UHD World Association
世界超高清视频产业联盟

三维沉浸视频技术白皮书（2024）

Three-Dimensional Immersive Video Technology White Paper

UHD World Association
世界超高清视频产业联盟

UHD Word Association
www.theuwa.com

前 言

本文件由 UWA 联盟 xxx 组织制订，并负责解释。

本文件发布日期：xxxx 年 xx 月 xx 日。

本文件由世界超高清视频产业联盟提出并归口。

本文件主要起草单位：

xxxx 公司、xxx 公司

本文件主要起草人：

xxx、xxx

免责说明：

1， 本文件免费使用，仅供参考，不对使用本文件的产品负责。

2， 本文件刷新后上传联盟官网，不另行通知。

目录

1. 三维沉浸视频概述

1.1 三维沉浸视频概念

三维沉浸视频通过一个或者多个视点拍摄，结合传统技术或深度学习方法，旨在模拟并重现真实场景的完整的视觉信息，使观众能够获得立体、真实、沉浸式的视频体验。

相对于传统的二维平面视频，三维沉浸视频具有以下特点：

从平面感知到立体感知：三维沉浸视频通过立体影像技术使观众能够感受到真实场景的三维结构。与传统的平面二维视频不同，三维沉浸视频通过立体显示设备，如虚拟现实头显、裸眼 3D 屏幕等，能够让观众感知到物体的空间位置关系及深度信息。

从单一视角到自由视角：传统视频限制了观众只能从固定的视角观看内容。三维沉浸视频致力于提供自由视角，观众通过触摸屏幕、转动头部、手势等交互方式，使观众视角能够在场景中自由移动，仿佛置身于真实世界。

从有限时空分辨率到任意时空分辨率：传统视频受时空分辨率的限制，三维沉浸视频力求提供更高的时间分辨率和空间分辨率，以更完整、精细地呈现场景。

复刻现实场景完整视觉信息：三维沉浸视频试图模拟并重现现实场景的所有视觉信息，包括颜色、光照、深度、运动等方面，通过照片级真实的渲染技术，创造更为逼真的体验。

总体而言，三维沉浸视频是一种持续发展创新的视觉技术，经历了一系列的发展阶段，最终目标是通过整合先进的技术和设计理念，使观众能够在虚拟环境中获得真实、身临其境的感受。

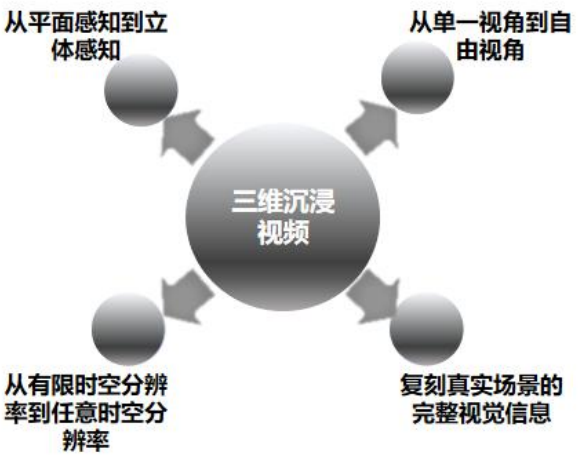


图 1 三维沉浸视频的特征

1.2 背景和意义

能够完整复刻真实场景全部视觉信息的三维沉浸视频一直存在于人们美好的想象中。相对于传统的 2D 视频，三维沉浸视频可以给予观众全新的体验。一方面，观众能够沉浸在近乎全真的虚拟环境中，获得深度的观看体验，为教育、文化、医疗和娱乐等诸多领域带来视觉上的变革。另一方面，三维沉浸视频使观众能与内容直接互动，观众从被动观看变成积极参与，拉近了数字世界与现实生活的距离，为个体提供更为个性化的体验。但是，硬件设备、软件算法、人机交互、内容创造等多方面的瓶颈都制约了三维沉浸视频的发展。

近年来，随着数字技术的高速发展，元宇宙的概念逐渐深入人心，相关研究和产品呈现爆炸性增长，为三维沉浸视频的最终实现创造了条件。一方面，图形处理单元的运算性能不断提高，能够实时渲染高质量的图像和视频；高分辨率的平面显示器、头戴显示设备、裸眼 3D 设备则能显示更加清晰、逼真和立体的图像；各种传感器设备如陀螺仪、加速度计、触摸传感器等，使得设备能够更准确地捕捉用户的动作和交互。这为三维沉浸视频的发展提供了硬件基础。另一方面，人工智能技术（AI）的浪潮，打破了传统算法所能达到效果的上限，降低了传统方法的硬件成本。例如，使用 AI 算法在改进图像的清晰度、颜色饱和度和帧率方面都超过了传统方法的效果，能够提供更真实和引人入胜的观看体验。AI 还可以用于新视点合成、三维重建等，实现了更好的实时性和更高的渲染质量，为沉浸视频的发展提供了软件算法的基础。

作为一项面向未来的技术，三维沉浸视频技术将对中国科技创新和产业升级产生巨大的推动作用，有利于强化国家战略科技力量、增强技术自主可控能力。因此，沉浸式视频的技术研究和产业应用已经写入国务院、工信部、科技部、广电总局和多个地方政府的政策性文件，《国家十四个五年规划和 2035 年远景目标纲要》提出要推进沉浸式视频应用。《“十四五”数字经济发展规划》则提出要发展互动视频、沉浸式视频、云游戏等新业态。2023 年 12 月 17 日，工信部等七部门联合印发《关于加快推进视听电子产业高质量发展的指导意见》，再次提出要加快 4K/8K 超高清、高动态范围、沉浸音视频、裸眼 3D 等技术应用。

根据《2024 中国沉浸产业发展白皮书》，到 2023 年，中国沉浸产业消费市场规模达到 927 亿元，总产值 1933.4 亿元，预计 2024 年能突破 2400 亿元。但作为一个新兴产业，三维沉浸视频的技术尚未成熟，产业正处于探索和发展阶段，大众对于三维沉浸视频的认知尚不足，亲身体验者更是寥寥。由于三维沉浸视频的技术复杂性，行业内缺乏完善的标准，硬件方面存在兼容性和互操作性问题，庞大的数据面临压缩和传输的挑战，制作和渲染未形成统一的解决方案。为此，本白皮书将梳理三维沉浸视频技术的演进路线和技术体系，展示典型的应用场景和产业需求，为三维沉浸视频技术提出标准化建议。

1.3 三维沉浸视频的技术演进路线

三维沉浸视频的技术发展不是一蹴而就的，需要经历多个阶段，不断引入新的技术和方法，以提供更为逼真、沉浸和交互的体验。如图 2 所示，三维沉浸视频的技术演进经历了以下几个阶段。

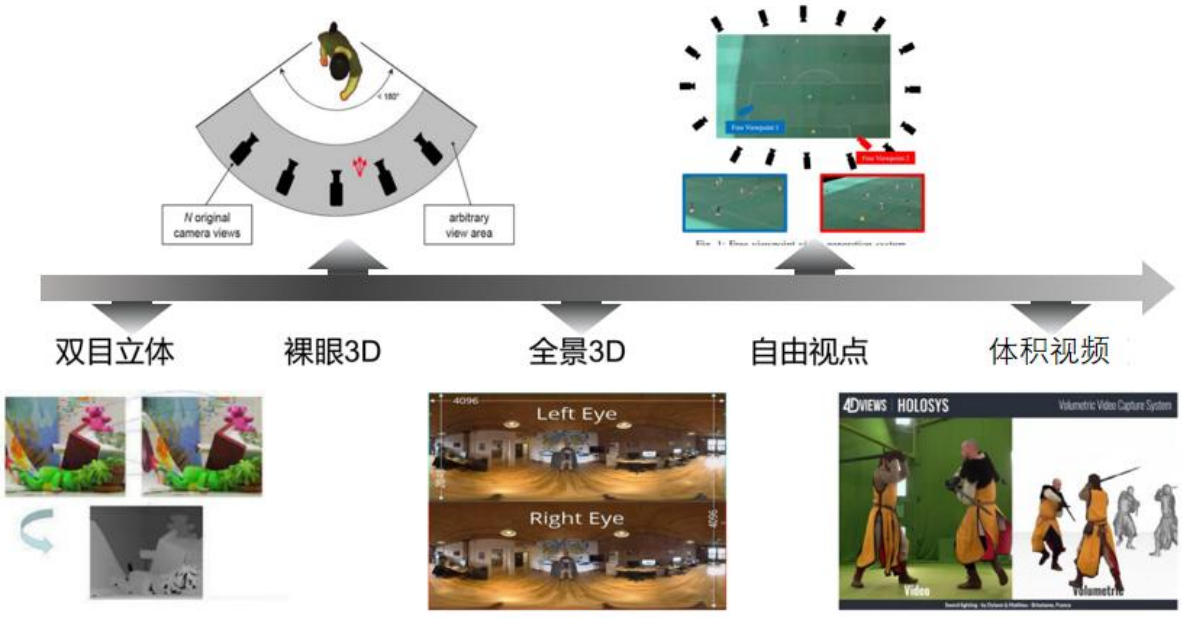


图 2 三维沉浸视频的技术演进

双目立体技术：双目立体是三维沉浸视频早期的实现方式。通过两台摄像机模拟人眼的双目视觉，使观众能够感知到深度，其缺点主要是需要佩戴特殊的眼镜或者头戴式显示设备，并且在视点数量和视觉舒适性方面存在局限性。

多视裸眼 3D 技术：裸眼 3D 技术突破了传统双目立体设备的限制。通过特殊的显示屏技术，观众在不使用额外辅助设备的情况下仍然能够感受到深度。但单视点的裸眼 3D 显示只能在屏幕前某个固定的位置上感受到 3D 图像，视点有局限性。为此，使用水平方向上多台摄像机围绕拍摄，同时生成并显示多个视点的图像，就能在屏幕前的多个位置或是一个较大的范围观看到 3D 图像。

全景 3D 技术：全景 3D 视频结合了全景视频和立体影像的技术，观众可以像身临其境一样感受到环绕式的观看体验，并且可以在不同的方向上自由转动视角。这种技术通常需要特殊的摄像设备来采集全景视频，并使用立体影像技术来处理和呈现立体效果。

自由视点技术：自由视点技术是一种允许观众从外部场景观看视频时自由选择视点的技术。它通过从多个视点采集视频或者利用计算机合成虚拟视点来实现。在观看视频时，观众可以通过交互式界面或者设备自由选择不同视角，从而获得更加个性化和沉浸式的观影体验。

体积视频（Volumetric video）技术：体积视频是一种采集和呈现三维空间中动态场景的技术。体积视频通常由一系列包含深度信息的帧组成，常见的表示形式包括点云、3D 网格等。这些三维模型在时间上连续排列，从而形成一个可以从不同视角观看的完整视频。当前，体积视频虽然能够提供六自由度（6DoF）的观看体验，但拍摄难度大，仅适用于室内小场景，渲染质量不够高，也没有形成广泛接受的标准。

2.三维沉浸视频技术体系

三维沉浸视频的技术体系包含了内容采集、三维重建、内容编码、内容传输、视频渲染、终端显示、质量评价等七个核心组成部分。内容采集阶段通过各类相机获取场景的视觉信息。然后利用三维重建技术将这些数据转化为具有立体感和几何结构的场景。接着，对数据进行压缩和编码，以便在传输和存储中减少数据量。渲染阶段将编码后的数据解码，并通过视点合成等技术渲染成沉浸式的视觉体验。最后，终端显示为用户提供了沉浸视频的观看方式。这六个部分协同作用，创造出立体、沉浸式的视觉体验。此外，质量评价可以帮助确定三维沉浸视频的整体质量。本章将对相关技术进行详细介绍。

