全景图像视频的场景分析与内容处理方法综述

谢红霞[1]，胡毓宁[1]，张赟[2]，王亚奇[2]，杜辉[2]，秦爱红[2]

(1. 浙大城市学院计算机与计算科学学院，浙江 杭州 310015；
2. 浙江传媒学院媒体工程学院，浙江 杭州 310018)

摘　　要：近年来，随着全景内容获取和交互的软硬件技术的快速发展，全景图像视频的数量激增，如何对全景内容进行高质量地分析和处理越来越成为虚拟现实领域的研究热点。当前，全景内容分析和处理在理论和应用方面面临着巨大挑战，关于该领域的关键问题在已有文献中未见系统全面地总结和研究。为了更好地促进该领域的研究和应用推广，针对全景图像视频的场景分析与内容处理近期的主要工作进行综述。在全景场景分析方面，分析了全景图像视频的深度学习网络、深度恢复、重要性检测、目标检测的研究工作；在全景内容处理方面，分析了全景图像视频的交互式浏览、去抖和校正、内容编辑的研究工作。最后，对综述进行了总结，并展望了未来在立体视图下全景图像视频的场景分析与内容处理方面的研究趋势。

关 键 词：虚拟现实；全景图像视频；场景分析；内容处理；立体视图

中图分类号：TP 391　　**DOI**：10.11996/JG.j.2095-302X.2023040640

文献标识码：A　　**文 章 编 号**：2095-302X(2023)04-0640-18

Survey of methods for scene analysis and content processing in panoramic images and videos

XIE Hong-xia[1], HU Yu-ning[1], ZHANG Yun[2], WANG Ya-qi[2], DU Hui[2], QIN Ai-hong[2]

(1. School of Computer & Computing Science, Hangzhou City University, Hangzhou Zhejiang 310015, China;
2. College of Media Engineering, Communication University of Zhejiang, Hangzhou Zhejiang 310018, China)

Abstract: In recent years, the rapid development of software and hardware technologies for acquiring and interacting with panoramic content has led to a significant increase in the number of panoramic images and videos. Immersive media with 360-degree panoramic images and videos as the main content has been widely used in the field of virtual reality and enhancement implementation. Compared with traditional 2D images and videos, panoramic images and videos can provide users with a new immersive experience. With wearable devices, users can freely watch the content from all perspectives through head movement. At present, the number of panoramic images and videos has soared, but it is usually difficult to obtain satisfactory panoramic images and videos, due to the difficulty in obtaining panoramic content and the lack of effective editing tools. Therefore, analyzing and processing panoramic content with high

收稿日期：2023-02-03；**定稿日期**：2023-04-11

Received：3 February，2023；**Finalized**：11 April，2023

基金项目：浙江省基础公益研究计划项目(LGG22F020009，LGF21F020002，LGF22F020015)；国家自然科学基金项目(62206242)；浙江省影视媒体技术研究重点实验室开放课题(2020E10015)；浙江传媒学院2021年第十六批教学改革项目(jgxm202131)

Foundation items：Zhejiang Province Public Welfare Technology Application Research (LGG22F020009，LGF21F020002，LGF22F020015); National Natural Science Foundation of China (62206242); Key Laboratory of Film and TV Media Technology of Zhejiang Province (2020E10015); The 16th Teaching Reform Project in 2021 of Communication University of Zhejiang (jgxm202131)

第一作者：谢红霞(1971–)，女，讲师，硕士。主要研究方向为计算机技术应用、虚拟现实。E-mail：xiehx@zucc.edu.cn

First author：XIE Hong-xia (1971–), lecturer, master. Her main research interests cover computer technology application, virtue reality. E-mail：xiehx@zucc.edu.cn

通信作者：张赟(1984–)，男，教授，博士。主要研究方向为计算机图形学、虚拟现实等。E-mail：zhangyun@cuz.edu.cn

Corresponding author：ZHANG Yun (1984–), professor, Ph.D. His main research interests cover computer graphics, virtue reality, etc. E-mail：zhangyun@cuz.edu.cn

quality has become an increasingly important research topic in the field of virtual reality. However, both in theory and application, the analysis and processing of panoramic content face significant challenges. Despite this, there is a lack of systematic and comprehensive summaries and research on the key issues in this field in existing literature. In order to better promote research and application in this area, a survey was provided on the recent works of scene analysis and content processing of panoramic images and videos. In terms of panoramic scene analysis, this survey reviewed the research on depth learning networks, depth recovery, importance detection, and target detection for panoramic images and videos. In terms of panoramic content processing, the survey analyzed the research on interactive browsing, stabilization and correction, and content editing of panoramic image video. Finally, the overview was summarized, with an outlook on future research trends in scene analysis and content processing of panoramic images and videos under the stereo view.

Keywords: virtual reality; panoramic images and videos; scene analysis; content processing; stereoscopic views

近年来，以 360°图像和视频作为主要内容的沉浸式媒体在虚拟现实(virtue reality，VR)和增强现实(augmented reality，AR)领域得到了广泛关注。人们借助穿戴式设备，如 Vive XR elite[1]和 Meta quest pro[2]等(图 1)，通过简单的头部运动自由地在所有视角观看全景图像视频，获得沉浸式的交互体验。目前，VR 技术经过多年积累已经日趋成熟，并且在教育、医疗、影视娱乐、数字游戏等诸多领域得到了成功应用。

图 1　最新穿戴式设备的用户体验效果图[1-2]及 360°全景图像的不同平面投影方式

Fig. 1　User experience renderings of latest wearable devices[1-2] and different planar projections of 360° panoramas

在 VR 世界中，为人们带来沉浸式体验的关键因素是 VR 内容及其交互体验方式。VR 内容通常包括全景图像视频、计算机生成的 3D 模型和虚拟场景等。本文的研究对象是 360°全景图像视频，与视野受限的传统 2D 媒体相比，全景图像视频定义在球面，包含了 360°×180°视野范围，能够完全覆盖用户观看的所有视角。如图 1 所示，为了处理方便，定义在球面的全景图像经常被投影到 2D 平面，如等矩形投影(equirectangular projection，ERP)、立方体投影(Cub map)，并同时保留了全部视角信息。

在 VR 系统中，全景内容是核心，当前巨大的应用需求，迫切要求更加强大的 VR 内容分析和处理能力。尽管传统的 2D 图像和视频处理技术已相对成熟，但仍难以直接应用于全景图像视频。近年来，已有相关文献对全景内容处理中的难点问题进行研究，如扭曲变形[3-4]、球面网络结构[5-6]、场景交互[7-8]等，但是未见系统全面的总结。本文针对全景内容处理中的场景分析和内容处理这 2 个重要且基础性问题展开研究与综述，讨论各类方法的优缺点，并对未来的发展趋势进行展望。

1　全景图像视频的场景分析

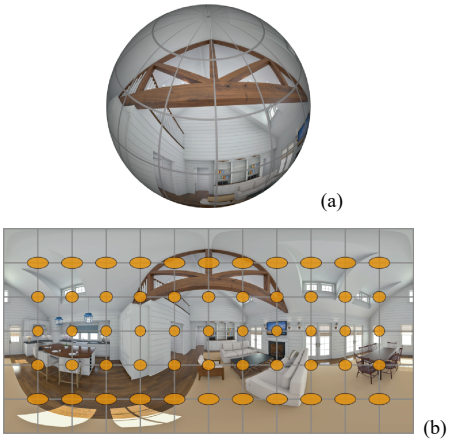
表 1 给出了全景图像视频场景分析主要研究工作的总结及代表性文献。本文首先在场景分析方面，对以下研究热点如：全景图像视频的深度学习网络、深度恢复、重要性检测和目标检测进行了总结和分析；然后，在内容处理方面，对全景图像视频的交互式浏览、去抖和校正及内容编辑进行了总结。

全景图像视频的场景分析是 VR 场景内容编辑和处理的重要基础，涉及深度学习网络、深度恢复、重要性检测、目标检测等领域。传统 2D 图像视频场景分析的理论和应用的研究已经非常深入，但仍难以直接应用于全景图像视频。如图 2 所示[9]，定义于球面的全景图能够直接映射到 2D 平面，但不可避免地会引入几何扭曲，图中椭圆的长短轴长度之比越大表示扭曲越大。此外难以表示球面上不同点之间的真实空间关系。因此全景图像视频的场景分析需要在 2D 媒体的基础上，进一步考虑高分辨率、球面几何属性、观看视角、扭曲等因素。下面将对近年来提出的代表性方法进行综述。

表 1 全景图像视频场景分析研究工作总结

Table 1 Summary of research work on scene analysis of panoramic images and videos

研究大类	研究内容	代表文献	主要特点
1 全景图像视频的场景分析	1.1 全景图像视频的深度学习网络	文献[14-15]	基于立方体投影表示
		文献[16-17]	基于球面的投影表示
		文献[18]	基于等矩形的投影表示
	1.2 场景深度恢复	文献[24-26]	单目深度恢复——基于卷积神经网络
		文献[33-34]	双目深度恢复——基于全局优化的方法
		文献[35-38]	双目深度恢复——基于深度学习的方法
	1.3 场景重要性检测	文献[45-46]	全景图像重要性检测——基于平面投影
		文献[48]	全景图像重要性检测——基于球面
		文献[51]	全景视频重要性检测——基于立方体填充
		文献[52-53]	全景视频重要性检测——基于球面卷积
		文献[55]	全景视频重要性检测——基于视点预测
	1.4 场景目标检测	文献[60-61]	基于单一投影的方法
		文献[62-63]	基于视野并交集的方法
		文献[64-65]	基于多种投影的方法



(a) (b)

图 2 360°全景图扭曲的可视化表示[9]((a) 360°全景图像；(b)具有 Tissot 指示的等矩形图像)

Fig. 2 Visualization of distortions in 360° panoramas[9] (a) Spherical 360° image; (b) The equirectangular image with Tissot's indicatrices)

1.1 全景图像视频的深度学习网络

近年来，随着全景相机的不断普及，全景图像视频的获取越来越方便，如何对其进行内容分析和高效处理是当前的研究热点。当下，深度学习已经在传统 2D 图像和视频中得到了广泛应用，包括目标检测[10]、光流分析[11]、语义分割[12]、重要性检测[13]等。可以预见，未来基于深度学习的全景图像视频处理也将成为研究热点。近年来，卷积神经网络(convolutional neural networks，CNN)由于其特征提取与表示的天然特性，广泛应用于平面图像和视频处理。然而，全景图投影后的扭曲问题极大地限制了其应用。当前基于全景图像视频的深度学习方法首先需要解决的是扭曲问题，根据全景表示的方式主要分为 3 类：立方体投影[14-15]、球面投影[16-17]、ERP[18]，如图 3 所示。

(1) 基于立方体投影表示方式。立方体投影是一种多视角的且扭曲最少的平面投影方式。现有方法主要采用平面 CNN 或结合立方体填充(cube padding，CP)方式以减少扭曲的影响。文献[14]认为在传统的平面投影下，如立方体投影，通过智能地选择视角也可以减少扭曲。并提出了一个基于循环神经网络(recurrent neural network，RNN)的自动预测立方体旋转方法，使得投影到立方体面上的结果能够更好地保证全景内容的完整性，即减少投影到立方体边缘上的前景物体。文献[15]提出了基于生成对抗网络(generative adversarial network，GAN)的 360°图像补全方法。其将输入的等矩形格式的全

