliency maps for omni-directional images with CNN. Signal Processing: Image Communication, 69: 26-34 [DOI: 10.1016/j.image.2018.05.005]

Nguyen A, Yan Z and Nahrstedt K. 2018. Your attention is unique: detecting 360-degree video saliency in head-mounted display for head movement prediction//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Republic of Korea: ACM: 1190-1198 [DOI: 10.1145/3240508.3240669]

Perazzi F, Krähenbühl P, Pritch Y and Hornung A. 2012. Saliency filters: contrast based filtering for salient region detection//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 733-740 [DOI: 10.1109/CVPR.2012.6247743]

Peters R J, Iyer A, Itti L and Koch C. 2005. Components of bottom-up gaze allocation in natural images. Vision Research, 45(18): 2397-2416 [DOI: 10.1016/j.visres.2005.02.004]

Qiao M L, Xu M, Wang Z L and Borji A. 2020. Viewport-dependent saliency prediction in 360 video. IEEE Transactions on Multimedia, 23: 748-760 [DOI: 10.1109/TMM.2020.2987682]

Qiu M and Shao F. 2021. Blind 360-degree image quality assessment via saliency-guided convolutional neural network. Optik, 240: 166858 [DOI: 10.1016/j.ijleo.2021.166858]

Rai Y, Gutiérrez J and Le Callet P. 2017. A dataset of head and eye movements for 360 degree images//Proceedings of the 8th ACM on Multimedia Systems Conference. Taipei: ACM: 205-210 [DOI: 10.1145/3192974]

Ramenahalli S. 2020. A biologically motivated, proto-object-based audiovisual saliency model. AI, 1(4): 487-509 [DOI: 10.3390/ai1040030]

Sitzmann V, Serrano A, Pavel A, Agrawala M, Gutierrez D and Masia B. 2018. Saliency in VR: how do people explore virtual environments? IEEE Transactions on Visualization and Computer Graphics, 24(4): 1633-1642 [DOI: 10.1109/TVCG.2018.2793599]

Song J, Hu S and Chen C. 2025. Breaking the dataset shackles: Data-efficient learning with Mamba network for 360° salient object detection. Journal of Physics: Conference Series, 3072(1): 012004 [DOI: 10.1088/1742-6596/3072/1/012004]

Su Q, Lin C Y, Zhao Y, Li Y R and Liu M Q. 2018. Saliency detection of 360 panoramic images based on multi-angle segmentation. Journal of Graphics, 39(6): 1055 (苏群, 林春雨, 赵耀, 李雅茹, 刘美琴. 2018. 基于多角度分割的360全景图的显著性检测. 图学学报, 39(6): 1055 [DOI: 10.11996/JG.j.2095-302X.2018061055]

Sui X J, Fang Y M, Zhu H W, Wang S Q and Wang Z. 2023. Scan-DMM: a deep Markov model of scanpath prediction for 360° images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 6989-6999 [DOI: 10.1109/CVPR52729.2023.00676]

Tliba M, Sayah M and Djilali Y. 2021. 2D-based saliency prediction framework for omnidirectional - 360° video//Proceedings of the 11th International Conference on Pattern Recognition Systems. Lille: IET: 1-6 [DOI: 10.1049/icp.2021.1459]

Wan Z, Qin H, Li Z, Fan X, Zuo W and Zhao D. 2025. CASP: Consistency-aware audio-induced saliency prediction model for omnidirectional video//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 12605-12614 [DOI: 10.1109/CVPR52734.2025.01176]

Wan Z, Qin H, Xiong R, Li Z, Fan X and Zhao D. 2024. Predicting 360 video saliency: a ConvLSTM encoder-decoder network with spatio-temporal consistency. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 14(2): 311-322 [DOI: 10.1109/JETCAS.2024.3377096]

Wang J, Xu M, Jiang L and Song Y. 2020. Attention-based deep reinforcement learning for virtual cinematography of 360° videos. IEEE Transactions on Multimedia, 23: 3227-3238 [DOI: 10.1109/TMM.2020.3021984]