图 3 三维沉浸视频技术体系

图 4 技术体系与技术演进关系图

2.1 三维沉浸视频内容采集

内容采集是三维沉浸视频制作的第一步，旨在捕捉场景的视觉和几何信息，为后续的三维重建和渲染提供基础数据。不同的采集方式适用于不同的场景和需求，能够提供不同范围的场景信息，从而影响到对视频的处理方式以及最终呈现效果的真实性和沉浸感。三维沉浸视频内容采集包括图像的采集和深度信息的采集，图像采集可以通过多视点的方式，使用双目相机、阵列相机或全景相机完成。而深度信息的采集既可以通过被动式采集即多目相机通过后期计算获得，也可以通过深度相机、激光扫描仪等通过物理的方式直接获得。

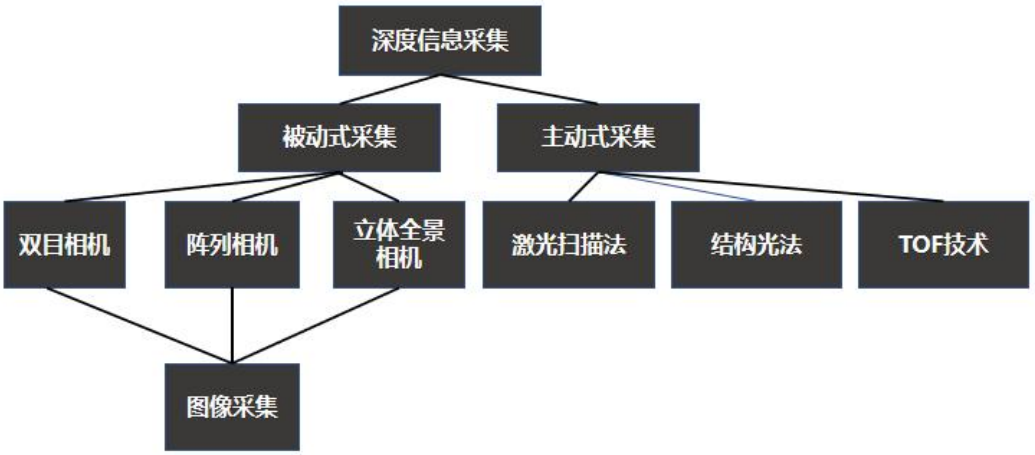


图 5 内容采集方式

1.双目相机采集

单目相机通常基于针孔模型来描述相机的成像过程。它假设相机具有一个光学中心和成像平面，光线从物体通过光学中心投影到成像平面上形成图像。焦距表示光学中心到成像平面的距离，而透视投影描述了物体在图像中的投影位置。相机参数包括焦距和成像平面尺寸等，用于校准相机并计算像素与物理空间之间的关系。单目相机的针孔模型是理解和分析相机成像的基础，它与多目相机系统共同构成了计算机视觉和摄影学中的重要工具。

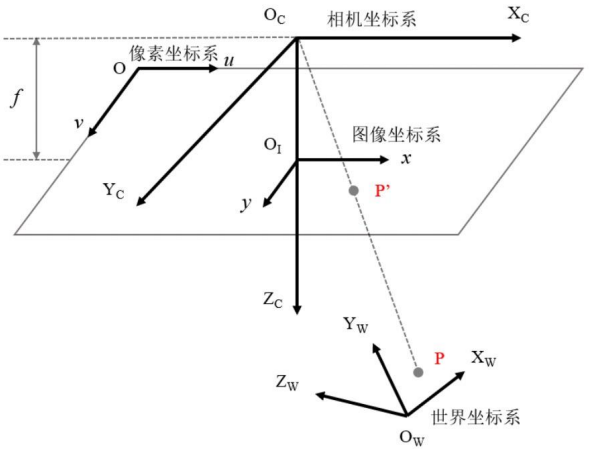


图 6 针孔相机模型

双目相机内容采集是一种利用两个单目相机同时拍摄同一场景的方法，以获取更加丰富和准确的视觉信息。在双目相机系统中，左右两个相机分别模拟人类的两只眼睛，两个镜头通常被安装在一个固定的平台上，以保证它们的空间位置和朝向一致。在内容采集过程中，双目相机需要同时获取两个摄像机的图像数据，并确保它们的时间同步和空间校准，以保证后续处理的准确性。另外，为了实现对场景深度的感知，需要通过分析两个摄像机图像之间的视差信息来计算物体到相机的距离。因此，双目相机内容采集不仅可以提供立体感觉，还能够实现对场景深度的测量和感知。下图所示为理想的双目深度相机成像模型，只需要获得同一个像素点在左右相机中的视差就可以计算出该像素点的深度信息。

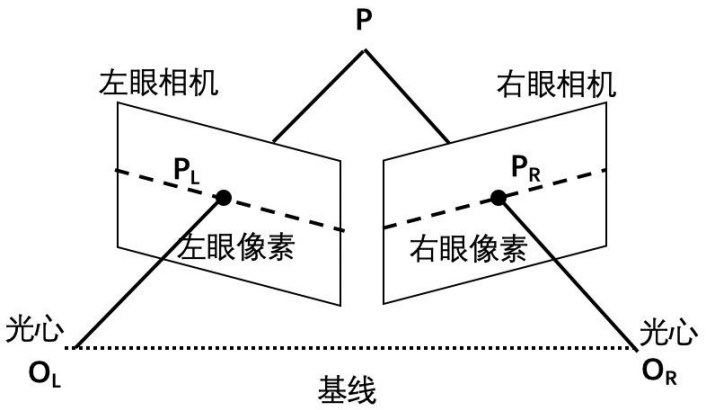


图 7 理想双目相机模型

2.阵列相机采集

阵列相机是一种使用多个摄像机排列在一起的成像系统，旨在捕捉更广阔的视野和更丰富的场景信息。如下图所示为阵列相机的几种典型几何排列方式，包括水平或垂直的线性排列、环绕排列、2 维平面式排列、3D 阵列等。阵列相机的工作原理类似于双目相机，但通过更多的摄像机增加了视点个数，便于获得整个场景的三维结构。在内容采集过程中，阵列相机需要确保所有摄像机拍摄的图像在时间上保持同步，并通过精确的空间校准来将它们的视野对齐。通常，这需要使用高精度的硬件同步和精确的摄像机标定技术。通过对多个摄像机图像进行融合和处理，阵列相机还能够合成密集的虚拟视点，或者通过多视点视图实现对真实场景的三维重建。

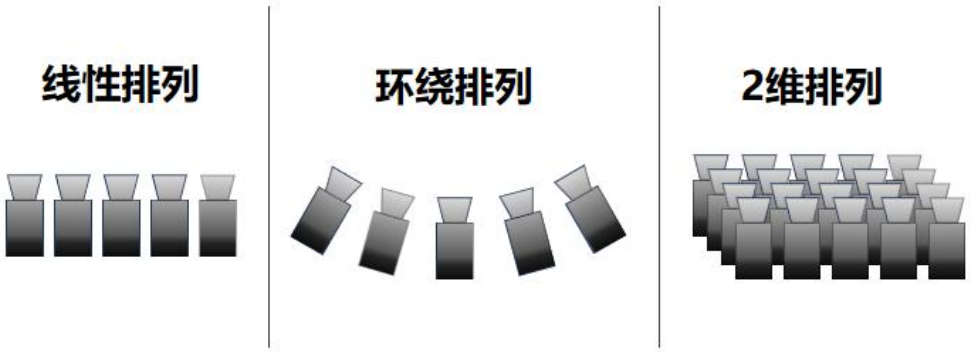


图 8 相机阵列几何排布方式

图 9 3D 阵列

3.全景相机采集

全景相机采集系统通常由多个摄像头组成，可以同时拍摄多个方向的视频，并通过软件或硬件的方式将这些视频拼接成全景视频。全景相机的镜头可以采用普通镜头，也可以使用鱼眼镜头，镜头的排列可以按环形、球形或其他几何形状的方式布局，以确保它们的视野可以实现全方位覆盖。摄像头的布局方式取决于相机系统的设计和应用需求，如使用两个 180 度鱼眼镜头组合或者使用 8 个普通镜头排成环形组合，也可以使用鱼眼镜头和普通镜头组合的方式。为了确保拍摄到的图像在时间上是同步的，多摄像头全景相机通常采用同步拍摄的方法。这可以通过硬件同步或软件同步来实现，以保证多个摄像头采集到的图像在后续拼接过程中能够对齐。通过多个摄像头同时工作，系统可以捕捉到更多的细节，并且在图像拼接后提供更高分辨率和高质量的全景图像。但它们也面临复杂的图像处理和同步控制挑战。因此，在使用这种相机系统时需要充分考虑其技术要求和应用场景。

VR 全景视频为了呈现立体效果，需要为左右两个视点分别生成全景图，这可以通过全方向立体投影（Omnidirectional Stereo Projection, ODS）模型来描述。ODS 给出了一种 3D 全景的紧凑表示方法，将空间中与一个半径为人眼瞳距的观察圆（Viewing Circle）相切的光线映射为两组（左眼光线和右眼光线）光线，对于同一个方向空间光线，它们在观察圆上的投影中心恰好落在观察圆的一条直径上。可以想象将人眼绕着中轴旋转 360°，并把每一个时刻记录下来的图片中与观察圆相切的一条光线拼接成一个完整的图像。如下图所示，ODS 对空间中所有与观察圆相切的光线进行采样，图中蓝色的光线对应于右眼观测到的光线，红色对应左眼观测到的光线。

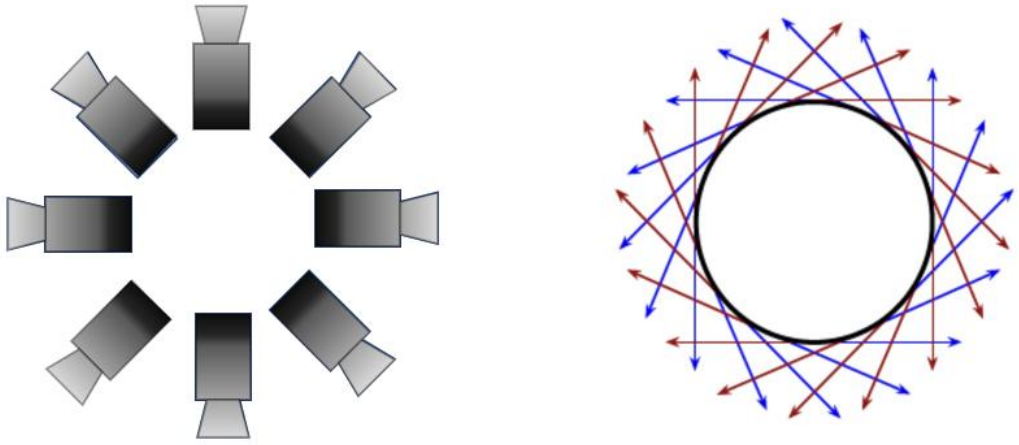


图 10 全景相机共光心环形排布、ODS 模型

4.深度信息采集

RGB-D 相机是一种同时提供彩色图像（RGB）和深度信息（D）的相机。它结合了传统 RGB 摄像头和深度传感器的功能，可以在多个像素位置上提供距离信息，从而使得获取的图像不仅具有颜色信息，还具有场景中物体距离相机的深度信息。这种深度信息通常以距离值或者点云的形式表示，能够帮助计算机视觉系统更准确地理解场景的几何结构。RGB-D 相机的使用极大地推动了许多领域的发展，使得以往依赖于复杂传感器和设备的任务变得更加简单和实用。RGB-D 相机获取深度信息的方法包括被动式方法如双目立体视觉，以及主动式方法如结构光相机、TOF 相机等。