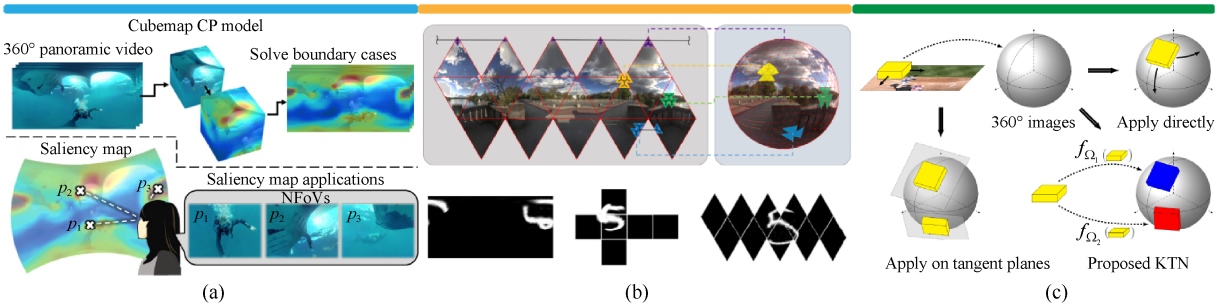


图 3 基于立方体、球面和等矩形平面表示的 CNN ((a)立方体投影表示[14]；(b)球面多面体表示[16]；(c)等矩形表示[18])

Fig. 3 CNN based on cube map, sphere and equirectangular representations ((a) Cubmap representation[14]; (b) SpherePHD representation[16]; (c) Equirectangular representation[18])

景图转换成立方体投影格式。为了构建立方体各面之间的关联性，在网络训练时，首先采用整个鉴别网络查看立方体的每个平面，以评估平面之间是否为相联的立方体；然后采用切片区分网络查看立方体映射的每个面，以确保局部一致性。

(2) 基于球面的表示方式。这类表示方式与全景图像的几何结构一致，但难以直接应用传统的平面 CNN，近年来，很多学者针对如何在球面上建立 CNN 进行了研究。文献[16]为了将 CNN 应用于全景图，提出通过球面多面体来表示所有方位的视图。该方法最小化了球面上空间分辨率的方差，并为球面多面体表示提供了新的卷积池化方法。基于二十面测地多面体的性质，提出一种几何结构进行360°图像的投影。与其他投影相比，该方法的空间分辨率和失真的变化较小。此外，该结构具有旋转对称性以及连续的特性。受图学和计算机图形学技术的启发，文献[17]提出了切线图像的球面图像表示方法。该方法将球面图像渲染成一组与细分二十面体相切的局部平面图像网格，便于可转换和扩展的计算机视觉应用。即基于指向性和较少扭曲的切线图像，提出了一种分辨率和细分层次相分离的方案，并且可以用标准网格卷积运算进行滤波。使用切线图像时，标准的 CNN 性能比专业网络更好，且可以有效地扩展到高分辨率球面数据，为透视数据和球面数据之间的保持性能的网络传输提供了可能性。此外，基于标准 CNN，能够方便地实现高度优化的卷积操作。

(3) 基于 ERP 表示。这类投影引入的扭曲较大。为了处理该问题，需要重新设计 CNN 及相关计算方法。文献[5]通过转换平面 CNN 直接在等矩形表示的全景图上学习球面 CNN。该方法能够在 360°全景图上通过学习来重现平面滤波输出，并且能够有效地处理全景图球面上的扭曲问题。该方法的主要优势在于：①能够高效地提取 360°全景图和视频中的特征；②能够有效地利用平面透视图像中已有的强大的预训练网络，从而快速增强 360°全景图的处理能力。为了解决从透视 CNN 到球面 CNN 转换过程中出现的计算代价大和准确度不高的问题，文献[18]进一步提出了通过内核转换网络(kernel transformer network，KTN)高效地将透视图像上训练的卷积核转换到 360°等矩形图像的方法。该方法在 360°图像上学习了一个新的 CNN，只需要通过 KTN 学习一个以源 CNN 内核为输入的函数，并将其转换为适用于 360°图像的 ERP。该函数很好地考虑了 360°图像中的扭曲，根据极角(Polar angle)和源 CNN 的内核返回不同的变换。经过训练的模型能够再现基于 360°图像每个切平面透视投影的源 CNN 输出，因此与源 CNN 的表现相似，能够避免图像重复投影。

以上提出的各种表示方式，在特定的场景下取得了成功应用，但仍在计算消耗、扭曲、以及适用范围等方面存在问题。此外，以上方法大多专注于如何将普通平面卷积核应用于球面，并未直接与全景图像视频中的人体视觉相结合，因此仍然具有一定的局限性。

1.2　全景图像视频的场景深度恢复

深度信息是计算机视觉领域长期以来的研究热点，有非常多的应用场景，如自动驾驶、场景三维重建等，其主要目标是理解真实的外部世界。深度恢复主要分为 2 类：单目和双目立体深度恢复。

单目深度恢复基于单个视角的图像或视频恢复场景的深度信息，这是一个病态且难以求解的问题。主要的解决方案是基于几何先验、图像特征或 CNN 来进行深度估计[19-21]。当前用于普通透视图像的单目深度恢复已经取得了较大进展[22]，然而直接应用于全景图像的性能将下降很多，原因在于全景图像视频的几何扭曲将大大影响场景结构的分析。如图 2 所示，ERP 的全景图中的几何扭曲在垂直方向上从中间向两边不断递增。WANG 等[23]提出了基于双重融合网络的全景深度估计方法，通过融合等矩形和立方体映射 2 种投影来有效地平衡几何扭曲和边界不连续性。文献[3]提出了基于立方体映射投影的全景图深度估计方法，包含双重立体映射深度估计模型和边界估计模型。现有的 CNN 适用于平面透视图像，难以直接处理包含几何扭曲的全景图[20-22]。文献[4]提出了扭曲注意的 CNN，如图 4 所示，其主要思想是根据图像扭曲模型对采样网格进行变形，从而对感受视野进行修正。JIN 等[24]根据 360°室内场景深度和几何结构之间的相互关系，提出了基于机器学习的框架以用于深度估计。具体地，将角点、边界和平面等几何结构信息，用于指导深度估计。PINTORE 等[25]采用切片的表示方式，通过深度学习框架估计室内全景的稠密深度，提出的网络模型能够直接在矩形投影下运行，并具有较高的准确度。SHEN 等[26]针对 CNN 的固定感受野在全景结构感知方面的不足，提出了全景变换器用于深度估计。如图 5 所示，通过切向块来减少全景

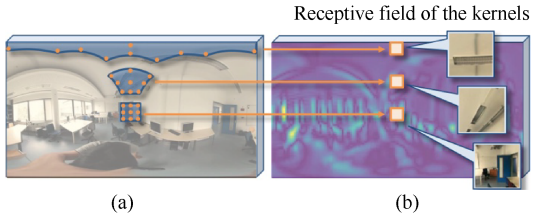


图 4 根据图像扭曲模型对采样网格进行变形，从而对感受视野进行修正[4] ((a)输入特征图；(b)输出特征图)

Fig. 4 The sampling grid is deformed according to the image distortion model, so that the receptive field is rectified[4] ((a) Input feature map; (b) Output feature map)

扭曲，利用令牌流使令牌位置更好地适配室内场景几何结构。最近，LI 等[27]为了处理全景图的扭曲并有效地提取出场景的全局背景信息，提出了一个端到端的网络结构在单位球面上估计单个全景图的全局深度信息的方法。即从 ERP 中提取的特征图投影到均匀分布网格采样的单位球面上，以大大减少特征图中的扭曲；进一步地，提出了一个基于全局交叉注意力的融合模块，用于融合来自跳过连接的特征图，并增强了获取全局上下文的能力。最终该方法在公开数据集上得到了优于之前方法的结果。文献[28-29]针对高清全景图的深度图的高效计算展开研究，重点解决传统方法难以在 GPU 上训练高分辨率全景图的问题。其将全景图分割成多个透视图，然后用传统方法计算出各个视图的深度，最后再将多个透视图的深度进行拼接得到最终结果。在视差拼接过程中，文献[28]采用了泊松(Possion)融合的方法来解决视差无缝对齐的问题，但该方法非常费时。为了提升效率，文献[29]等采用基于注册的方法高效地解决视差的全局一致性问题。

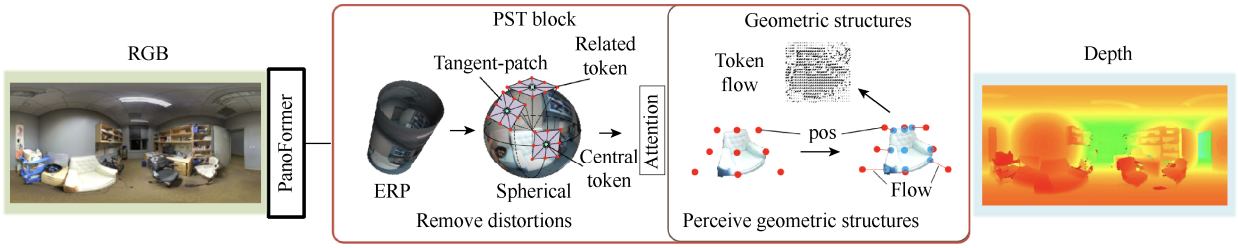


图 5 提出基于切向块的方法来消除全景扭曲，并通过令牌流更好地适配室内场景几何结构[26]

Fig. 5 The tangent-patch is proposed to remove panoramic distortions, and the token flows force the token positions to fit the structure of the sofa better[26]

以上方法存在的不足在于：其需要较多的透视投影操作，这会造成模型训练阶段的不稳定；对于基于球面的表示，尽管会减少扭曲，但不可避免地增加网络的计算开销，从而难以实用。此外，这些方法对于数据集中较少的场景如户外风景，难以取得稳定、高质量的结果。

与单目深度恢复相比，双目立体深度恢复输入的是多个视图，需要考虑多个视图之间的特征对应。经典的立体匹配算法包括局部和全局 2 种方法。全局的方法[30]能够更好地估计深度，但是一个需要求解复杂优化的问题。局部的方法，如梯度域导向滤波[31]、双边滤波[32]等方法，其计算速度快，但是匹配的准确性和稳定性不够，如在异质区域存在不明确的匹配问题。针对全景的立体匹配，LI[33]提出了以“顶部-底部”的相机配置方法来定义球面视差。KIM 和 HILTON[34]同样采用了“顶部-底部”的相机配置，通过偏微分方程的规则化方法来提升深度估计的结果。尽管这些方法直接在球面投影下进行深度估计，但仍存在不明确的匹配和稳定性不够等问题，因此当前的主流是采用基于机器学习的方法。深度学习的方法大多包含：特征提取、代价聚合、代价函数构造、视差优化步骤。ŽBONTAR 和 LECUN[35]通过训练 CNN 来预测图像块之间的匹配，并计算立体匹配的代价，最后通过代价聚合和全局匹配进行结果提升。CHEN 和 JUNG[36]提出了基于 3D CNN 和块匹配的深度估计方法，通过构造多标签分类问题对所有可能的视差值进行分类，最后通过导向滤波对深度估计结果进行提升。CHANG 和 CHEN[37]针对之前方法在“病态”区域难以找到对应关系的上下文信息的问题，提出了一个金字塔网络，主要包含空间金字塔池化(atrous spatial pyramid pooling，ASPP)和 3D CNN。当前多数立体匹配方法是基于平面透视视图的，并未考虑扭曲等问题，难以直接应用到 360°全景立体视图。为了实现全景立体场景的深度恢复，文献[38]同样采用了“顶部-底部”的相机设置，首次提出了端到