Wang S, Yang S, Su H, Zhao C, Xu C and Qian F. 2023. Robust saliency-driven quality adaptation for mobile 360-degree video streaming. IEEE Transactions on Mobile Computing, 23(2): 1312-1329 [DOI: 10.1109/TMC.2023.3235103]

Wang W, Lai Q, Fu H, Shen J, Ling H and Yang R. 2021. Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 3239-3259 [DOI: 10.1109/TPAMI.2021.3051099]

Wen H F, Zhu Z J, Zhou X F, Zhang J Y and Yan C G. 2025. Consistency perception network for 360° omnidirectional salient object detection. Neurocomputing, 620: 129243 [DOI: 10.1016/j.neucom.2024.129243]

Wu J, Xia C, Yu T and Li J. 2022. View-aware salient object detection for 360° omnidirectional image. IEEE Transactions on Multimedia, 25: 6471-6484 [DOI: 10.1109/TMM.2022.3209015]

Xie C X, Xia C Q, Ma M C, Zhao Z R, Chen X W and Li J. 2022. Pyramid grafting network for one-stage high resolution saliency detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11707-11716 [DOI: 10.1109/CVPR52688.2022.01142]

Xu J W, Zhou Q Q, Li Z P, Shi Y J, Yi Y G and Yu J C. 2025a. HVPNet: A Unified Bio-Inspired Network for General Salient and Camouflaged Object Detection[EB/OL]. [2025-11-06]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5637952

Xu J W, Zhou Q Q, Yu J C, Liao C and Zhu D D. 2025b. Semantic-orthogonal multi-modal attention network for RGB-D salient object detection. The Visual Computer, 41(3): 1-13 [DOI: 10.1007/s00371-025-04013-5]

Xu M, Song Y H, Wang J, Qiao M G, Huo L Y and Wang Z L. 2018a. Predicting head movement in panoramic video: a deep reinforcement learning approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(11): 2693-2708 [DOI: 10.1109/TPAMI.2018.2858783]

Xu M, Yang L, Tao X, Duan Y Y and Wang Z L. 2021a. Saliency prediction on omnidirectional image with generative adversarial imitation learning. IEEE Transactions on Image Processing, 30: 2087-2102 [DOI: 10.1109/TIP.2021.3050861]

Xu Y, Dong Y, Wu J, Sun Z, Shi Z and Yu J. 2018b. Gaze prediction in dynamic 360 immersive videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 5333-5342 [DOI: 10.1109/CVPR.2018.00559]

Yamanaka T, Suzuki T, Nobutsune T and Wu C. 2023. Multi-scale estimation for omni-directional saliency maps using learnable equator bias. IEICE Transactions on Information and Systems, E106-D(10): 1723-1731 [DOI: 10.1587/transinf.2023EDP7055]

Yan J, Tan Z, Fang Y, Chen J, Jiang W and Wang Z. 2025. Omnidirectional image quality captioning: a large-scale database and a new model. IEEE Transactions on Image Processing, 34: 1-14 [DOI: 10.1109/TIP.2025.3539468]

Yang F, Yang C, An P and Huang X. 2024a. 360 video quality assessment based on saliency-guided viewport extraction. Multimedia Systems, 30(2): 89-102 [DOI: 10.1007/s00530-024-01285-0]

Yang Q, Gao W X, Li C L, Wang H, Dai W R, Zou J N, Xiong H K and Frossard P. 2024b. 360SPred: saliency prediction for 360-degree videos based on 3D separable graph convolutional networks. IEEE Transactions on Circuits and Systems for Video Technology, 34(10): 9979-9996 [DOI: 10.1109/TCSVT.2024.3407685]

Yi K, Li Y M, Xu J and Zhang J. 2025. Dual mutual learning network with global-local awareness for RGB-D salient object detection. Circuits, Systems, and Signal Processing, 44(2): 1-28 [DOI: 10.1007/s00034-025-03143-4]

Yun H, Lee S H and Kim G. 2022. Panoramic vision transformer for saliency detection in 360° videos//Proceedings of the European Conference on Computer Vision. Tel Aviv: Springer: 1-17 [DOI: 10.1007/978-3-031-19833-5_25]