结构光 RGB-D 相机通常采用特定波长的不可见的红外激光作为光源，发射出来的光投射在物体表面。使用相机拍摄被测物体的结构光图像，通过一定的算法获得物体的位置和深度信息。这种方式在静态场景和非透明物体具有较好的性能，适用于室内环境，但在较远距离和透明物体上的性能较差，深度测量精度可能下降。

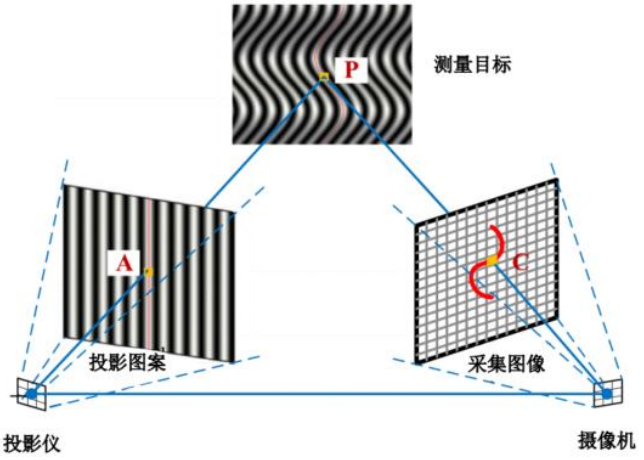


图 11 结构光 RGB-D 系统组成

TOF（Time-of-Flight）相机是一种利用飞行时间原理获取深度信息的 RGB-D 相机。它的工作原理是通过发射连续光脉冲到场景上，然后测量光线从相机发射到物体表面再返回相机的时间，从而计算出物体到相机的距离。其优点是速度快，可以实时采集深度信息，适用于动态场景。但分辨率相对较低，深度图像的精度和准确性可能较差，尤其是在长距离和低反射率表面上的性能较差。

此外，还可以通过激光扫描的方法获取物体或者场景的三维结构信息。该技术利用激光扫描设备向目标物体发射激光，并记录激光束反射回来的时间差确定距离，从而生成大量离散的点云数据。通过对这些点云数据进行处理和分析，可以实现对物体的三维重建，包括几何形状和表面细节。激光扫描法生成模型精度相对较高，被广泛应用于工程设计、文物保护、建筑测量等领域。激光扫描法的缺点是受环境影响较大、无法扫描特殊表面、时间长、成本高、应用范围有限等。

2.2 三维沉浸视频内容重建

三维沉浸视频的内容重建是通过从单个或者多个视角采集的视频或几何信息，利用计算机视觉和图像处理技术，对场景进行三维结构的重建。获取场景三维结构的方法可以分为主动式和被动式两个大类，基于主动视觉的三维重建方法需要人工设置特别的照明光源，光源信号投射到场景后，图像传感器获取返回的信号，通过比较发射前后信号的差异来计算物体的深度信息生成三维结构。这类方法适用范围比较受限，超出一定距离后误差很大，在深度图质量、图像分辨率和时间分辨率等参数上也存在一些劣势。基于被动视觉的三维重建技术不需要进行人为的增加光源，相机在自然光下采集图像，包括双目立体视觉技术、基于相机运动的三维重建（SFM）技术、多视立体视觉技术（MVS）等。对于平面视频，可以通过明暗度恢复形状法（Shape from shading，SFS)、光度立体视觉法（PS，Photometric stereo）、纹理法（SFT，Shape from texture)、轮廓法（Shape from silhouettes/contours，SFS/SFC）、调焦法（Shape from focus，SFF) 等通过图像中的特征信息进行三维重建，也可以通过深度学习的 2D 转 3D 技术实现立体效果。对于全景 3D 视频，还需要使用到全景图的拼接和合成技术。

1 三维沉浸视频表示

如下图所示，3D 图像的表示方式决定了相机设置，数据安排，发送端，接收端处理数据获取场景和信号处理的方式。另一方面，3D 图像表达方式也决定了内容合成、编码和传输方式。3D 图像表示和渲染的方法有基于几何模型表示如点云（point cloud）、体素（voxel）、网格（mesh），有基于图像的表示方式如光场合成、还有混合表示如多视点加深度图、分层深度图等，此外还有隐式表示的方式。常见的表示方式如下：

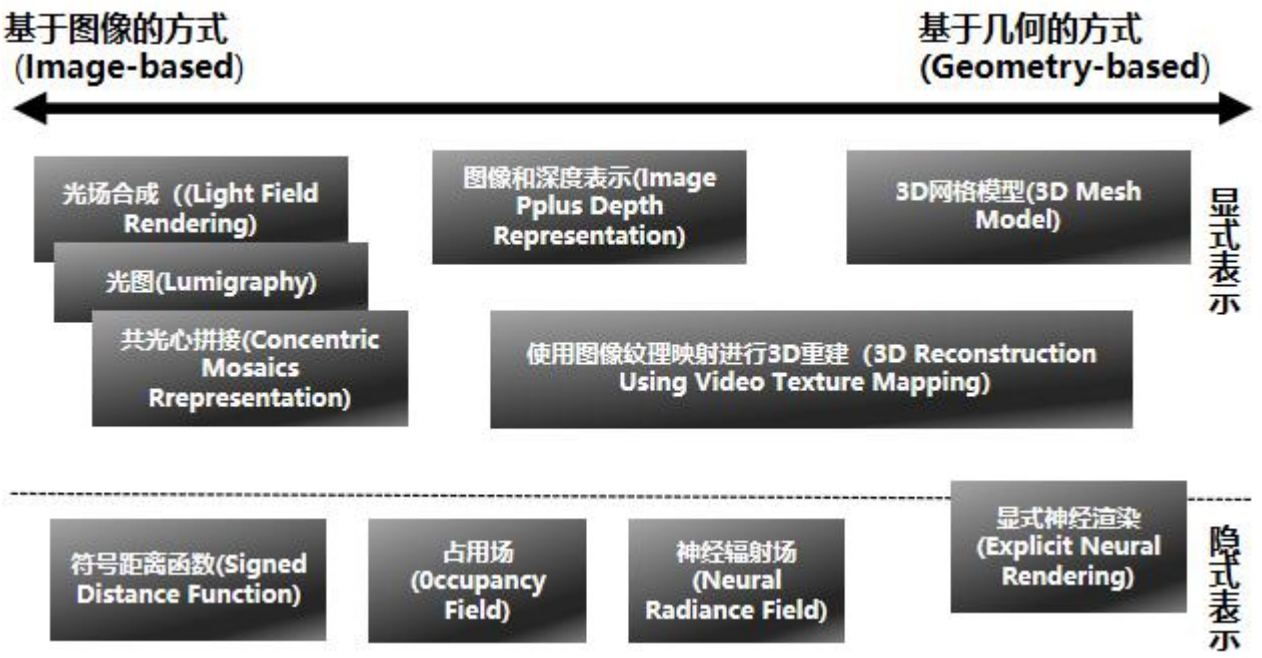


图 12 3D 图像的表示法方式

点云是由三维空间中的一组点组成的集合，每个点都有其在空间中的坐标。点云通常用于表示和模拟三维对象或场景，是计算机图形学、计算机视觉、机器人学和虚拟现实等领域中的常见数据表示形式。通过激光扫描、结构光扫描、立体视觉等方式可以获取原始点云数据，然后通过对原始点云数据进行预处理、特征提取、配准、滤波、拟合等实现对物体的建模、分析和处理。

体素是三维空间中的体积像素，是三维点云的三维等效物。与二维像素类似，体素是三维图像或三维场景的最小可分辨单元。体素通常由立方体表示，具有三维坐标以及可能的属性信息，如颜色、密度等。体素在三维重建中扮演着重要的角色，它们是一种对三维空间进行离散化表示的方式，有助于建立物体或场景的模型。

Mesh 网格是由一系列连接的顶点、边和面组成的三维结构，在三维重建中扮演着重要的角色。一些三维重建算法的输出就是一个网格，以表示被重建物体的外表面。通过从点云数据、体素网格等形式转换为 mesh，可以更容易地对重建后的对象进行可视化、分析和编辑。

以上三类表示较为精确，便于渲染和显示任意视点，但建模和匹配相对复杂，耗时大。

基于图像的表示不需要几何信息和匹配信息。这类方法包括光场合成(Light field rendering)，光图（Lumigraphy），共光心拼接（Concentric mosaics representation）等。通过多角度的图像采集，使用基于像素或者块的图像处理方式来产生虚拟的中间视点。该方法的主要优点是无需 3D 场景重建即可完成高质量的虚拟视点合成(Virtual view synthesis)。然而该优点也必须付出一些昂贵的代价: 一方面，必须通过大量的相

机对场景进行稠密的过采样，另一方面，为了合成高质量的虚拟视点，大量的图像被处理和传输。如果对场景采样过小，那么闭塞区域（Disocculusions）的插值伪影（Interpolation Artifacts）会相当明显，极有可能影响合成质量。

两者混合的表示兼顾了基于几何和基于图像表示方法的优势，主要的表示方式有多视点加对应的深度图、视差图（MVD）等，这种方式一般只需要很少的几个视点加对应的深度信息，通过基于深度图绘制的视点合成方法可以生成一定范围内的虚拟视点。由于这种表示方式合成效果较好，且相对容易获得，因此成为了三维视频重要的描述方式。但这种方式在合成视图中依然面临伪影和空洞修复的问题。

隐式表示是一种基于函数的方法，通过隐式函数来描述三维空间中的物体或场景。在隐式表示中，物体的表面不是显式地表示为点、面或体素，而是通过满足某种隐式函数的点集来定义，隐式函数不直接提供关于三维空间的几何信息，而是输出三维空间中所有几何特征满足的关系。常见的隐式表示有符号距离函数（Signed Distance Funciton，SDF），占用场（Occupancy field），神经辐射场（Neural radiance field，NeRF）等。隐式表示方法可以直接从观测到的二维图像数据中学习对象的三维结构和属性，而无需显式地提供三维信息作为监督信号。同时，隐式表示方法还能够产生高质量、高分辨率的图像，使其在图像生成、渲染和视觉重建等任务中具有广泛的应用前景。

2.2D 转 3D 技术

对于 3D 电影，两个摄像机可以模拟人眼，并同步采集视频。然而，多摄像机采集的成本和难度仍然高于单个相机拍摄，而且依靠专业的 3D 相机拍摄的视频资源仍然很少。经过几十年的积累，已经存在大量宝贵的平面视频资源，通过使用立体视觉转换技术将平面视频直接转换为 3D 视频具有重要价值。

传统方法：在深度学习之前，对于单目深度估计的研究较少，为了从单眼输入生成 3D 视频，传统方法首先利用各种深度线索如运动信息、阴影、遮挡等从原始 2D 视频中提取深度信息，生成深度图。然后对深度图进行滤波等后处理，接着使用基于深度图的渲染（DIBR）算法合成新视图，最后通过一些算法对 DIBR 产生的空洞区域进行填补。但传统方法在单目深度估计、填补场景被遮挡区域以及处理速度方面仍然不能令人满意。

深度学习方法：相对于传统的 2D 转 3D 算法，基于神经网络合成新视角的方法具有速度快，效果好的特点，可以很好地应用在现有的视频上。典型的 3D 视频合成算法，如基于神经网络的端到端算法 DEEP3D 以及分阶段的 3D 视频合成算法 3DP。

DEEP3D 设计了一个深度神经网络，将左眼的视图作为输入，内部估计一个软（概率）视差图，然后为右眼渲染一个新图像。该网络通过端到端训练，使用高质量的 3D 电影大片作为训练数据，直接从一个视图