端的训练网络用于 360°立体全景的深度估计，并创建了包含立体全景图像对和真实深度 360°场景的立体数据集。网络结构如图 6 所示，主要包括：①以立体等矩形图像和极角为输入的双分支特征提取器；②用于放大感受野的巨大 ASPP 模块；③用于构建具有最佳步长的成本体积的可学习移位滤波器。最后，使用 3D 编码器-解码器来提取更深层次的上下文并回归出最终的视差图。WEGNER 等[39]首次针对 360°立体视频的深度恢复展开研究，并引入了基于圆投影的 360°立体视频模型，且构造了全景立体视频深度恢复问题，最后将现有的深度估计算法应用于全景立体视频。

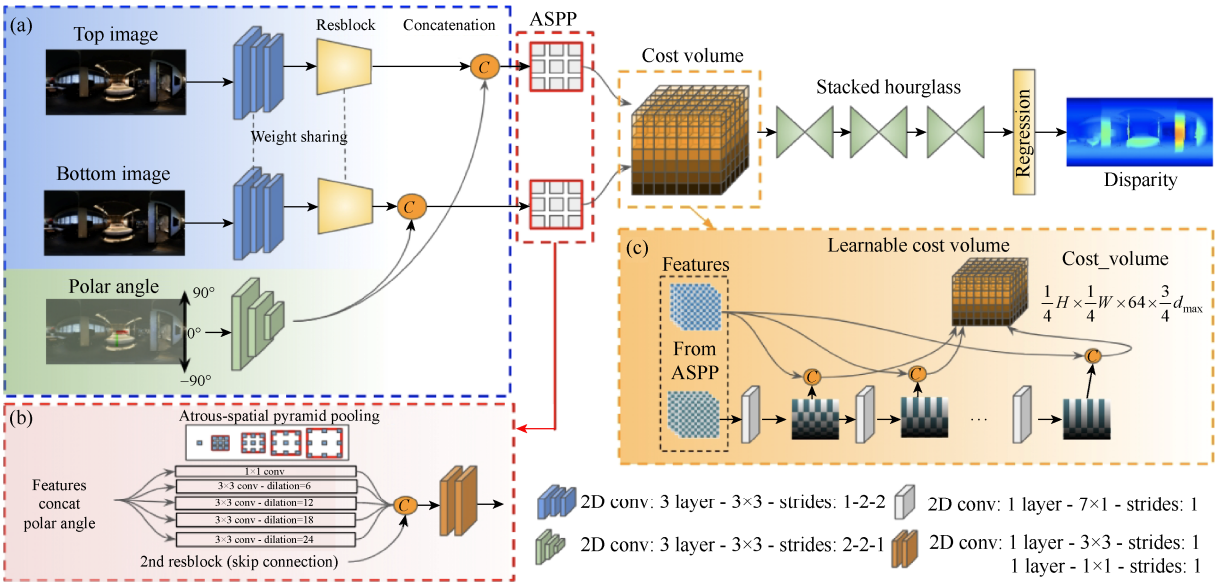


图 6　360°立体全景的深度估计网络[38] ((a)基于双分支的特征提取；(b)用于扩大感受野的 ASPP 模块；(c)用于非线性球面投影的学习代价计算)

Fig. 6　360° stereo depth estimation network[38] ((a) Two-branch feature extractor; (b) The ASPP module to enlarge the receptive field; (c) The learnable cost volume to account for the nonlinear spherical projection)

当前，直接研究 360°立体全景深度恢复的文献并不多，主要是因为存在直接将已有方法应用于 360°立体全景难以处理好扭曲问题，此外当前基于 360°的立体全景方法主要为上下视图，与人们日常观察的左右视图差异较大，应用场景和效果受限。

1.3　全景图像视频的场景重要性检测

场景重要性检测的目标是在图像和视频场景中检测出吸引人眼注意力的区域。多年来，这一课题吸引了大量认知和视觉领域的学者，因为这是场景理解和内容编辑的重要基础。BORJI 等[40]对多年来重要性区域检测的研究进展进行了深入全面的综述，包括核心概念、任务、核心方法、建模趋势、数据集、评价方法等，讨论了评价指标、数据集对建模性能的影响等开放性问题，并展望了未来的研究方向。先前的大多数重要性检测方法主要针对限制视野范围的视频和图像，然而真实世界中，人们通过主动的头部运动来全方位地感知周围环境，因此在全景图像视频场景中需要模拟这个过程来检测 360°全景图像视频中重要性区域。

当前，基于深度学习的方法，如 CNN-RNN 等，在平面图像的重要性检测方面取得了重要进展[41-44]，然而，直接用于全景图像视频的重要性检测的研究工作是比较少的。直接将 2D 透视图像的重要性检测方法应用于全景图像是有问题的，因为全景图中包含大量几何扭曲，难以将常用的 2D 图像深度学习模型应用于 360°场景。直接根据局部视角对全景图进行投影能够避免扭曲，但通常需要进行大量投影，使得计算量激增。针对 360°图像的重要性检测，MONROY 等[45]提出了端到端的 CNN。为解决数据集不足的问题，对每个全景图进行随机视角投影，生成多个无扭曲的图像块，并将每一块的重要性图作为 CNN 网络的标签。网络的特征包括：用于 2D 重要性检测的 CNN 网络特征，以及相应的球面坐标。该网络架构能够正确计算球面不同位置的重要性值。LI 等[46]为了减少 ERP 所带来的扭曲，构造了一个能够适应扭曲变形的模型来实现重要性检测。基于渐进式深度学习网络，其进一步提出了多

尺度上下文整合块，以用于感知和区分 360°方向上的丰富场景和对象。MA 等[47]针对 360°全景图的重要性检测在训练数据、特征表示、多任务处理等方法存在的问题，提出将复杂的 360°图像重要性检测问题分解成多阶段的子任务，从而只需要小尺度的训练数据。如图 7 所示，第一步，采用基于对象级语义的显著性计算方法粗略地估计出显著的视角；第二步，基于上一步估计出的显著性视角，将其投影到无扭曲的平面图像块中，并进行第二轮基于对象级语义的更加准确的显著性检测；第三步，在对象级重要性检测结果基础上进一步计算得到像素级别的计算结果。

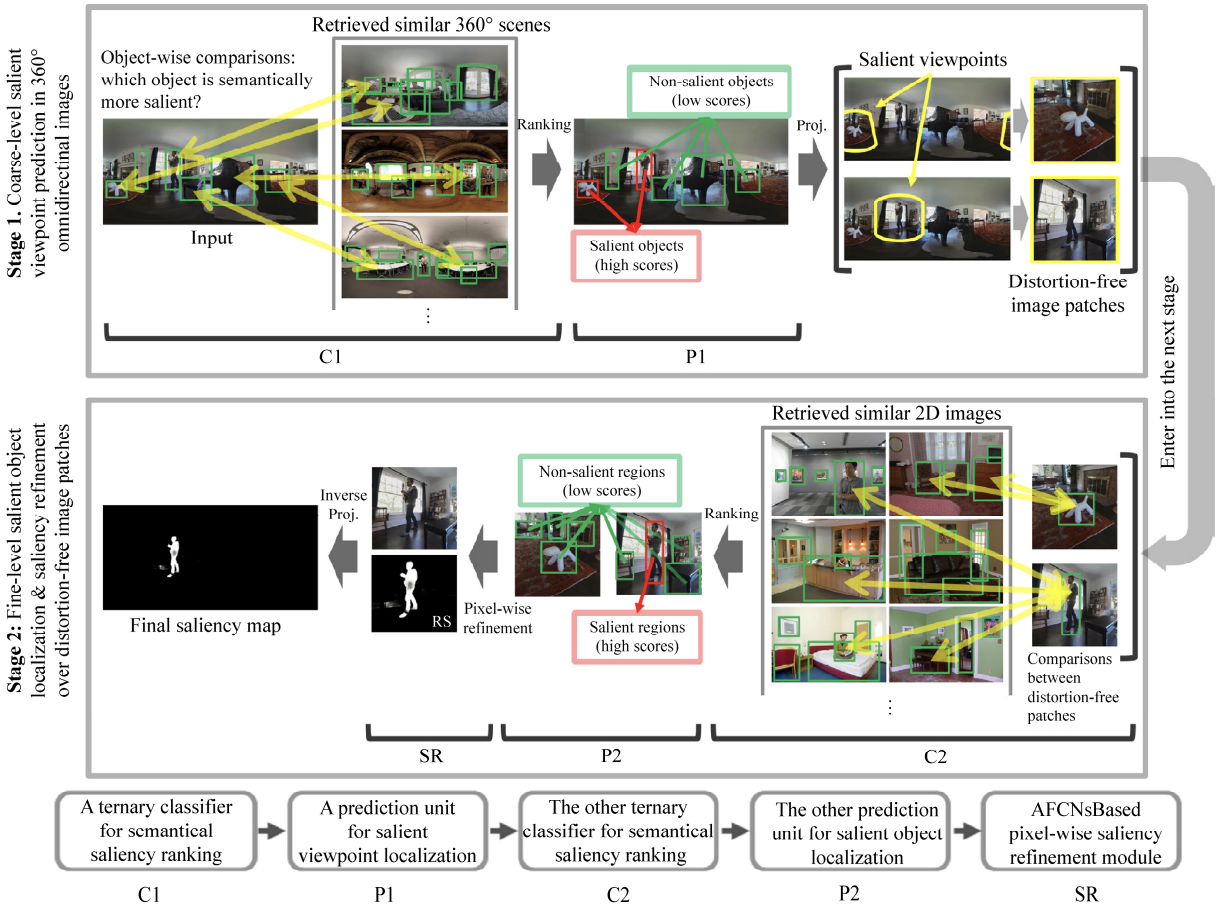


图 7 360°图像重要性检测方法流程图[47]

Fig. 7 Flowchart of saliency detection of 360° images[47]

LV 等[48]针对传统 CNN 在全景图像重要性检测方法的不足，提出了基于图的 CNN 来提取全景图像的球面特征，并生成球面图的重要性检测映射。最近，ZHANG 等[49]针对 360°全景图像重要性检测中映射卷积计算量大的问题，提出了快速的映射卷积方法，取消了卷积核访问邻接列表的过程，并基于整个邻接列表直接对全景图像只进行一次球面自适应采样，以适合 2D 卷积计算。进一步地提出了新的适应性的赤道偏移函数，可以更准确地模拟观众观看时的赤道偏向行为。与先前的方法相比，该方法具有更快的速度和更好的可扩展性。GAO 等[50]同样针对 CNN 在 360°全景图像重要性检测中计算代价大的问题展开研究，并基于人体视觉感知的过程，提出了一种新的多阶段递归生成对抗性网络。在每个阶段，预测模型以原始图像和上个阶段的输出结果为输入，然后逐步输出越来越准确的重要性检测结果。为了提高计算效率，还提出了在所有阶段共享权值参数的轻量级网络架构。