Zhang R P, Chen C Y and Peng J. 2024. Multi-scale graph feature extraction network for panoramic image saliency detection. The Visual Computer, 40(2): 953-970 [DOI: 10.1007/s00371-023-02825-x]

Zhang R P, Chen C Y, Zhang J C, Peng J and Alzbier A M T. 2023a. 360-degree visual saliency detection based on fast-mapped convolution and adaptive equator-bias perception. The Visual Computer, 39(3): 1163-1180 [DOI: 10.1007/s00371-021-02395-w]

Zhang Y, Chao F Y, Hamidouche W and Deforges O. 2023b. PAV-SOD: a new task towards panoramic audiovisual saliency detection. ACM Transactions on Multimedia Computing, Communications and Applications, 19(3): 1-26 [DOI: 10.1145/3565267]

Zhang Y, Hamidouche W and Deforges O. 2022. Channel-spatial mutual attention network for 360 salient object detection//Proceedings of the 26th International Conference on Pattern Recognition. Montreal: IEEE: 3436-3442 [DOI: 10.1109/ICPR56361.2022.9956354]

Zhang Y, Zhang L, Hamidouche W and Déforges O. 2020. A fixation-based 360 benchmark dataset for salient object detection//Proceedings of the 2020 IEEE International Conference on Image Processing. Abu Dhabi: IEEE: 3458-3462 [DOI: 10.1109/ICIP40778.2020.9191158]

Zhang Y, Zhang L, Wang K, Hamidouche W and Déforges O. 2021. SHD360: a benchmark dataset for salient human detection in 360 videos[EB/OL]. [2021-05-18]. https://arxiv.org/abs/2105.08923

Zhang Z H, Xu Y Y, Yu J Y and Gao S H. 2018. Saliency detection in 360 videos//Proceedings of the European Conference on Computer Vision. Munich: Springer: 488-503 [DOI: 10.1007/978-3-030-01234-2_30]

Zhao Y J, Zhao L C, Yu Q, Zhang J, Sheng L and Xu D. 2023. Distortion-aware transformer in 360° salient object detection//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM: 499-508 [DOI: 10.1145/3581783.3612234]

Zhu C, Huang K and Li G. 2017. Automatic salient object detection for panoramic images using region growing and fixation prediction model[EB/OL]. [2017-10-11]. https://arxiv.org/abs/1710.04071

Zhu D D, Zhang K W, Min X K, Zhai G T and Yang X K. 2025a. From haziness to clarity: a novel iterative memory-retrospective emergence model for omnidirectional image saliency prediction. IEEE Transactions on Image Processing, 34: 1-14 [DOI: 10.1109/TIP.2025.3578264]

Zhu D D, Zhang K W, Min X K, Zhai G T and Yang X K. 2025b. ScanDTM: a novel dual-temporal modulation scanpath prediction model for omnidirectional images. IEEE Transactions on Circuits and Systems for Video Technology, 35(3): 1-14 [DOI: 10.1109/TCSVT.2025.3545908]

Zhu D D, Zhang K W, Zhang N N, Zhou Q Q, Min X K, Zhai G T and Yang X K. 2023. Unified audio-visual saliency model for omnidirectional videos with spatial audio. IEEE Transactions on Multimedia, 26: 764-775 [DOI: 10.1109/TMM.2023.3271022]

Zhu Y, Duan H, Zhang K, Zhu Y, Zhu X, Teng L, Min X K and Zhai

G T. 2025c. How does audio influence visual attention in omnidirectional videos? Database and model. IEEE Transactions on Image Processing, 34: 3447-3462 [DOI: 10.1109/TIP.2025.3567842]

Zhu W J, Liang S, Wei Y C and Sun J. 2014. Saliency optimization from robust background detection//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 2814-2821 [DOI: 10.1109/CVPR.2014.360]

Zou Z Z, Ye M, Li S, Li X and Dufaux F. 2023. 360° image saliency prediction by embedding self-supervised proxy task. IEEE Transactions on Broadcasting, 69(3): 704-714 [DOI: 10.1109/TBC.2023.3254143]