预测另一个视图。DEEP3D 可以隐式地对遮挡区域进行图像填补，不需要后期处理。3DP 是微软 2020 年发表在 CVPR 上的一项技术。该技术根据一张图片和对应的深度图生成了这张图片的 3D 场景信息，并转换为点云，可以自由的渲染观看视角。将观看的视角固定在左右眼即可得到 3D 视频需要的图片对。

3. 双目立体视觉技术

双目立体视觉是一种基于双目相机的深度感知技术，它模拟了人类双眼的视觉系统。通过安装两个摄像头并调整它们之间的间距，双目立体视觉系统可以同时采集同一场景的两个不同视角的图像。这两个图像之间存在一定的视差，利用这个视差信息，可以计算出场景中物体的深度信息。在双目立体视觉中，使用传统方法获得左右眼相机的视差，通常使用如下图所示的几个步骤：

图 13 双目立体视觉流程

相机标定：相机的内外参数是描述相机成像过程中的重要参数，包括内部参数和外部参数。内部参数主要描述了相机的内部几何特性，如焦距、主点、畸变系数等，这些参数通常在相机制造时就固定下来，并且通常由相机制造商提供。外部参数则描述了相机与世界坐标系之间的几何关系，包括相机的位置、朝向、旋转角度等，这些参数需要通过相机标定或者视觉定位算法来获取。内外参数的准确性对于计算机视觉任务至关重要，它们在三维重建、摄像机运动估计、立体视觉等方面起着关键作用，能够影响到最终结果的精度和稳定性。

对于相机内参（如焦距、主点位置、镜头畸变），通常可以使用棋盘格标定法估计，让相机拍摄多角度的印有棋盘格的物体，通过角点检测找到棋盘格上的角点，再通过平面约束求解相机内参。对于相机外参（如相机位置、朝向），可以使用稀疏光束平差(Sparse bundle adjustment，SBA)对多相机系统进行标定，该方法假定给定多个视角下二维点对应三维坐标初始估计，以及每个相机的内参估计，利用这些信息完成一个优化问题，包括所有相机的内外参数以及三维点坐标，使得重投影误差最小。

图像矫正：在相机相对位置一致的情形下，场景点在两个相机上的投影满足极线约束 (Epipolar constrain)，即一副图像中的特征点在另一幅图像上的所有可能的对应点的轨迹构成一条二维直线，这条二维

直线称为极线（Epipolar line），通过极线约束可以极大缩小立体匹配的范围，提高立体匹配的鲁棒性和稳定性，减少计算复杂度。

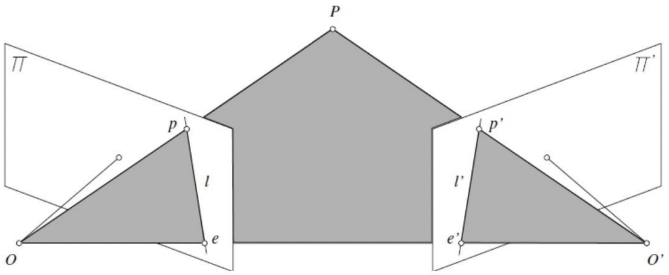


图 14 对极几何约束（P 是场景点，O 和O'分别是两个相机的光心，p 和p'分别是 P 在两个相机上的投影）

理想情况下若两相机成像面平行且对齐，参数相同，那么像素的极线就在水平方向上。但实际拍摄中，由于相机的安装误差、成像面不平行等原因，通常不满足理想条件。因此，需要通过旋转和平移相机的成像面使得它们与两个相机的基线(Baseline)平行，以实现极线矫正。

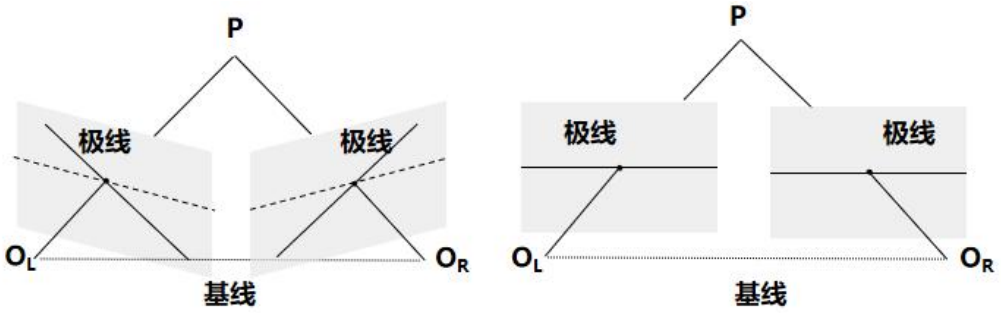


图 15 旋转相机成像面使其与基线平行

立体匹配：对于一组经过矫正的图像对，通过立体匹配来获取图像对上像素点的对应关系。立体匹配的算法有全局算法、局部算法、深度学习的方法等。全局（半全局）立体匹配算法采用全局优化理论，通过最小化全局能量函数（包含一个数据项和平滑项）得到最优视差值。常用算法有动态规划（Dynamic programming）、图割法（Graph cuts）、置信度传播方法（Belief propagation）等。

局部立体匹配的方法通过对参考图像中的每个像素周围定义一个窗口，并利用该窗口内的像素信息与目标图像中对应窗口进行匹配，进而计算像素间的视差。这种算法仅利用局部区域的约束数据进行匹配，不考虑全局一致性，具有计算复杂度低的优点，但在低纹理、重复纹理、视差不连续和遮挡等情况下，匹配效果可能较差。近年来，随着深度学习技术的发展，基于神经网络的立体匹配方法逐渐流行。这类方法利用卷积神经网络学习左右图像之间的特征表示和匹配关系，例如神经网络的特征匹配、立体神经网络（Stereo neural networks）等。

三维重建：根据匹配的结果计算图像中每个像素的视差值，然后利用视差值和相机参数进行三角测量，计算出场景中每个像素点的深度信息，从而可以得到稠密的三维空间点云。最后，对获取的三维空间点云进行预处理、表面重建 、模型优化、后处理等步骤就可以得到更为光滑和精细的三维模型。

4.多视立体视觉技术

多视立体(Multiple view stereo，MVS)视觉是双目立体视觉的推广，它通过多个摄像头或者摄像头组从不同的角度观察同一场景，以获取场景的多个视角图像。视角之间的差异性提供了丰富的深度信息，使得系统能够更准确地理解场景的三维结构和物体的位置。

多视立体视觉的方法包含体素重建法、点云扩散法、深度图融合法等：

体素重建法对计算机内存设备的要求较高，如果想要表示较大的重建场景，则体素的数量只能增多，也即用硬件换取重建场景的范围，而如果想要更精细的细节，则需要更小但分辨率更高的体素，同时这也意味着更高的硬件要求。在有限的硬件资源下，如果想要表示大场景，只能降低体素的分辨率。

点云扩散法将稀疏重建中得到的稀疏点云投影到各个视角的图像，并向投影点周围区域扩散。对于某个视角，在扩散的过程中，如果深度值与其他视角不一致或一致性较低，则视角间选出一致性最高的点作为新的深度值，这样就能逐渐重建完整的点云模型。点云扩散法优势是重建得到的点云精度较高，且在模型上的分布均匀，但是劣势是其在对弱纹理区域的处理能力较弱，容易造成空洞。

基于深度图融合的方法：对于每张纹理图估计对应的深度图，然后依次融合成点云。由于深度图的计算可用 GPU 进行加速，其在视角数量众多的场景下具有其他方法不可比的优势。此外，深度图融合的方法相比其他方法，点云密度高，这也将有助于网格生成等下游任务。目前，大部分的开源 MVS 软件以及商用 MVS 软件均采用此方法。

图 16 MVS 重建效果（来源：https://github.com/cdcseacave/openMVS）

基于深度图融合的多视立体视觉通常经过稀疏重建和稠密重建两个阶段。稀疏重建阶段可以使用基于相机运动的三维重建（Structure from Motion, SfM）技术，在未知相机姿态的情况下恢复场景的稀疏三维结构。稠密重建的主要任务是从已估计的相机姿态和稀疏三维点云出发，进一步细化和丰富场景的三维结构，构建场景的稠密三维模型。此外，随着深度学习的发展，通过深度学习方法实现多视立体视觉也成为一种有效的手段。

1）基于相机运动的三维重建

从图像中恢复出场景的三维结构是计算机视觉的基本目标。其中一种特别有效的三维重建方法使用静止场景的众多图像来进行场景重建，也就是基于相机运动的三维重建。SfM 主要分为增量式和全局式。增量式 SfM 采用逐步的方式处理图像序列，一次处理一对或一小组图像，然后逐步积累姿态信息来重建整个场景。全局式 SfM 会同时考虑所有的图像，并在整个图像集上进行优化，以最大程度地提高重建结果的准确性和稳健性。SfM 通常包括以下几个步骤：

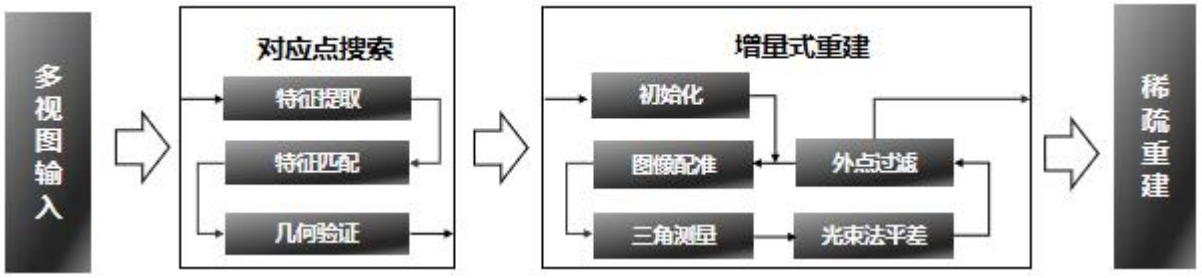


图 17 典型的 SfM 算法流程

特征点提取与匹配：特征点提取的目的是在图像中识别具有显著信息的点，这些点在视角变化、尺度变化和光照变化等情况下能保持稳定性。常用的特征点如角点、边缘或者 SIFT（尺度不变特征变换，Scale invariant feature transform）、ORB（Oriented fast and rotated brief）等局部特征点。特征点匹配的目标是找到不同图像中对应的特征点，即代表相同物理点的特征点。匹配过程中，首先计算特征点的描述子，然后使用距离度量（如欧氏距离、汉明距离等）来衡量它们的相似性，最后通过最近邻搜索、暴力匹配等策略找到匹配的特征点。

几何验证：特征匹配仅是基于特征点的外观描述，因此无法保证对应的特征点实际上映射到相同的场景点。为此需要利用图像间的几何关系，来验证这些特征匹配的正确性。常见的方法有随机采样一致性方法

(Random sample consensus，RANSAC)、8 点法、5 点法等。这些方法能够有效估算图像之间的基础矩阵和本质矩阵，丢弃错误的匹配点并准确计算相机之间的相对运动。

初始化：选择一对合适的图像来初始化模型是非常关键的，因为很有可能无法从错误的初始化中重建三维模型。此外，选择密集、信息丰富的初始图像对能够提升重建的精度和鲁棒性，而选择稀疏的初始化位置可以减少计算复杂性，但重建质量可能下降。

图像配准：增量式 SfM 重建需要在初始模型的基础上逐步加入新图像，并通过图像配准和三角测量的方式更新模型。图像配准过程从一个度量重建（metric reconstruction）的模型开始，通过解决 PnP 问题，估计新图像的相机位姿（位置和朝向）并将新图像配准到当前模型中。PnP（Perspective-n-Point）过程利用特征点的对应关系，将新图像中的特征点与已引入模型的图像的三角测量点（2D-3D 对应关系）进行匹配，得到新图像的相机位姿以及未标定相机的内参。