尽管以上全景图像重要性检测取得了一定成果，但仍存在以下问题：①无论是基于 ERP 还是立方体投影，投影扭曲仍无法避免；②很多方法在处理扭曲过程中，产生了太大的计算代价；③为了适应球面的几何特征，需要在卷积过程中对特征进行插值，容易造成误差的累积以及预测准确率下降。

文献[51]面向 360°视频的视角引导的应用需求，提出了将一个基于弱监督学习的时空网络架构用于重要性检测，并采集了新的 360°视频数据库用

于模型训练。为了减轻扭曲和图像边缘问题的影响，还提出了 CP 方法。该方法首先在立方体的六个面上进行透视投影，然后连接六个面，并利用立方体各个面之间的连接在卷积层、池化层、卷积长短时记忆网络层(long short-term memory，LSTM)进行填充。该方法通过 CP 方法实现了图像的无边界化且适用于传统的 CNN。ZHANG 等[52]提出了一个基于球面 CNN 的解决方案，用于 360°视频的重要性检测。考虑到 360°视频通常以 ERP 的方式储存，通过拉伸和旋转卷积核的方式实现基于球面的卷积操作。与现有的球面卷积相比，该方法具有参数共享特性，能够大大减少参数学习的数量。为了保证时间连续性，进一步提出了基于改进的球面 U-Net 的方法进行序列重要性检测。QIAO 等[53]认为 360°视频中的重要性检测受相应视点的内容和位置影响，提出了深度学习网络结构并基于视频内容和视点位置来预测 360°视频的重要性区域。并通过 2 个最近的基准数据库研究了 30 多个受试者观看 200 多个 360°视频的过程，证实了以上发现。进而提出了多任务的深度神经网络进行 360°视频重要性检测模型训练，其中每个任务对应于在类似的 360°区域中预测视口内的显著性。DU 等[54]采用球面调和方法和 GPU 流水线提出了一个简单高效的 360°视频的重要性计算和虚拟拍摄方法。其解决了以下几个关键问题：①通过基于调和方法在球面上构造重要性检测问题；②通过去除低频信息加快计算速度；③基于重要性图的自动且平滑的 360°视频浏览。最近，BERNAL-BERDUN 等[55]提出了一个新的 360°视频重要性预测模型，该模型联合了 CNN 和 RNN 对 360°视频的内在时空特征进行提取和建模。如图 8 所示，以上方法的模型采用了编码和解码的网络架构，其中编码模块由一个球面卷积长短时记忆网络(convolutional LSTM network，ConvLSTM)和一个球面最大池化层组成；解码模块由一个 ConvLSTM 和上采样层组成，用于解码特征向量，以得到最终的结果。YUN 等[56]针对 360°视频扭曲、不连续、不明确等公认的难点问题，提出了全景视觉转换器。在编码阶段，采用可变形卷积将 360°视频表示为一系列小块，并通过局部切线投影使得其几何误差最小；然后，采用视觉转换器来消除从正常视角(normal field of view，NFoV)域转移预训练权重时额外微调参数的需要。该方法与以前在每一层都执行的基于深度 CNN 的方法不同，通过单次几何近似，减轻了模型的分层几何误差累积。当前的研究在适应性和稳定性等方面仍然存在问题。未来建议关注以下方面：①将 360°全景和普通视频作为输入，以训练统一的体系结构，并利用这 2 种格式的互补性进行重要性检测；②研究基于动态相机拍摄的 360°视频的重要性检测；③综合考虑场景的深度信息。

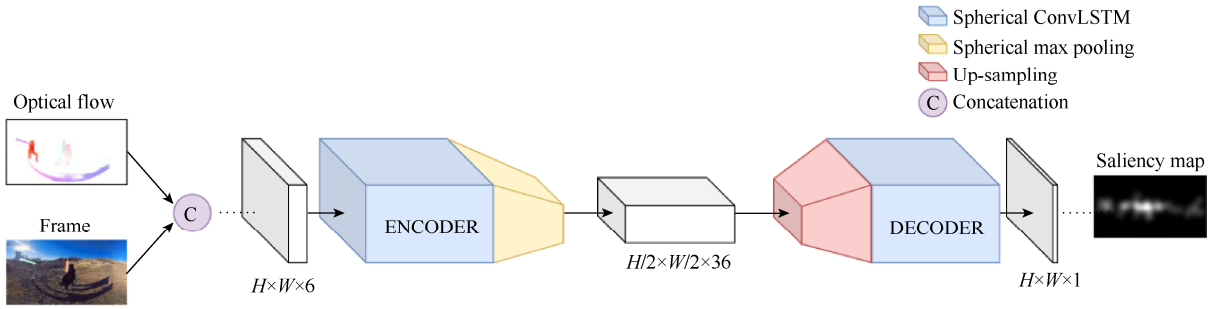


图 8　360°视频重要性预测的网络模型[55]

Fig. 8　Network for the saliency prediction of 360° images[55]

1.4　全景图像的场景目标检测

全景图像的目标检测在 VR、自动驾驶、安全监控等方面有着广泛地应用，其本质是对场景语义的理解。最直观的方法是将 360°图像投影到透视平面上，如 ERP 应用标准的平面图像目标检测方法[57]。然而该方法难以检测处于透视图像边缘附近的目标，但投影带来的几何扭曲也使得常规的目标检测方法得以奏效。

为了处理以上问题，需要在现有的目标检测方法中考虑全景图像的球面几何特征。文献[5]总结并分析了 2 种主要的将 CNN 应用于 360°图像的策略。如图 9 所示，常用的方法包括 2 类：①将球面全景图投影到平面，并将 CNN 应用于 2D 图像，然而由于球面到平面的投影引入了扭曲，会使得卷积结果不准确[58]。②将球面图像不断投影到切平面上，且应用到 CNN 中[59]。该方案通过多个切平面投影，提升了卷积准确性，但有着较高的计算开销。为了兼顾效果和效率，同时提出了学习一个 CNN 并将

平面 CNN 进行转化，从而能够以 ERP 的方式处理 360°图像。该方法能够高效地提取全景图像的特征，并可有效利用已有透视图像的强大预训练结果实现高效准确的目标检测。文献[6]提出了一个新的深度学习框架，并利用全景结构中的不变量来对抗几何扭曲，同时将其编码到 CNN 中，能够有效地将现有的 CNN 模型应用到全景图像中，在图像分类和目标检测方面有着成功地应用。YANG 等[60]提出视野包围盒(bounding FoV，BFoV)用于创建真实场景下的高分辨率等矩形全景图像数据集用于全景目标检测的评估。还提出了多立体投影方法(Multi-stereographic)减轻等矩形投影的扭曲问题，并融入了用于透视图像的 YOLO 检测器中。WANG 和 LAI[61]在引入了 360°街景数据集，并且基于快速的区域性卷积神经网络(region-based CNN，R-CNN)进行全景目标检测。为了增加训练数据，并提出了训练数据增强方法，并在模型中引入了多核的层和位置信息用于提升检测的准确率。ZHAO 等[62]提出了一个新颖的球面准则用于快速准确的 360°图像的目标检测，主要包括球面包围盒和球面并交集(intersection of union，IoU)。基于以上准则，结合 ERP 和多透视投影的优势提出两步 360°检测。最近，CAO 等[63]提出了视野并交集(field-of-view IoU，FoV-IoU)和 360°增强用于 360°图像中的目标检测。该方法能够有效地将针对透视图像的目标检测方法结合到 360°图像的目标检测中，实现在 ERP 图像中准确地进行目标检测。ZHENG 等[64]针对现有投影方法在目标检测中的不足，提出了基于“双重投影”的 360°图像目标检测网络，其由双投影特征提取器、交叉投影感兴趣区域(region of interest，RoI)搜索器以及分类和回归预测器组成。该方法结合了 ERP 的宽广视角和立方体投影扭曲少的优点，并通过 RoI 搜索器将 2 种投影方式结合于统一的框架中。CAO 等[65]为了应对 ERP 中高纬度区域的扭曲问题，提出了双重 ERP，即多视角的 ERP。该方法结合了单个 ERP 和多重投影表示的优点，能够方便地与现有的目标检测方法相结合。

虽然当前的全景图像的场景目标检测方法采用了球面卷积网络来适应全景图的几何特性，但这些方法由于计算复杂度高难以应用于高分辨率图像或层次较多的模型。如，当前最新的基于 FoV-IoU 和基于球面并交集(sphere IoU，Sph-IoU)的方法[63]仍然只是对球面多边形的近似，不可避免地会错误地评估相交的区域。未来的研究中需要不断提高近似的精度，实现更加准确的目标检测。

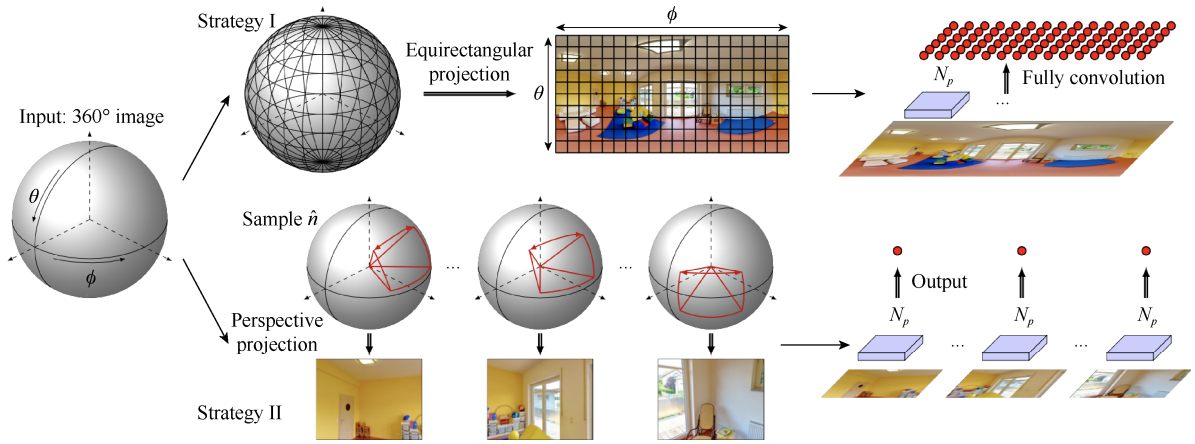


图 9 2 种主要的将 CNN 应用于 360°图像的策略分析[5]

Fig. 9 Two existing strategies for applying CNNs to 360° images[5]