三角测量：如下图所示，三角测量是通过从不同视角的图像中对同一场景点进行观测，来确定该点的三维空间位置。通过这个过程，可以在三维空间中定位新点，并将其添加到现有模型中。三角测量是 SfM 的关键步骤，因为它不仅可以扩展场景模型，而且提供了多视角的冗余信息，从而增强了模型的稳定性。

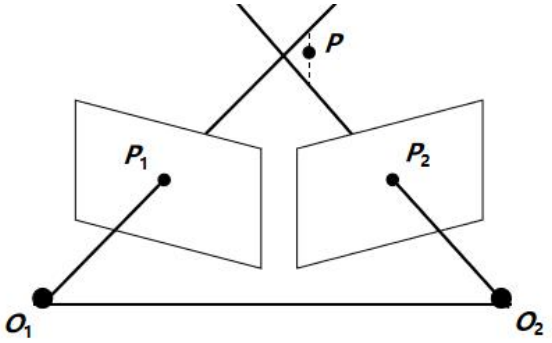


图 18 三角测量获得三维点的深度

光束法平差（Bundle adjustment）：光束法平差是一种用于优化相机位姿和三维点位置的技术。它通过最小化图像中观察到的特征点与根据当前估计的相机位姿和三维点位置计算出的重投影位置之间的误差来改善重建结果的准确性。在这个过程中，相机的位姿和三维点的位置被视为优化变量，目标是使重投影误差尽可能小。光束法平差通常使用迭代优化算法来解决这个非线性优化问题，并且通过反复迭代，不断更新相机位姿和三维点的位置，直到达到收敛条件为止。最终的优化结果可以提高三维重建的精度和稳健性。

2）稠密重建

SfM 主要关注从图像中提取相机参数和生成稀疏点云，而稠密重建则在此基础上，通过深度估计和多视图立体技术进一步细化和完善三维模型，以实现更精确和全面的场景重建，其主要技术包括：

深度估计：利用空间几何一致性约束，即空间中一个点/块区域在不同视角是颜色、形状一致的，计算获取得到每一张图片每一个像素的估计深度。MVS 的深度估计可以分为 Plane Sweep 与 PatchMatch 两类。Plane Sweep 算法可以比较容易地实现，并且在一些场景中具有较好的性能。它可以并行化地处理每条扫描线，从而提高匹配的速度。PatchMatch 算法通常需要更复杂的实现，并且在处理大规模图像时可能会变得较慢。但它通常能够产生更准确的匹配结果，并且具有更好的鲁棒性。

点云融合（Fusion)：根据上步骤获取的深度图，将二维像素点反投影到三维重建中，并进行重复点云的融合，获得一个统一的稠密点云表示。

网格化（Meshing）和纹理贴图(Texturing)：根据稠密点云，通过三角化等方法将点云结构转换成网格结构，并将纹理映射到网格模型上，最终获得一个完整的场景/物体模型。

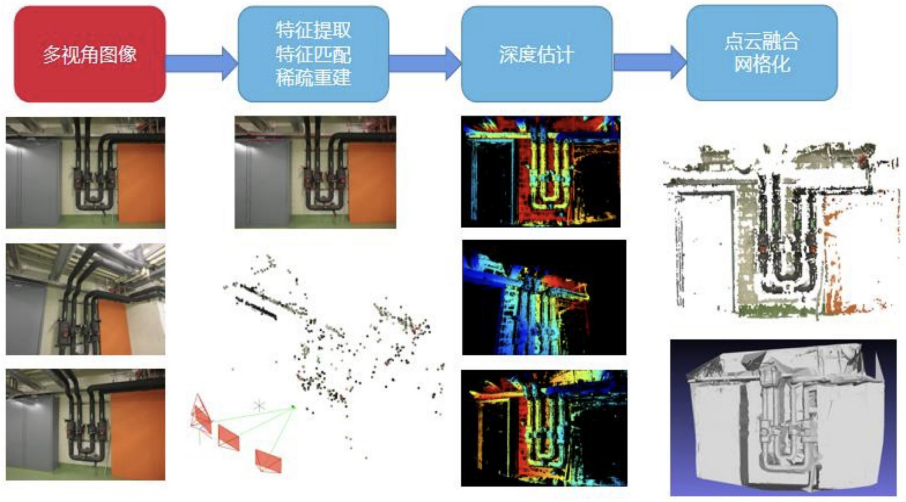


图 19 基于深度图融合的多视立体匹配流程

3）基于深度学习的多视立体视觉

自 2018 年起，多视立体视觉与深度学习结合的方法开始涌现。MVSNet 借鉴了传统方法中 Plane Sweep 的方法来构建匹配代价体(Matching Cost volume)，然后对匹配代价体使用 3D 卷积正则化代价体（可以类比于传统方法里代价聚合），得到初始深度图，最后通过 2D 卷积精细化深度图细节或者去除噪声。此后，许多方法在 MVSNet 的基础上进行了优化，如 CasMVSNet 在 MVSNet 的基础上构建了金字塔结构，从粗略到精细逐渐优化深度图；R-MVSNets 使用循环网络代替 3D 卷积优化代价体，减少深度网络的显存需求，使得深度网络也可以处理高分辨率图像等。

从 PatchMatchStereo 中获得启发，双目深度估计网络 DeepPruner 抛弃 PlaneSweep 的代价构建方式，使用可微分的 PatchMatch 方式获取深度图，在推理速度上进一步提升，而 PatchatchNet 将其拓展到多视角深度估计中。但是，有监督的多视立体视觉深度学习模型需要真实标签（Ground Truth），也即真实的深度图作为监督。在目前的技术条件下，真实深度的采集设备通常为激光，然而激光采集设备不但昂贵，而且深度图也较为稀疏。Khotetal，MvS2 与 MVS3 提出无监督的多视角立体匹配模型，解决了深度网络依赖于物理设备采集的真实深度数据的难题。

基于深度学习的多视立体视觉，无论是有监督模型还是无监督模型，与其他领域的深度学习模型一样，同样面临着场景变换情况下泛化性的问题，相比之下，传统方法则不需要训练集，这是其最大的优势。无监督的深度学习模型解决了真实标签难以获取的难题，但其效果仍然与 SOTA 有监督模型存在一定差距。无论是有监督还是无监督，都面临场景变换中泛化性问题。其次，关于深度图生成速度问题，监督与无监督模型训练耗时极大，而训练完成后模型的推理速度又极高。相对而言，传统方法生成深度图的速度仍然很慢。

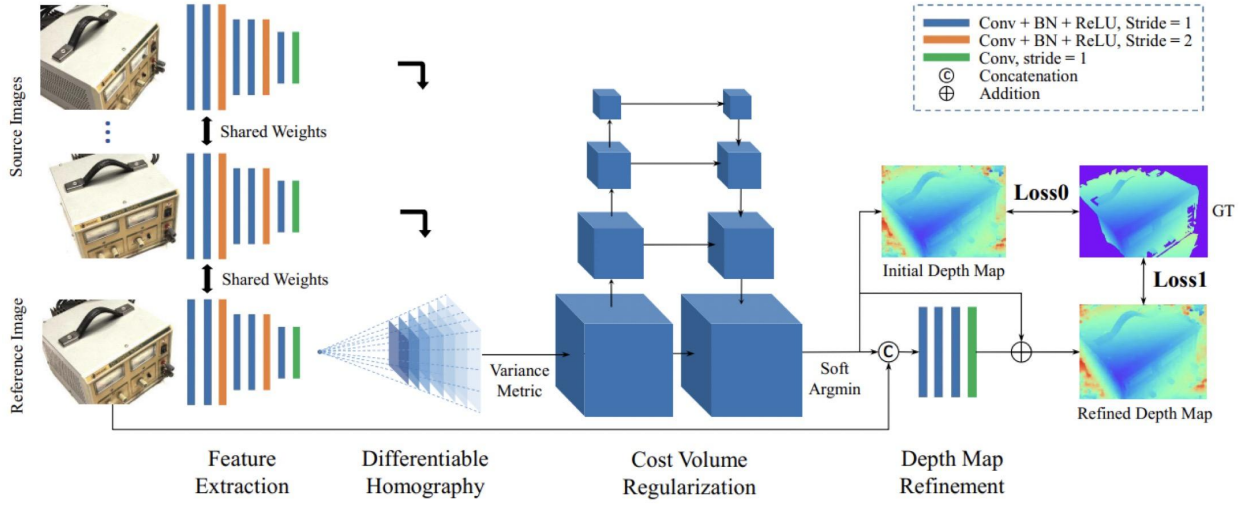


图 20 MVSNet 系统流程图

5.全景立体视频技术

全景视频展示的是某个观察点周围 360 度的全部景象，因此需要存储的信息由经过这个观察点的所有光线组成。由于经过一个点的光线有无穷多条，全景视频会对这些光线进行离散化采样，并将其中的一部分保存，再映射到某个 2 维平面进行存储。用户的有效视野在水平方向上通常约为 150 度，而戴上 VR 头显后，实际可见范围会变得更小，通常在 110 度以下。因此，全景视频数据中有一部分内容无法在用户有效视野内显示。当用户使用头显或其他观看工具时，全景图像会被反映射成 3D 光线，根据用户观看的角度重新成像，以适应头盔的限制，并在用户的眼睛中呈现出一种沉浸感。

全景立体视频技术结合了全景视频和立体视觉，可以为观众提供 360 度环绕和立体的视频感受。构建立体全景视频会比平面全景视频更加复杂，需要使用到全景视频的拼接技术以及全景立体视频合成技术。

1）全景视频的拼接技术

该技术是用来将多个成像设备在不同位置拍摄到的视频内容对齐并拼接为全景图像的方法。常用的拼接算法包括：传统的基于单映射的全景拼接技术、基于双单应矩阵的的拼接算法、基于动态直接线性变换法的拼接技术等。

基于单应矩阵的拼接方法：单应矩阵通常描述处于共同平面上的一些点在两张图像之间的变换关系。若所有相机采集到的视图共面或近似共面，或者视角变化不大时，则可以通过单应来进行相机位姿估计。这种方式适用于相机之间仅有旋转，没有平移的情况。然而在实际拍摄过程中，多个相机的成像中心并不重合，对应不同景深的图像区域带有不同的视差(Parallax)，无法正确地通过单应矩阵对齐，可能会出现拼缝或者失真等情况。

基于双单应矩阵的拼接算法：该技术使用两个单应矩阵分别拟合近景平面和远景平面，并且对这两个单应矩阵进行融合，从而更好地对齐图像。

基于动态直接线性变换法的拼接技术：当场景为平面的或者相机位姿为纯旋转，单应矩阵的拼接模型是合理的，然而实际情况中该前提很难满足，因而会产生伪影(Ghosting Artifact)。基于动态直接线性变化法的拼接技术（As-projective-as-possible with moving DLT, APAP）不再采用全局投影，而是允许局部存在相对于全局投影的偏差。APAP 基于动态直接线性变换法 （Moving direct linear transformation，Moving DLT）可以无缝地桥接与投影模型不一致的图像区域。该算法产生了高度准确的图像拼接结果，显著少了伪影现象，大大降低了算法对后处理阶段去伪影的依赖性。

2）全景立体视频合成技术

上述方式拼接出来的全景视频只能给双眼提供相同的内容，缺乏 3D 深度感。为了提供六自由度的内容，需要从有限的真实视图合成虚拟视图。这可以使用稠密的光流算法实现，待合成的连续虚拟视点不是某个空间视点位置对应的完整图像，而是分别针对左右眼视点且满足 ODS 模型的像素列，这里模拟了用户双眼观看现实世界的过程。

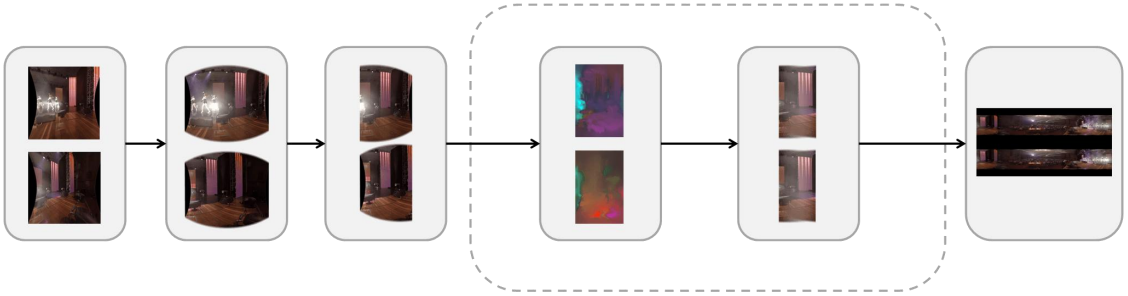


图 21 全景立体视频内容合成

典型的全景立体视频合成算法，首先从相邻相机镜头之间提取重叠区域。然后逐像素计算左右重叠区域之间的双向稠密光流信息，这里可以使用传统方法如 Lucas-Kanada 方法、LK 金字塔光流算法等，也可以使用深度学习的方法获得光流如 RAFT、GMA 等。最后，根据光流信息及 ODS 模型合成左右眼的虚拟像素列。

2.3 三维沉浸视频内容编码

三维沉浸视频内容编码涉及将三维场景的内容编码成数字数据，以便在沉浸式视频系统中传输、存储和呈现。三维视频和二维视频很大的不同在于其表示格式、编码技术和三维显示技术之间是相互关联的，不同的三维显示需要使用不同编码方案，如传统的平面视频编码、双目立体视频编码、多视点编码、全景立体视频编码、模型编码等。

1.传统平面视频编码技术

视频编码是指用于将数字视频压缩以便于存储和传输的一系列规范和算法。视频编码标准只规定了码流的语法语义和解码器，只要码流符合相应的标准语法，解码器就可以正常解码。如下图所示，从 1980 年代至今，视频编码标准的发展已经超过了 40 年。目前，由国际电信联盟（ITU-T）和国际标准化组织（ISO）/国际电工委员会（IEC）制定的 H.26x 标准，由开放媒体联盟 AOM 制定的 AVx 标准以及由我国数字音视频编解码技术标准工作组（AVS 工作组）制定的 AVS 系列标准是国际上三个主流的视频编解码标准。随着技术的不断发展和应用需求的不断变化，视频编码标准将继续发展和演进，以满足视频压缩效率、视觉质量和实时性等方面的不断提升的需求。

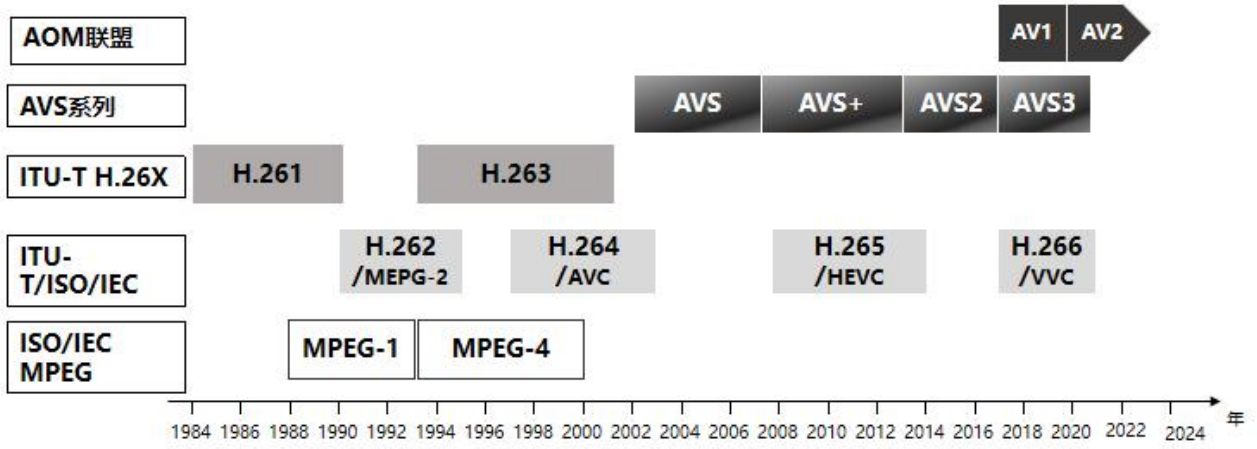


图 22 主要编解码标准发展历史

主流视频编码标准通常采用基于块的混合编码框架，以实现高效的视频压缩。如下图所示为 AVS3 标准的编码框架，可以看到，对于输入视频需要经过帧内帧间预测、变换、量化、反变换反量化、熵编码、环路滤波等步骤，最终输出编码后的码流。

当前，支持多视点立体视频编解码标准的设备及应用仍然较少，传统的平面视频编码标准在三维沉浸视频中依然发挥着重要的作用。

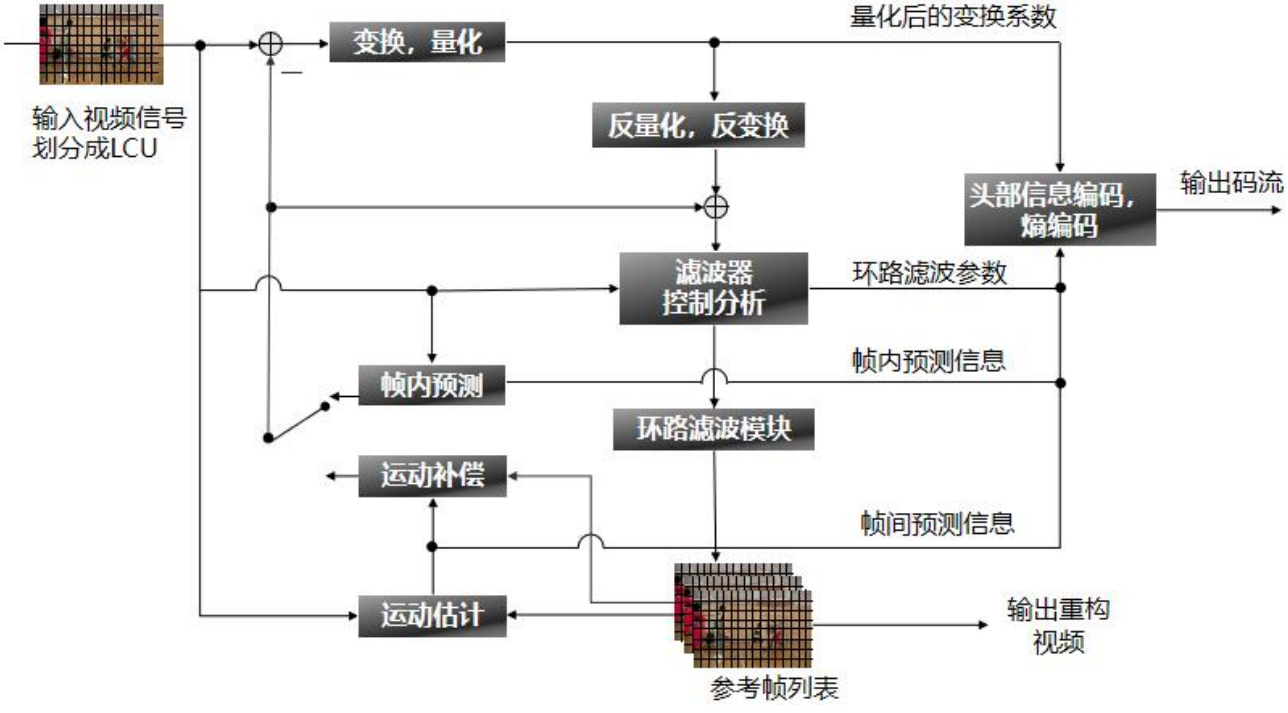


图 23 AVS3 编码框架

2.双目立体视频编码技术

双目立体视频通常以两种方式进行表示。一种方式是将左右两个视点分别作为两个视频序列进行编码和传输，两路视频可以选择任意平面视频的编码标准如 AVC、HEVC、AVS3 等。这种方式可以使用两个相机进行简单校准后拍摄，每个相机采集到的画面代表了左右眼的视点。由于左右两路视频序列独立存在，因此可以很容易的兼容现有的 2D 显示器，只需丢弃其中一路视点即可实现 2D 显示。

另一种方式如下图所示，将左右两个视点拼接成一个视频序列进行存储或传输，两眼图像可以按照左右或者上下的格式进行排列。使用这种方式的立体视频可以通过专门的立体摄像机采集，也可以在两个摄像机分别采集后再进行拼接处理。其优点是可以直接使用现有的信道进行传输，也可以使用通用的平面视频解码器进行解码。现有的立体视频服务多采用上述两种方式对双目立体视频进行编码和解码。

图 24 左右视点拼接

事实上，立体视频的两幅图像通常具有比较强的视点相关性，存在着大量的空间冗余。对这种立体视频格式进行编码时，可以采取视点间预测方式。一种简单的实现方式是使用普通的平面视频编码器对基准视点的图像进行压缩，然后利用已经编码的基准视点图像来预测和编码另一视点的图像。这种预测编码的过程类似于利用运动矢量和残差图像进行运动补偿预测，从而实现对另一视点图像的高效压缩，减少存储空间和传输带宽的需求，同时保持较好的视频质量。

3.多视点视频编码技术

随着裸眼 3D 立体显示器、自由视角电视等设备不断进步，并走入家庭消费场景，多视点编码成为研究热点。与普通立体显示设备不同，多视点显示设备需要同时传输多个视点的画面来提供多角度、立体的观看体验，因而传统的平面编码方式所需要的编码码率和与视图的数量近乎等比例增长。一种比较好的替代方案是以多视点加深度图的方式（MVD）来传输 3D 视频。在 MVD 格式中，只需对少数几个视图进行编码，但每个视图都有对应的深度图，通过这些深度数据可以还原出所采集场景的基本几何结构。基于传输的视频图像和深度图，可以使用基于深度图像的虚拟视点合成（DIBR）技术在接收端生成任意视角的 3D 视图。

为满足上述需求并充分利用现有的平面视频编码标准提供最先进的压缩能力，动态图像专家组（MPEG）成立了一些专门的小组（如 JCT-3V）并开发了一系列现有编码标准的多视点扩展如 MVC+D、MV-HEVC、3D-HEVC、MIV 等，以下做简单介绍。

MVC+D 和 MV-HEVC 的设计原则是重复使用基础平面视频编码标准 AVC 和 HEVC。因此只需要更改现有标准的切片头（SLICE）或者更高级的语法元素就可以实现。MV-HEVC 还采用了层（layer）间处理的方式，其高级语法允许各层之间通过参考图像列表进行关联，允许预测层的图像使用参考层的图像进行预测编码。此外，通过辅助图片层（Auxiliary picture layers）机制来支持深度视图，而有关深度辅助层的更详细属性，可以通过 SEI 消息提供。

3D-HEVC 通过引入新的块级编码工具进一步降低了码率，这些工具有效地利用了视频纹理与深度之间的统计依赖，并专门适应了深度图的特性。由于深度图通常包含由锐利边缘分隔的均匀区域，因此采用了新的帧内预测和残差编码方法，来处理这些独特的信号特征。此外，还指定了新的深度图编码工具，允许进行视点间运动预测，或从纹理层预测运动和分块信息。新引入的预测技术通过使用子块分区（SBP）来提升预测精度。在某些情况下，这些子块分区可以将一个预测块（PB）细分为具有非矩形形状的两个部分，从而进一步优化编码效果。在需要视频纹理与深度的应用场景中，3D-HEVC 提供了更大的优势。