2 全景图像视频的内容处理

表 2 给出了全景图像视频内容处理的主要研究工作总结，并列出了代表性文献。与普通 2D 视频和图像相比，全景内容通常覆盖了更加宽广的视野，其分辨率更高，结构也更复杂，因此拍摄困难且编辑处理的难度也更大。全景图像视频通常定义于球面，因此需要采用基于球面的度量来定义邻域、特征向量、神经网络等。直接应用传统方法中基于 2D 平面的度量和方法难以得到正确的处理结果。本文针对全景图像视频的交互式浏览、去抖和校正、内容编辑展开分析和研究。

2.1 全景视频的交互式浏览

随着全景拍摄设备的普及，360°全景视频在互

联网的各大社交平台上盛行，人们通过佩戴 360°头盔能够自由地选取不同方向的影像进行主动式浏览。然而，头盔的携带不便且价格昂贵，难以普及使用，当前大多通过鼠标拖动的方式选择不同视角进行浏览，该操作复杂且难以快速找到感兴趣的视角和内容，容易错过精彩内容。

表 2　全景图像视频内容处理研究工作总结

Table 2　Summary of research work on content processing of panoramic images and videos

研究大类	研究内容	代表文献	主要特点
2 全景图像视频的内容处理	2.1 交互式浏览	文献[72-73]	自动视频导航
		文献[75-77]	交互式和可视化浏览
	2.2 去抖和校正	文献[78]	自动全景图像校正
		文献[79-80]	全景视频去抖
		文献[81]	全景视频去抖+校正
	2.3 内容编辑	文献[83-85]	全景内容补全
		文献[87-88]	全景克隆与颜色编辑
		文献[89-90]	内容增强——全景图超分辨率

与全景浏览密切相关的是全景内容的存储、压缩、传输等技术，360°全景图像视频包含了全部视角，因此数据量通常更大，需要进行有效地压缩和解压缩以保证后续浏览的质量和效果。尽管当前的压缩和解压缩方法在传统的 2D 视频图像上已经取得了较好的效果，但将 360°全景图像视频从球面投影到平面，并进一步地存储、处理和传输，难以保持全景图像视频的球面特征。因此，当前有很多工作研究了如何高效地对 360°全景图像视频进行压缩，如 MPEG-1[66]和 JPEG360[67]等，且进一步地提出了如何进行压缩质量评估的方法。由于 360°全景图像视频的观看大多基于可穿戴设备，因此压缩存储时重点应关注不同视角下的图像质量，以及当前视角之外的场景内容的压缩[68]。一般说来，评价 360°全景视频质量和压缩时要综合考虑人体视觉感受[69]。最近，CHEN 等[70]针对 360°实时视频的高带宽、低延迟、多用户视点的要求，提出了视点注意的 360°实时视频流框架，以优化端到端的视频流传输。并根据用户的实时观看兴趣对 360°摄像头进行优先排序，并以更高的比特率上传更有吸引力的内容。同时重定义了 360°实时视频的观众体验质量指标，并用动态编程的方式解决了优化问题。为了满足用户多种不同的沉浸式体验的需求和不同级别的视图预测精度，HU 等[71]提出了具有动态传输时间间隔的双层 360°视频传输帧结构。结果表明灵活帧结构能够达到更高的用户体验质量(quality of experience，QoE)，并提升了多用户沉浸式通信的系统性能。为了提升 360°视频的观看体验，近年来，学者们对自动和交互式 2 类方法进行了大量研究。

(1) 基于自动的方法。主要通过追踪场景中重要性目标的摄像机路径来实现。为了提升 360°体育视频的观看体验，HU 等[72]提出了基于深度学习的 360°体育视频自动导航。该方法通过目标探测器得到候选的目标包围盒，再通过基于 RNN 的选择器过滤得到最佳包围盒，最后通过基于 RNN 的回归模型得到选中帧间包围盒的平滑过渡。为了更好地训练并评估算法模型，其首次构造了 360°体育视频数据集。KANG 和 CHO[73]提出了交互式和自动的 360°视频回放，该方法通过计算包含显著性区域的最优路径，并基于此为用户提供一段正常视角范围的视频内容，如图 10 所示。用户在观看视频时，能够像观看 360°视频一样，通过拖动鼠标改变观看方向，然后系统能够立刻更新观看路径，以体现用户的观看意图。该方法能够成功追踪运动目标，且无需在多目标之间来回切换。最近，为了提升用户观看 360°视频的体验并减少网络带宽的消耗，IRFAN 等[74]提出了基于深度学习的框架来预测用户视觉注意力，并能自动生成虚拟相机，为用户展示 360°视频中感兴趣的观看视角。即使用 2 种高效的 CNN 在一个联合框架下进行重要性目标检测和记忆性计算。为了更好地控制虚拟相机，利用实时目标检测(you only look once，YOLO)和 LSTM 网络进行特征提取和序列模式学习，并得到目标位置；最后基于目标位置和稠密光流来控制 360°视频中虚拟相机的视角。然而，该方法的不足之处在于，难以应对场景发生变化以及目标处于虚拟相机边缘的情况；此外由于需要稠密流，难以处理突发性运动和遮挡等情况。

(2) 基于交互的方法。利用可视化技术引导用户浏览 360°全景视频。PAVEL 等[75]利用按钮驱动

的镜头重定向技术，能够帮助观看者浏览 360°视频的所有重要内容。同时提出的视角定向技术能够在每个镜头切换处对镜头进行重定向，使得最重要的内容能够出现在用户视野中。WALLGRÜN 等[76]在基于图像的教育类 VR 浏览交互过程中，综合比较了箭头、蝶形导轨、雷达制导机制这 3 种视觉引导的用户体验，实验结果表明基于箭头的引导方式是最广泛且易于让用户接受的。但仍未解决视觉缩略图和正常视野观察的问题。为了进一步提升包含复杂场景的 360°视频的导航和回放体验，文献[7]提出了 Transitioning360°，能够由用户指定正常视野范围的摄像机路径数目，然后在同时关注内容重要性和路径的多样性的前提下计算出虚拟的相机路径。为了简化摄像机路径优化问题，引入了一个新的多样性能量项来度量路径之间的相互作用，并引入了一种基于动态规划的从粗到精的优化策略来计算多样化的相机路径。用户在浏览与文献[73]类似的主相机路径时，其他可选浏览视角的计算和可视化的计算已经完成，此时能够通过简单的交互在多个相机路径间实现空间注意的平滑切换。

图 10 交互式和自动的 360°视频导航[73] ((a) 360°视频；(b)由(a)转换来的正常视野的视频；(c)由(a)转换到(b)的相机路径)

Fig. 10 Interactive and automatic 360° video navigation [73] ((a) 360° video; (b) Normal field-of-view (NFoV) video converted from (a); (c) Camera path for converting (a) to (b))

尽管这些方法取得了一定成功，但还存在以下问题：①当前方法注重相机路径数量却忽视了路径质量；②相机的视野不够灵活，不支持变焦操作；③内容注意的优化过程需要更多地考虑用户的交互意图等因素。针对 360°全景视频的特性，结合实际应用，学者们提出了新颖的交互的协作方式。文献[8]在 VR 2022 会议上首次提出了基于穿戴设备交互的 360°视频的弹幕插入和展示技术。并针对 VR 头盔设计了 4 种展示方法，包括一种基于平面和 3 种基于球面的方法，以及 2 种基于插入控制的方法(选择+拖动、定位+选择)。主客观度量表明，球形显示方法由于能够为用户提供更多的互动性和参与性，更加优于平面显示方法。此外，球形滑动评论由于其更具生动性，比静态评论更受欢迎。

KUMAR 等[77]提出了 Tourgether360° (图 11)，能够让多用户以协作和共享空间的方式浏览和导航 360°旅游视频。同时可与现有的时间导航机制结合使用，实现高效的 360°视频浏览。尽管该交互方式存在着用户间的额外沟通和协调等限制，但仍受到了喜欢共同探索新的共享空间用户的欢迎。实验和用户报告表明，Tourgether360°能够为用户提供与互动社交视频游戏类似的社交共享的乐趣，且进一步揭示了未来设计师应该考虑的基于伪空间导航方法的大量新的交互挑战和机遇。

图 11 360°视频的协同浏览[77]

Fig. 11 Collaborative navigation of 360° videos[77]

2.2 全景视频的去抖和校正

全景视频的去抖和校正主要针对 360°视频的预处理，其目标是提升用户观赏的沉浸感和视觉舒适度。JUNG 等[78]针对场景朝向和图像轴心不对齐所致的视觉不舒适问题，提出了一个自动的全景图像校正方法。该方法基于亚特兰大世界假设，利用场景中的水平和垂直线段约束构造代价函数，并充分考虑了异常情况，能够高效、稳定且准确地进行 360°全景图像的垂直调整。如图 12 所示，手持相机拍摄的 360°视频通常存在抖动，且视频帧有着较大的几何扭曲，此时容易引起严重的不舒适感。KOPH[79]提出了混合的 3D-2D 方法进行 360°视频去抖。该方法采用 3D 分析方法来估计有一定间隔的关键帧之间的可靠旋转，并通过 2D 优化的方法保持关键帧间特征跟踪轨迹的平滑性。进一步提出了一个灵活的变形旋转运动模型，能够有效地减少由小尺度的平移运动、视差等引起的抖动残留。该方

法在准确性、稳定性、平滑保持和速度等方面比先前基于 2D 和 3D 的方法更优，并首次提出了 360°视频的去抖方法，为后续的 360°视频处理奠定了坚实的基础。SHEN 等[80]强调了保持 360°视频平滑性旋转的重要性，提出了将旋转和其他类型的运动分开处理，并采用不同方法进行运动平滑。同时提出将 2.5D 方法用于估计 3D 旋转且无需使用 3D 运动结构恢复方法，其去抖性能更加稳定。

(a)　(b)　(c)　(d)　(e)

图 12　360°手持相机、全景视频以及视频去抖前后的运动轨迹[79] ((a)手持 360°相机；(b) 360°×180°的全球面视频；(c)原始的运动轨迹；(d)基于旋转的去抖；(e)基于变形旋转的去抖)

Fig. 12　Hand-held 360° camera, panoramic video, and the tracked motion before and after stabilization[79] ((a) Hand-held 360° cameras; (b) Full-spherical 360°×180° video; (c) Original motion tracks; (d) Pure rotation stabilization; (e) Deformed rotation stabilization)

TANG 等[81]通过研究用户观察全景视频的行为习惯和偏好，提出了 360°联合去抖和方向校正的方法，能够通过联合优化对用户正前方的内容进行方向校正，并同时保持相机的平滑运动。在全景视频去抖方面，还提出了非线性优化的方法优于文献[78]采用的五点优化法。此外，基于运动估计的 3D 球面变形模型通过更好地处理旋转和平移，比文献[79]的方法更具可控性。当前去抖和校正的研究仍然存在一些问题：首先当多相机拍摄系统的设置发生变化或发生不精确的同步时，场景的建模方法将失效，从而难以实现有效地去抖。此外，当具有连续运动的前景物体占据了球面的较大视野时，难以实现准确的特征跟踪。未来仍将在这方面继续探索适用于 360°视频去抖和校正的有效模型和高效的优化方法，并进一步研究基于数据驱动的方法。