MIV 是为了支持沉浸式视频内容的压缩而开发的。一个真实或虚拟的 3D 场景可以由多个真实或虚拟摄像机采集得到，该标准使得沉浸式视频内容可以通过现有和未来的网络进行存储和分发，并支持以 6 自由度（6DoF）的视点位置和方向进行播放。MIV 是一个灵活的标准，支持带有深度图的多视点视频（MVD）和多平面视频（MPI），并利用强大的硬件支持来对体积视频进行编码。所有配置文件都有符合性比特流，MIV 主配置文件用于 MVD，MIV 扩展配置文件支持 MPI，此外还有适用于基于云端和解码器端深度估计的 MIV Geometry absent profile 文件。除了符合性测试外，MIV 的验证测试也已完成。

下图展示了 MIV 的编码和解码过程。在编码器阶段，包含纹理及深度组件的多个源视图以及相机参数列表输入 MIV 参考软件—TMIV 编码器。编码器将输入视图标记为基本视图和附加视图，后者根据视图间的冗余进行修剪。然后，所有视图以补丁（patch）的形式按光栅顺序打包到视图集中，并使用 HEVC 编码器对视图集进行编码，子码流与包含 patch 信息的元数据一起复用形成 V3C 格式的码流。在解码器端，码流被解复用和解码，获取视图集和元数据，并传递给播放终端，从而根据客户需求渲染出场景的任意视点。当前的 MIV 标准使用 HEVC 技术，由于 V3C 格式与视频编码标准无关，实际上可以使用任意编码格式如 AVC、VVC、AVS3 等。MIV 码流还包括高级语法，用于对齐视图集和相机，从而对视角相关的流进行解码和渲染。

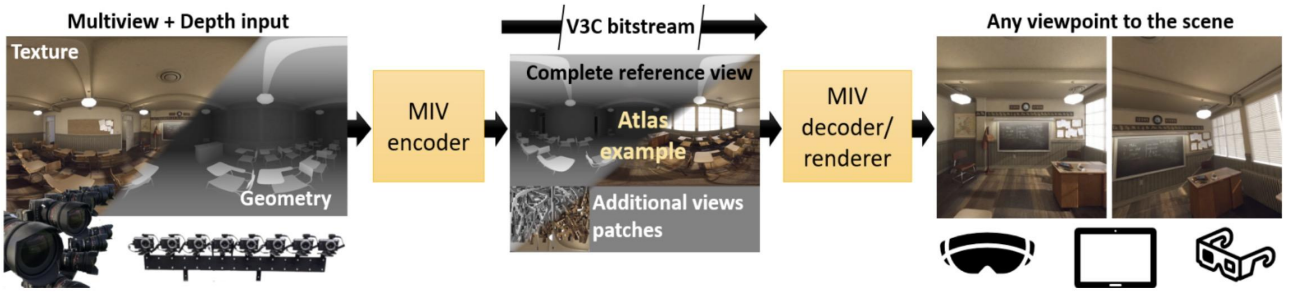


图 25 MIV 编码流程

4.全景立体视频编码

由于全景视频是包含了 360 度视频内容的球面视频，而现有的视频编码和存储技术并不支持对球面视频直接进行处理，因此全景视频在编码或存储前需要通过全景视频投影技术将其投影为二维的平面视频。然后在二维平面视频上进行编码。因此全景立体视频编码技术分为全景视频的投影技术和全景立体视频编码技术。

1）全景视频投影技术

常见的全景视频投影方式包括等距圆柱投影、多面体投影、非均匀投影等。

等距圆柱投影（Equirectangular Projection，ERP）：是一种简单的地图投影方法，在这种投影方法中：假设球面和圆柱面相切于赤道, 将球面上的经纬线投影到圆柱面上, 然后沿圆柱面的一条母线展开成平面的一种投影。ERP 投影使用纬度和经度组成的方形网格来表示，具有矩形且直观的优点，使用现有的视频编辑工具相对易于操作。但是在用于视频传输时，它有一些严重的问题。首先，极点得到了大量的像素，而赤道得到相对较少。因为球面视频的重要内容通常分布在赤道地区周围，也就是观看者的水平视野上。它还具有高失真，这对现有的视频压缩技术造成了一些困难。

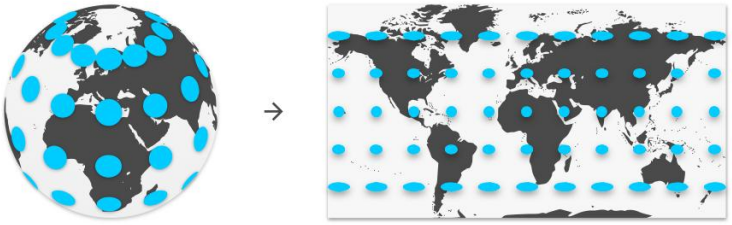


图 26 等距圆柱投影

多面体投影：多面体投影使用球心透视方位投影的方法将球面投影到外切的多面体上，然后再将多面体展开成二维平面。多面体投影包括正四面体投影、立方体投影、正八面体投影、正十二面体投影和正二十面体投影等，下图给出了上述投影的 3D 模型和展开后的二维平面的示意图。在全景视频投影中，立方体投影、正八面体投影和正二十面体投影均有研究和使用。相较于等距离圆柱投影，多面体投影在采样密度上有明显的

改善，但其对球面的采样密度仍旧不是完全均匀的。为此可以使用等角投影的方式，确保多面体面上采样长度相等，从而创建均匀分配的像素。

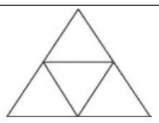
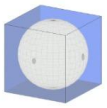
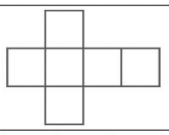
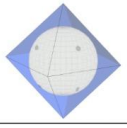
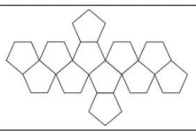
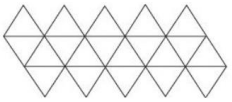
投影方式	3D 模型	投影得到的 2D 平面
四面体投影 (4 面)		
立方体投影 (6 面)		
八面体投影 (8 面)		
12 面体投影 (12 面)		
20 面体投影 (20 面)		

图 27 多面体投影方式

非均匀投影方式：全景视频均匀投影技术适用于各种场景的全景视频应用中，但是其编码效率较低。为了取得更高的编码效率，可以采用非均匀投影技术。非均匀投影技术将球面分为主视点区域和非主视点区域，投影时对主视点区域使用较高的采样密度以保证主观质量，而对非主视点区域采用较低的采样密度以节省码率。非均匀投影主要适用于一对多的基于全景视频流切换的应用中。在基于全景视频流切换的应用中，服务器端编码多路不同主视点的非均匀投影的视频，服务器和客户端之间根据用户头部视点实时选择最近的一路流进行传输。

2）全景立体视频编码技术

全景立体视频编码技术是针对全景立体视频的特殊需求而设计的一种编码方案。与普通全景视频相比，全景立体视频需要在左右眼分别渲染有视差的全景图像。与双目立体视频类似，全景立体视频也涉及左右眼的视差处理，但是全景立体视频的视图是以球面形式呈现的，因此需要先将球面图像分别投影成两个平面视图，然后再进行编码。

全景立体视频的左右眼图像可以使用传统的平面视频编码方式分别编码，也可以采用上文提到的双目立体视频编码方式或者多视点编码的方式。传统的平面视频编码方式将左右眼图像分别处理，然后采用常规的

视频编码算法对其进行编码和压缩，最终生成左右眼各自的视频流。而双目立体视频编码方式则将左右眼图像作为一对立体图像进行处理，通过特定的立体视频编码算法对其进行联合编码，以充分利用左右眼之间的相关性和视差信息，从而实现更高效的压缩和传输。此外，也可以将左右眼视频做为两个视点输入多视点编码器如3D-HEVC 或者 TMIV 进行编码。选择哪种编码方式取决于具体的应用场景、编码效率、解码和渲染终端等因素。

5.体积视频编码

体积视频通常由一系列包含深度信息的帧组成，常见的表示形式包括点云、多视点加深度、三维网格等。

点云：是三维沉浸视频领域广泛使用的数据格式之一，但其原始数据通常过于庞大，难以高效存储和传输。此外，更高分辨率的点云采集技术对点云数据量的大小提出了更高的要求。为了使点云数据可用，压缩是必要。为此，MPEG-I 发布了基于几何的点云压缩 G-PCC（Geometry-based Point Cloud Compression）标准和基于视频的点云压缩 V-PCC（Video-based Point Cloud Compression）标准，可以显著减少点云的数据量，推动点云格式在各个领域的广泛应用。V-PCC 的基本思想是将三维（3D）点云投影到二维（2D）图像上，如占位图像、几何图像、属性图像。然后使用传统的 2D 视频编解码器（如 AVS、HEVC、和 AV1）编码这些 2D 图像，以减少数据量。

随着深度学习技术在图像视频压缩等方面的应用进展，基于深度学习的点云压缩迎来一系列发展。基于八叉树的点云编码方法迭代地把包含点云的最小立方体划分为八个子正方体，然后用一个字节编码八个子正方体是否包含“点“这一信息，最后使用基于上下文的算术编码进一步去除相关性。基于此，一些方案利用神经网络来估计八叉树节点的熵模型，并且运用到动态场景中。还有一些方案结合八叉树架构与体素结构的各自优势，提出利用相邻节点的体素化的信息来增强对时空信息的利用，进一步提升点云的压缩效率。也有一些方案利用基于深度神经网络的变分自编码器来高效地压缩点云几何信息。

动态 mesh 网格：相较于保持固定连接性的跟踪网格序列，具有时变连接性的动态网格虽然能够提供更好的生成质量和更简化的生成过程，但也代表了庞大的数据量和复杂的压缩需求。为此，MPEG 发布了新的动态网格标准，称为基于视频的动态网格编码（V-DMC）。这一技术使用低分辨率网格序列（称为基网格）及其附加信息，如位移信息和纹理图，以重建高分辨率的输入网格序列。基网格可以使用任意网格编码器进行编码，而根据体积视频编码（V3C）格式的标准，可以使用任意视频编码器对附加信息进行编码。

2.4 三维沉浸视频内容传输

1.多视点视频

双视点、多视点双目立体视频，利用左右眼视差带来视觉上的立体感。可以是两个或多个带有视差的多路视频，也可包含深度信息。由于每个视点都有自己的视频流，多视点视频数据总量非常庞大，这对网络带宽、存储和处理能力都提出了高要求。为此可以根据用户当前的视点和潜在的移动方向，只传输相关视点的视频流，减少不必要的数据传输。

经测试，Iphone 15 Pro 使用主摄和广角摄像头拍摄的 1080p@30fps 双视点双目立体视频典型码率约 15Mbps。

2.FOV视频

4K 全景视频在 VR 眼镜上看起来也就只相当于 540P，所以 8K 分辨率视频的分发也仅仅是超高清画质体验的“入门级需求”。另外，一些游戏、体育赛事等内容的视频对帧率也有很高的要求，达到 120fps 才会有较好的体验；8K@120fps 全景视频码率在 150Mbps 以上，对网络要求过高，全解码方案也超出了芯片性能。

FOV（Field Of View）视频技术将根据视角对 VR 的 360°视频进行分段。用户无需从全视角下载和解码 360°视频。可以根据当前视角实时获取对应的视频段，并进行相应的解码，同时编码一个 2K 的全景图，它可以在接收端做“兜底”，以降低传输带宽和解码能力。为了确保 XR 用户体验良好，并避免出现眩晕等不良症状，整个系统需要将视角切换时延 P95 控制在 150ms 以内，即时延的 95 分位满足 150ms。