2.3　全景图像视频的内容编辑

随着 VR 技术的快速发展，360°全景图像视频的获取越来越容易，如何高质量地对其进行快速编辑是当前的主要挑战。近年来，全景图像视频处理方面的主要工作集中于全景内容的生成、传输和压缩[66-68,82]，内容编辑方面的工作并不多。一般说来，直接将 2D 平面图像编辑的技术应用于 360°全景图易在性能和效果上产生问题，原因在于：①360°全景图定义于球面，直接应用并定义于 2D 平面的度量方式是有问题的，易造成编辑结果的不一致和不连续；②全景图由于覆盖了 360°视野范围，因此通常包含更多的像素，编辑过程中需要消耗大量内存和计算资源。

本文涉及的全景图像视频的内容编辑，包括交互式内容修正和补全、内容克隆、颜色编辑、内容增强。为了修复全景拼接过程中产生的瑕疵，GAO 和 BROWN[83]提出了一个交互式的全景图像编辑工具对全景图进行对齐和修正。该系统的主要功能包括：①基于局部操作的混合接缝的编辑工具：可将接缝编辑问题构造成基于马尔可夫随机场的分割问题，并通过图割算法高效求解；②是一种内容注意的捕捉工具，能帮助用户在图像重叠区域更好地对齐局部图像内容。用户只需在图像重叠区域将需要编辑的区域拖动到目标区域，通过基于变形的优化得以实现。ZHU 等[84]针对城市 360°街景图像的内容补全进行了研究，考虑到传统 2D 图像补全方法无法处理全景扭曲的问题，于是提出了一个基于优化的投影方法，能够产生视觉满意的补全结果。但也存在一些不足，如，难以处理没有明显直线结构的全景图，以及那些光照不一致的情况。SHANG 等[85]针对全景图在可穿戴设备浏览时的视角注意的特性，提出了将视角注意的 GAN 用于全景图像的内容补全。如图 13 所示，首先引入一个纬度自适应特征融合模块，以引导生成器自适应地捕获和生成判别特征。该模块采用双层结构设计，融合了 ERP 图像中的纬度级特征和失真较小的块级(Patch-level)视域特征。然后，进一步提出了一个跨域鉴别器，以促使网络在 ERP 之外的视角生成理想的结果。

XU 等[86]进一步研究了时空一致的全景视频内

容补全问题，提出了从粗到精的优化方法通过颜色和运动信息的补全能够准确地恢复全景视频中遮挡的背景区域。如图 14 所示，运动信息通过不断补全和迭代能够保证遮挡区域的时空一致性，通过不断地传播来自相邻帧的像素颜色达到帧间的连续性。

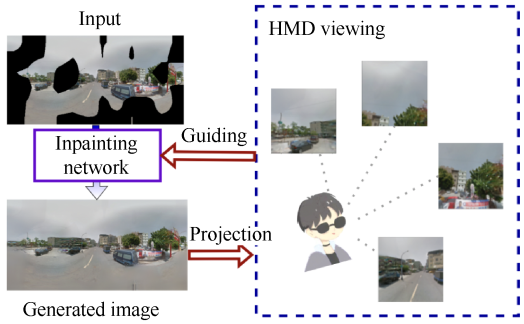


图 13 ERP 视角注意的全景图像补全[85]

Fig. 13 Viewport-aware inpainting based on the ERP image[85]

ZHAO 等[87]首次提出解决基于球面的全景图内容克隆问题。考虑到全景图的球面几何特征，并提出了一个基于球面坐标的方法实现梯度域全景图像克隆。还通过两步旋转估计的方法有效地保持球面上克隆图像块的朝向，利用基于分裂的方法来处理超过 180°视野范围的图像块克隆，并有效地解决了颜色瑕疵问题。文献[9]首次提出了基于笔画交互的 360°全景图编辑扩散方法，并构造了球面流形保持的能量优化，能够高质量地完成用户编辑在球面的扩散，为了提高计算效率，进一步提出了多分辨率的方案在保持球面流形的前提下实现高效扩散。为了进一步提高效率，ZHANG 等[88]提出了基于径向基函数(radial basis function，RBF)插值的方法，更加高效地实现了球面编辑扩散。同时在 RBF 插值过程中创新地提出了适应性采样方法，能够根据局部颜色分布决定采样数量，从而大大减少了计算量，算法流程如图 15 所示。

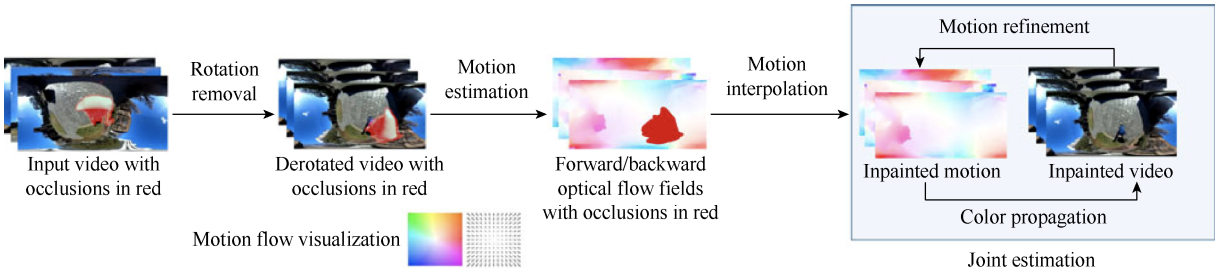


图 14 360°视频内容补全流程图[86]

Fig. 14 Flowchart of 360° video completion[86]

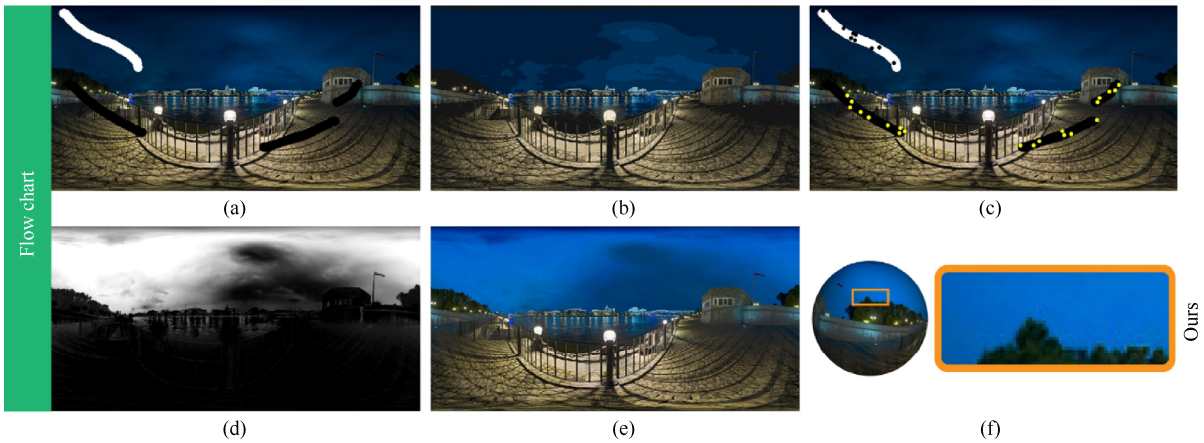


图 15 基于 RBF 插值的 360°全景图编辑扩散流程[88] ((a)输入图像；(b)颜色量化；(c)适应性采样；(d)径向基函数插值；(e)结果渲染；(f)球面和放大视图)

Fig. 15 360° panorama edit propagation based on the RBF interpolation[88] ((a) Input image; (b) Color quantization; (c) Adaptive sampling; (d) RBF interpolation; (e) Result rendering; (f) Spherical & Zoom-in views)

近年来，基于深度学习的方法开始应用于全景图像视频的内容增强。ZHANG 等[89]基于深度神经网络实现了全景图像的增强。其首先构造了第一个真实场景的全景图像数据集，然后基于 GAN 设计了一个采用多频率结构的紧凑型网络，该网络具有压缩的残余密集块中的残余(residual in residual

dense blocks，RRDB)以及来自每个密集块的卷积层。该方法能够高效地实现超高清全景图像增强，并能获得视觉和量化指标俱佳的增强结果。LIU等[90]首先提出了基于深度学习的 360°全景视频的超分辨率(super resolution，SR)方法。该方法基于单帧和多帧的联合网络架构，采用可变形卷积网络消除目标帧及其相邻帧的特征图之间的运动差异，从而充分利用了像素级帧间一致性。同时设计了一种混合注意力机制来增强特征表示能力，应用双重学习策略来约束解的空间，并进一步提出了一种新的基于加权均方误差的损失函数，用于强调全景图赤道附近的超分辨率区域，以便实现更好的超分效果。

最近，YOON 等[91]针对 ERP SR 方法存在的不足，提出了基于球面的 SR 框架，能够从低分辨率的 360°图像生成基于连续球面表征的高分辨率360°图像。算法流程如图 16 所示，输入为二十面体表示的低分辨率全景图，首先对二十面体表示的球面进行特征提取；然后，提出球形局部隐式图像函数(spherical local implicit image function，SLIIF)，通过提取的特征预测 RGB 颜色值，以便进一步用任意投影类型灵活地重建高分辨率图像；最后，利用 SLIIF 的优势，提出了基于其他投影特征的损失函数，如 ERP、鱼眼、透视投影等。该方法的最大优势在于，能够基于任意投影类型和超分尺度因子灵活地重建出高分辨率的 360°图像。当前，360°视频图像内容编辑的相关工作并不多，但可以看出基于数据驱动的方法是未来研究趋势，如基于 GAN 的内容合成与补全、基于深度神经网络(deep neural network，DNN)的编辑等。

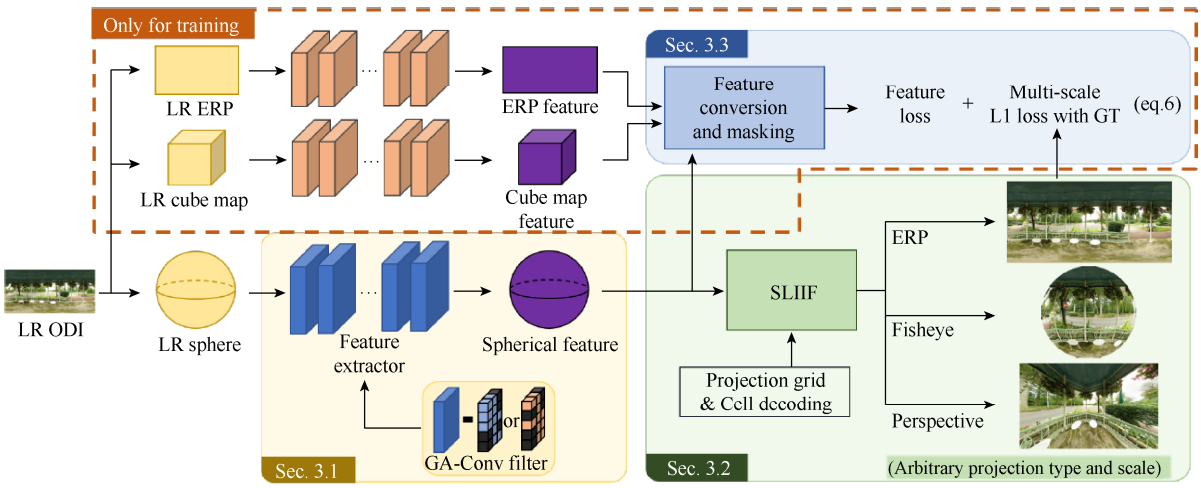


图 16　SphereSR 的总体框架[91]

Fig. 16　Overall framework of the proposed SphereSR[91]

3　结论

3.1　研究总结

360°全景图像视频的场景分析和内容处理是 VR 领域的重要研究方向，未来有着非常广阔地应用前景。本文在场景分析方面，首先回顾了深度学习在场景分析中的作用，介绍了典型的网络结构，并分析了深度恢复、重要性检测、目标检测等方面的研究工作；在内容处理方面，总结和分析了全景图像视频的交互式浏览、去抖和校正、内容编辑等方面的研究进展。进一步地分析了现有工作存在的不足、面临的主要难点以及未来可能的研究方向。与 2D 图像视频相比，360°全景图像视频在场景特征、几何结构、内容组织、数据量级等方面有明显不同，因此在场景分析和内容编辑方面有着更多挑战和难点问题。近年来，以深度学习为代表的 AI 技术的快速发展，有力推动了全景内容处理技术的进步与发展[92]，能够为全景图像视频的场景分析和内容处理中扭曲、特征表示等问题提供新的思路和技术框架，未来也将成为 VR 领域诸多应用的主流方法。

受限于本文作者的专业能力和学术视野，本文也存在一定不足之处，如综述的广度和深度有待加强，内容组织有待优化。未来将深入研究全景图像视频的处理算法，解决 VR 应用中的难点问题；并进一步对全景内容分析与处理的研究工作进行更加全面的分析和深入的思考。

3.2　未来的研究方向

与单目 360°全景图像视频相比，双目立体全景

能够提供场景深度信息，从而大大提升场景的真实感和沉浸式体验。未来的研究可以更加专注于解决立体全景图像视频的场景分析、内容处理方面存在的问题和挑战如下：

3.2.1 立体全景图像视频的场景分析

(1) 立体全景的深度估计。研究全视域结构保持的立体全景深度估计。解决球面立体全景中由于多视角拼接引起的深度错位、投影表示中的几何形变等对深度信息估计所带来的挑战，并研究如何在深度估计中保持全局几何结构和深度感知的一致性，从而实现全景立体深度信息的提取和表达。建议的解决方案为：使用左右视角全景立体图像进行立体深度估计的端到端可训练深度神经网络，解决多视角拼接的深度错位、几何失真等因素对深度估计的影响。

(2) 立体全景的场景流估计。研究基于运动层次分析的立体全景场景流估计。通过对深度分布和表观特征进行分析与建模，研究立体全景中的场景流估计，并且通过分析动态立体全景中的运动层次关系，实现立体全景中不同运动模式中的时空连续的场景流表达。建议的解决方案为：采用基于多视角投影融合的深度神经网络，利用自注意力和互注意力机制构造表观和深度相似性理解机制，建立端到端的场景流估计网络，实现基于运动层次分析的立体全景场景流估计。

3.2.2 立体全景图像视频内容处理

(1) 立体全景视频拼接和去抖。研究运动平滑和深度一致协调的立体全景视频去抖。基于球面进行视频关键帧特征的帧间跟踪以及左右视图间的稀疏特征对应、根据以上特征匹配结果及视差约束，计算视频帧间的相机相对运动；最后通过全局优化计算出立体全景视频运动平滑和深度一致协调的去抖结果。这里的主要难点在于全景扭曲的处理，建议的解决方案为：在 ERP 的视图下，对左右视图中的视频帧进行稀疏的特征匹配，通过基于单应性的随机抽样一致算法(random sample consensus，RANSAC)排除异常匹配，并将过滤后的特征点投影回单位球面；然后对左视图的全景视频进行单独去抖；最后，根据左视图去抖后的结果，对目标视差进行优化，并根据目标视差和运动平滑的约束，通过优化实现右视图的去抖。

(2) 交互式立体全景图像视频编辑。沉浸式交互的全景图像视频编辑是指用户在沉浸式环境中通过笔划、拖拽等简单直观的交互实现全景媒体内容编辑，以满足用户的个性化需求。

3.2.3 场景协调一致性全景内容交互式编辑

通过简单笔画交互进行场景亮度、颜色等特征的高效编辑，以实现源、目标场景的协调一致，以及左右视图的深度一致协调。为保证实时反馈，拟研究基于深度神经网络的方法实现高效计算。建议的方案是：将编辑扩散看成多标签分类问题，然后设计一个端到端的深度神经网络来解决问题。为达到该目标，从输入立体图像中采集多层次颜色和深度特征信息，并根据用户交互笔画提取空间信息，然后将颜色特征信息和空间信息作为深度神经网络的输入，构建全卷积网络结构，并同时加入左右视图的一致性编辑约束。以端到端的方式训练 DNN，以估计与交互笔画适应的概率图，并通过条件随机场(conditional random field，CRF)优化、细化概率图，以确保交互式编辑的像素级分类更准确。

3.2.4 基于深度学习的全景内容合成

与正常视野的图像和视频相比，360°全景图像和视频更难获取与合成，未来可以利用 GAN 在图像生成方面的优势[93-94]，进一步将其应用于 360°全景图像视频的合成。此时需要建立全景图像和视频的数据库，并进行语义标注。此外，需要考虑建立基于球面的 CNN 用于特征表示和模型训练。以立体全景图合成为例，其主要挑战在于如何找到一个合适的双目相机模型将待合成对象投影到立体全景图像中，并同时保持正确的深度感知。建议的方案为：首先估计前景对象和目标场景附近的视差信息，得到待合成对象的稀疏点云；然后进一步根据用户视角上点的方向对前景对象的点云进行分割；最后将每个分割片断与一个虚拟相机对关联，再进行片断的提升和融合。

参考文献 (References)

[1] HTC. VIVE XR Elite[EB/OL]. (2022-12-10]) [2023-01-02]. https://www.vive.com/uk/product/vive-xr-elite/overview/.

[2] Meta. Quest Pro[EB/OL]. (2022-11-10) [2023-01-02]. https://www.meta.com/gb/quest/quest-pro/.

[3] SHEN Z J, LIN C Y, NIE L, et al. Distortion-tolerant monocular depth estimation on omnidirectional images using dual-cubemap[C]//2021 IEEE International Conference on Multimedia and Expo. New York: IEEE Press, 2021: 1-6.

[4] TATENO K, NAVAB N, TOMBARI F. Distortion-aware convolutional filters for dense prediction in panoramic images[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 732-750.

[5] SU Y C, GRAUMAN K. Learning spherical convolution for

fast features from 360° imagery[EB/OL]. [2022-12-02]. https://arxiv.org/abs/1708.00919.

[6] COORS B, CONDURACHE A P, GEIGER A. SphereNet: learning spherical representations for detection and classification in omnidirectional images[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 525-541.

[7] WANG M, LI Y J, ZHANG W X, et al. Transitioning360: content-aware NFoV virtual camera paths for 360° video playback[C]//2020 IEEE International Symposium on Mixed and Augmented Reality. New York: IEEE Press, 2020: 185-194.

[8] LI Y J, SHI J C, ZHANG F L, et al. Bullet comments for 360° video[C]//2022 IEEE Conference on Virtual Reality and 3D User Interfaces. New York: IEEE Press, 2022: 1-10.

[9] ZHANG Y, ZHANG F L, LAI Y K, et al. Efficient propagation of sparse edits on 360° panoramas[J]. Computers & Graphics, 2021, 96: 61-70.

[10] WANG W G, LAI Q X, FU H Z, et al. Salient object detection in the deep learning era: an In-depth survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3239-3259.

[11] BAI S J, GENG Z Y, SAVANI Y, et al. Deep equilibrium optical flow estimation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 610-620.

[12] GUO M H, LU C Z, HOU Q B, et al. SegNeXt: rethinking convolutional attention design for semantic segmentation[EB/OL]. [2022-12-02]. https://arxiv.org/abs/2209.08575.

[13] WANG W G, SHEN J B, XIE J W, et al. Revisiting video saliency prediction in the deep learning era[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 220-237.

[14] XIONG B, GRAUMAN K. Snap angle prediction for 360° panoramas[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 3-20.

[15] WOO HAN S, YOUNG SUH D. A 360-degree panoramic image inpainting network using a cube map[J]. Computers, Materials & Continua, 2020, 66(1): 213-228.

[16] LEE Y, JEONG J, YUN J, et al. SpherePHD: applying CNNs on a spherical PolyHeDron representation of 360° images[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 9173-9181.

[17] EDER M, SHVETS M, LIM J, et al. Tangent images for mitigating spherical distortion[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 12423-12431.

[18] SU Y C, GRAUMAN K. Kernel transformer networks for compact spherical convolution[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 9434-9443.

[19] BHOI A. Monocular depth estimation: a survey[EB/OL]. (2019-01-27) [2022-12-23]. https://arxiv.org/abs/1901.09402.

[20] ZHANG Y R, GONG M G, LI J Z, et al. Self-supervised monocular depth estimation with multiscale perception[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2022, 31: 3251-3266.

[21] LEI Z Y, WANG Y, LI Z J, et al. Attention based multilayer feature fusion convolutional neural network for unsupervised monocular depth estimation[J]. Neurocomputing, 2021, 423: 343-352.

[22] MING Y, MENG X Y, FAN C X, et al. Deep learning for monocular depth estimation: a review[J]. Neurocomputing, 2021, 438: 14-33.

[23] WANG F E, YEH Y H, SUN M, et al. BiFuse: monocular 360 depth estimation via Bi-projection fusion[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 459-468.

[24] JIN L, XU Y Y, ZHENG J, et al. Geometric structure based and regularized depth estimation from 360 indoor imagery[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 886-895.

[25] PINTORE G, AGUS M, ALMANSA E, et al. SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 11531-11540.

[26] SHEN Z J, LIN C Y, LIAO K, et al. PanoFormer: panorama transformer for indoor 360° depth estimation[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2022: 195-211.

[27] LI M, WANG S B, YUAN W H, et al. S^2Net: accurate panorama depth estimation on spherical surface[J]. IEEE Robotics and Automation Letters, 2023, 8(2): 1053-1060.

[28] REY-AREA M, YUAN M Z, RICHARDT C. 360MonoDepth: high-resolution 360° monocular depth estimation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 3752-3762.

[29] PENG C H, ZHANG J Y. High-resolution depth estimation for 360° panoramas through perspective and panoramic depth images registration[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE Press, 2023: 3115-3124.

[30] WANG L, YANG R G. Global stereo matching leveraged by sparse ground control points[C]//2011 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2011: 3033-3040.

[31] YUAN W M, MENG C, TONG X Y, et al. Efficient local stereo matching algorithm based on fast gradient domain guided image filtering[J]. Signal Processing: Image Communication, 2021, 95: 116280.

[32] YANG Q X. Hardware-efficient bilateral filtering for stereo matching[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 1026-1032.

[33] LI S G. Binocular spherical stereo[J]. IEEE Transactions on Intelligent Transportation Systems, 2008, 9(4): 589-600.

[34] KIM H, HILTON A. 3D scene reconstruction from multiple spherical stereo pairs[J]. International Journal of Computer Vision, 2013, 104(1): 94-116.

[35] ŽBONTAR J, LECUN Y. Computing the stereo matching cost with a convolutional neural network[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2015: 1592-1599.

[36] CHEN B L, JUNG C. Patch-based stereo matching using 3D convolutional neural networks[C]//The 25th IEEE International Conference on Image Processing. New York: IEEE Press, 2018: 3633-3637.

[37] CHANG J R, CHEN Y S. Pyramid stereo matching network[C]//2018 IEEE/CVF Conference on Computer Vision

and Pattern Recognition. New York: IEEE Press, 2018: 5410-5418.

[38] WANG N H, SOLARTE B, TSAI Y H, et al. 360SD-net: 360° stereo depth estimation with learnable cost volume[C]//2020 IEEE International Conference on Robotics and Automation. New York: IEEE Press, 2020: 582-588.

[39] WEGNER K, STANKIEWICZ O, GRAJEK T, et al. Depth estimation from stereoscopic 360-degree video[C]//The 25th IEEE International Conference on Image Processing. New York: IEEE Press, 2018: 2945-2948.

[40] BORJI A, CHENG M M, HOU Q B, et al. Salient object detection: a survey[J]. Computational Visual Media, 2019, 5(2): 117-150.

[41] WU Y H, LIU Y, ZHANG L, et al. EDN: salient object detection via extremely-downsampled network[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2022, 31: 3125-3136.

[42] 张璐. 基于深度特征融合的显著目标检测算法研究[D]. 大连: 大连理工大学, 2021.
ZHANG L. Research on salient object detection algorithms based on deep feature integration[D]. Dalian: Dalian University of Technology, 2021 (in Chinese).

[43] WANG W G, SHEN J B, XIE J W, et al. Revisiting video saliency prediction in the deep learning era[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 220-237.

[44] LIU J J, LIU Z A, PENG P, et al. Rethinking the U-shape structure for salient object detection[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2021, 30: 9030-9042.

[45] MONROY R, LUTZ S, CHALASANI T, et al. SalNet360: Saliency maps for omni-directional images with CNN[J]. Signal Processing: Image Communication, 2018, 69: 26-34.

[46] LI J, SU J M, XIA C Q, et al. Distortion-adaptive salient object detection in 360° omnidirectional images[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(1): 38-48.

[47] MA G X, LI S, CHEN C, et al. Stage-wise salient object detection in 360° omnidirectional image via object-level semantical saliency ranking[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(12): 3535-3545.

[48] LV H R, YANG Q, LI C L, et al. SalGCN: saliency prediction for 360-degree images based on spherical graph convolutional networks[C]//The 28th ACM International Conference on Multimedia. New York: ACM, 2020: 682-690.

[49] ZHANG R P, CHEN C Y, ZHANG J C, et al. 360-degree visual saliency detection based on fast-mapped convolution and adaptive equator-bias perception[J]. The Visual Computer, 2023, 39(3): 1163-1180.

[50] GAO P, CHEN X L, QUAN R, et al. MRGAN360: multi-stage recurrent generative adversarial network for 360 degree image saliency prediction[EB/OL]. [2023-01-13]. https://arxiv.org/abs/2303.08525.

[51] CHENG H T, CHAO C H, DONG J D, et al. Cube padding for weakly-supervised saliency prediction in 360° videos[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 1420-1429.

[52] ZHANG Z H, XU Y Y, YU J Y, et al. Saliency detection in 360° videos[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 504-520.

[53] QIAO M L, XU M, WANG Z L, et al. Viewport-dependent saliency prediction in 360° video[J]. IEEE Transactions on Multimedia, 2021, 23: 748-760.

[54] DU R F, VARSHNEY A. Saliency computation for virtual cinematography in 360° videos[J]. IEEE Computer Graphics and Applications, 2021, 41(4): 99-106.

[55] BERNAL-BERDUN E, MARTIN D, GUTIERREZ D, et al. SST-sal: a spherical spatio-temporal approach for saliency prediction in 360° videos[J]. Computers & Graphics, 2022, 106, C: 200-209.

[56] YUN H, LEE S H, KIM G. Panoramic vision transformer for saliency detection in 360° videos[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 422-439.

[57] ZHANG D W, HAN J W, CHENG G, et al. Weakly supervised object localization and detection: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5866-5885.

[58] LAI W S, HUANG Y J, JOSHI N, et al. Semantic-driven generation of hyperlapse from 360 degree video[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(9): 2610-2621.

[59] XIAO J X, EHINGER K A, OLIVA A, et al. Recognizing scene viewpoint using panoramic place representation[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2012: 2695-2702.

[60] YANG W, QIAN Y, KÄMÄRÄINEN J K, et al. Object detection in equirectangular panorama[C]//The 24th International Conference on Pattern Recognition. New York: IEEE Press, 2018: 2190-2195.

[61] WANG K H, LAI S H. Object detection in curved space for 360-degree camera[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE Press, 2019: 3642-3646.

[62] ZHAO P Y, YOU A S, ZHANG Y X, et al. Spherical criteria for fast and accurate 360° object detection[J]. The AAAI Conference on Artificial Intelligence, 2020, 34(7): 12959-12966.

[63] CAO M, IKEHATA S, AIZAWA K. Field-of-view IoU for object detection in 360° images[EB/OL]. [2022-12-12]. https://arxiv.org/abs/2202.03176.

[64] ZHENG Z S, LIN C Y, NIE L, et al. Bi-projection for 360° image object detection bridged by RoI searcher[J]. Journal of Visual Communication and Image Representation, 2022, 89: 103660.

[65] CAO M, IKEHATA S, AIZAWA K. Dual-erp representation for object detection in 360° images[C]//2022 IEEE International Conference on Image Processing. New York: IEEE Press, 2022: 2016-2020.

[66] WIEN M, BOYCE J M, STOCKHAMMER T, et al. Standardization status of immersive video coding[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2019, 9(1): 5-17.

[67] JPEG Requirements Subgroup. JPEG 360 metadata use cases[EB/OL]. (2018-02-02) [2023-01-02]. https://jpeg.org/downloads/jpeg360/JPEG360-use-cases.pdf.

[68] XU M, LI C, ZHANG S Y, et al. State-of-the-art in 360° video/image processing: perception, assessment and compression[J]. IEEE Journal of Selected Topics in Signal

Processing, 2020, 14(1): 5-26.

[69] GUTIÉRREZ J, DAVID E, RAI Y, et al. Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360° still images[J]. Signal Processing: Image Communication, 2018, 69: 35-42.

[70] CHEN J Y, LUO Z X, WANG Z L, et al. Live360: viewport-aware transmission optimization in live 360-degree video streaming[J]. IEEE Transactions on Broadcasting, 2023, 69(1): 85-96.

[71] HU M, WANG L F, TAN B, et al. Two-tier 360-degree video delivery control in multiuser immersive communications systems[J]. IEEE Transactions on Vehicular Technology, 2023, 72(3): 4119-4123.

[72] HU H N, LIN Y C, LIU M Y, et al. Deep 360 pilot: learning a deep agent for piloting through 360° sports videos[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 1396-1405.

[73] KANG K, CHO S. Interactive and automatic navigation for 360° video playback[J]. ACM Transactions on Graphics, 2019, 38(4): 108:1-108:11.

[74] IRFAN M, MUHAMMAD K, SAJJAD M, et al. Deepview: deep-learning-based users field of view selection in 360° videos for industrial environments[J]. IEEE Internet of Things Journal, 2023, 10(4): 2903-2912.

[75] PAVEL A, HARTMANN B, AGRAWALA M. Shot orientation controls for interactive cinematography with 360 video[C]//The 30th Annual ACM Symposium on User Interface Software and Technology. New York: ACM, 2017: 289-297.

[76] WALLGRÜN J O, BAGHER M M, SAJJADI P, et al. A comparison of visual attention guiding approaches for 360° image-based VR tours[C]//2020 IEEE Conference on Virtual Reality and 3D User Interfaces. New York: IEEE Press, 2020: 83-91.

[77] KUMAR K, PORETSKI L, LI J N, et al. Tourgether360: collaborative exploration of 360° videos using pseudo-spatial navigation[J]. Proceedings of the ACM on Human-Computer Interaction, 2022, 6(CSCW2): 1-27.

[78] JUNG J, KIM B, LEE J Y, et al. Robust upright adjustment of 360 spherical panoramas[J]. The Visual Computer, 2017, 33(6): 737-747.

[79] KOPF J. 360° video stabilization[J]. ACM Transactions on Graphics, 2016, 35(6): 1-9.

[80] SHEN L C, HUANG T K, CHEN C S, et al. A 2.5D approach to 360 panorama video stabilization[C]//2018 25th IEEE International Conference on Image Processing. New York: IEEE Press, 2018: 3184-3188.

[81] TANG C Z, WANG O, LIU F, et al. Joint stabilization and direction of 360° videos[J]. ACM Transactions on Graphics, 2019, 38(2): 1-13.

[82] CHEN Z Z, LI Y M, ZHANG Y X. Recent advances in omnidirectional video coding for virtual reality: projection and evaluation[J]. Signal Processing, 2018, 146: 66-78.

[83] GAO J H, BROWN M S. An interactive editing tool for correcting panoramas[C]//SA '12: SIGGRAPH Asia 2012 Technical Briefs. New York: ACM, 2012: 31:1-31:4.

[84] ZHU Z, MARTIN R R, HU S M. Panorama completion for street views[J]. Computational Visual Media, 2015, 1(1): 49-57.

[85] SHANG Z Y, LIU Y W, LI G Y, et al. Viewport-oriented panoramic image inpainting[C]//2022 IEEE International Conference on Image Processing. New York: IEEE Press, 2022: 3031-3035.

[86] XU B B, PATHAK S, FUJII H, et al. Spatio-temporal video completion in spherical image sequences[J]. IEEE Robotics and Automation Letters, 2017, 2(4): 2032-2039.

[87] ZHAO Q, WAN L, FENG W, et al. 360 panorama cloning on sphere[EB/OL]. [2022-12-12]. https://arxiv.org/abs/1709.01638.

[88] ZHANG Y, ZHANG F L, ZHU Z, et al. Fast edit propagation for 360 degree panoramas using function interpolation[J]. IEEE Access, 2022, 10: 43882-43894.

[89] ZHANG Y P, ZHANG H Z, LI D J, et al. Toward real-world panoramic image enhancement[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE Press, 2020: 2675-2684.

[90] LIU H Y, RUAN Z B, FANG C W, et al. A single frame and multi-frame joint network for 360-degree panorama video super-resolution[EB/OL]. [2022-12-01]. https://arxiv.org/abs/2008.10320.

[91] YOON Y, CHUNG I, WANG L, et al. SphereSR: 360° image super-resolution with arbitrary projection via continuous spherical image representation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 5667-5676.

[92] WANG M, LYU X Q, LI Y J, et al. VR content creation and exploration with deep learning: a survey[J]. Computational Visual Media, 2020, 6(1): 3-28.

[93] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 5967-5976.

[94] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 2332-2341.