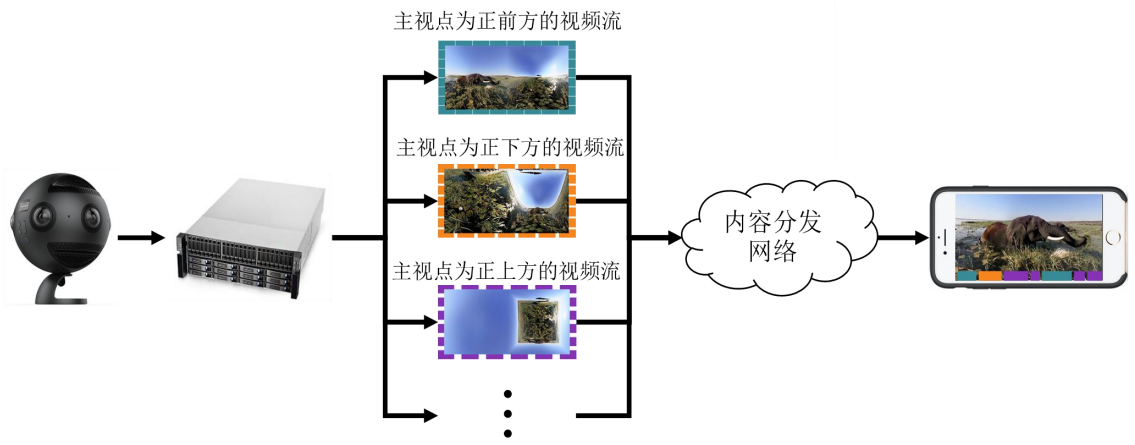


图 28 全景视频流切换过程

3.体积视频

体积视频本质是三维模型序列，可以由三维模型的 Mesh 序列和纹理贴图序列两部分组成。根据 2022 年 9 月对全球最大的 3D 资源分享社区 Sketchfab 超过 11 万的样本量统计，给出了体验分档下的典型模型参数。

表 1 体积视频典型参数 （来源：Wireless X Labs）

体验分档	帧率	纹理分辨率	纹理帧序列码率	Mesh 面数	Mesh 帧序列码率
入门	30FPS	2048Px2048P	10Mbps(H.265)	10 万面	70Mbps
良好	30FPS	4096Px4096P	25Mbps(H.265)	50 万面	125Mbps
优秀	60FPS	8192Px8292P	60Mbps(H.265)	100 万面	260Mbps

4.三维沉浸视频传输网络要求

经测试验证，当网络带宽达到视频码率的 1.5 倍时能够满足流畅播放需求，达到 5 倍时可满足“秒开”体验需求。视角切换时延需求是 100ms@95%，其中服务器时延 30ms，网络时延 100ms，客户端时延 20ms。

表 2 三维沉浸视频网络传输要求（来源：Wireless X Labs）

视频类型	分辨率	典型码率	流畅播放网络要求	“秒开”网络要求
双目立体视频 (双视点)	1080p	15Mbps	23Mbps	75Mbps
	2K	20Mbps	30Mbps	100Mbps
	4K	80Mbps	120Mbps	400Mbps
FoV 视频	4K	15Mbps	23Mbps @帧时延 P95 100ms	75Mbps @帧时延 P95 100ms
	8K	30Mbps	45Mbps @帧时延 P95 100ms	150Mbps @帧时延 P95 100ms
体积视频	2K	80Mbps	120Mbps	400Mbps
	4K	150Mbps	250Mbps	750Mbps

2.5 三维沉浸视频渲染交互

多视裸眼 3D、自由视点技术可以提供多个视点，因而观看者可以在任意允许的视点范围内观看，但视点越多，所需同时传输的数据量就越大，这对于带宽和存储都带来了巨大的压力。此外，视点越多，拍摄成本相应就越高，因此需要使用虚拟视点合成术。如下图所示，虚拟视点合成技术是一种利用已有视角的图像或视频信息，通过计算机图形学方法，在场景中生成新的视角的技术。该技术通常通过分析场景的几何和光学属性，结合视角之间的关系，使用插值、合成和空洞填补等算法，生成具有逼真效果的新视角，使用户能够以不同的角度和位置观察场景，从而提升观看体验和增强沉浸感。按合成原理，渲染虚拟视点的方法可以分为基于模型或者几何的渲染方法，即 MBR（Model Based Rendering）方法，以及基于图像的渲染，即 IBR（Image Based Rendering）方法。三维沉浸视频的交互是指在渲染三维沉浸视频时，用户可以与视频内容进行互动的过程。这种交互可以包括改变观看角度、调整视角位置、缩放或移动场景等操作。通过交互，用户能够更加自由地探索视频内容，增强沉浸感和参与感。这需要使用先进的渲染技术和交互设计，以确保用户体验流畅、直观和令人满意。

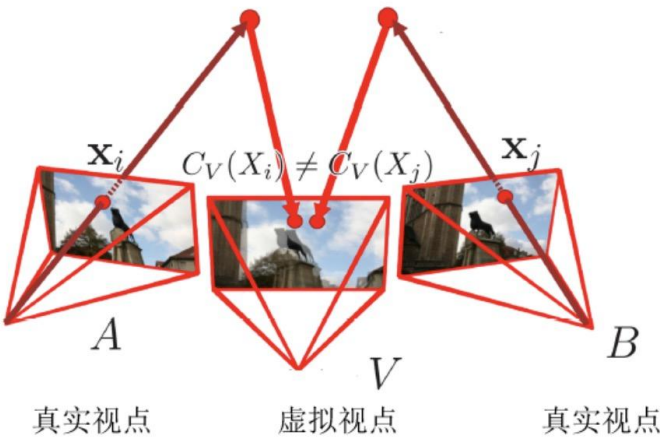


图 29 虚拟视点合成

1.基于模型的渲染技术

基于模型的渲染是通过使用三维场景模型（包括几何形状、材质属性、纹理等）来生成图像的过程。它涉及将三维场景投影到二维视图平面上，计算每个像素的颜色值，并考虑光照、阴影、反射等视觉效果，以创建符合真实或虚拟环境的图像。基于模型的渲染依赖于预先创建的三维模型，这些模型可以通过不同的方式获取，如高精度扫描仪、多视立体几何等。

虽然 MBR 方法在生成虚拟视点时提供了较高的自由度，但建模过程较为困难，渲染效果高度依赖于三维模型的精度和质量，难以达到照片级真实的效果。且基于模型的渲染适用于静态或预定义的场景，对于动态变化或实时生成的内容，其适应性较差。在需要实时互动和响应的应用中，基于模型的渲染可能难以迅速调整

和更新图像，影响用户体验。因此，纯建模的 MBR 方法虽然在计算机图形学、电子游戏等领域应用广泛，但不太适用于交互视频的应用。

2.基于深度图像的渲染

IBR 技术通常无需建模，通过二维图像及相应的几何信息即可渲染虚拟视点。通常来说，场景的图像信息容易获取且合成速度较快，但由于图像只包含二维信息，缺乏空间几何信息，导致合成的虚拟视点质量不理想。而使用基于深度图的合成技术(Depth Image Based Rendering, DIBR) 则可以平衡合成质量与速度。

DIBR 技术利用输入数据中提供的深度信息，结合参考视点和虚拟视点不同的相机位姿，生成虚拟视点所能看到的图像。相较于传统的 IBR 方法，DIBR 技术不需要精确的几何建模过程，只需要在前期为参考视点图像生成对应的深度图，即可完成虚拟视点图像的绘制。因此，DIBR 技术可以更高效地处理大规模场景。此外，DIBR 技术的输入数据都是二维图像信息，这使得它非常方便进行后续的压缩和编码传输，使其在实际应用中更为可行和有效。

基于深度图的虚拟视点合成首先将原图中的点反投影至真实世界中的 3D 坐标，接着，将 3D 点重投影到用户指定视角的成像平面上。在 DIBR 系统中，所有三维点的坐标、相机内外参数都需要作为元数据传递到接收端。多视点采集系统与虚拟视点合成系统都在相同的三维世界坐标系下，以便采集系统的真实摄像机和虚拟摄像机之间的相对关系能很好地定义。基于以上几何关系，合成步骤如下图所示：

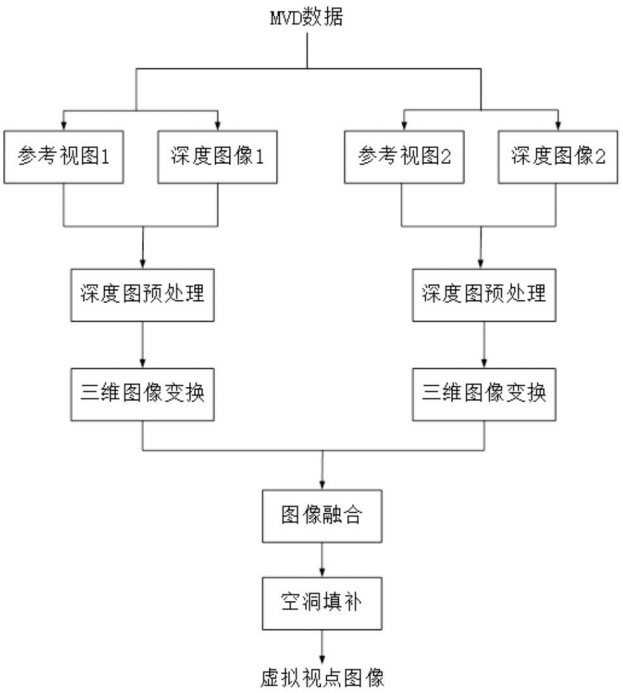


图 30 DIBR 合成虚拟视图的流程

深度图在获取时常常会受到噪声和深度与纹理不对齐等问题的影响，导致合成的虚拟视点出现前后景纹理错误等情况。为了减轻这些问题带来的影响，在进行图像变换之前通常需要对深度图进行预处理，例如使用中值滤波或高斯滤波等方法，以平滑深度图像的过渡，避免产生突兀的效果。接着，通过三维图形变换（3D Warp）的过程，利用图像纹理、深度信息以及相机内外参数，建立起参考视点和虚拟视点像素之间的联系，从而合成虚拟视点。在此过程中，需要确保虚拟视点能够准确地反映场景中的几何和纹理信息，以实现真实的合成效果。得到多张虚拟视点图像后，通常需要进行图像融合的步骤，以生成最终的虚拟视点图像。这一过程可以通过将同一位置的像素点根据其距离进行加权融合来实现。最后，由于虚拟视点的部分区域可能无法从任意视点获得，因此需要对虚拟视点图像进行空洞填补，以完善合成的虚拟视点。在经典的 DIBR 框架中，一种简单且快速的空洞填充方法是使用均值滤波，这种方式在速度上比较有优势。

虽然 DIBR 技术具有传输简便、节省带宽和合成速度快的优点，但合成虚拟视点的图像质量仍然是一个挑战，常见的问题包括空洞、伪影、边缘模糊和时域不稳定等现象。

3.基于图像域形变的渲染

基于图像域形变的虚拟视点合成是另一重要的视点合成方法。对比依赖稠密深度图或者视差图的 DIBR 技术，IDW 通过稀疏的视差关系即可合成新的视点。人眼并不能精确地估计绝对深度，对于看似合理的图像，人眼对图像失真并不十分敏感，因此可以将图像失真隐藏在非显著区域。受到这一点的启发，IDW 可以自动地估算图像对之间的稀疏视差以及的角点对，根据这些匹配关系，可以计算参考图到最终合成的虚拟视点图中的 warp，并且将失真隐藏在非显著的区域中。一类经典的 IDW 算法如下图，可以由两个视点合成多个视点。

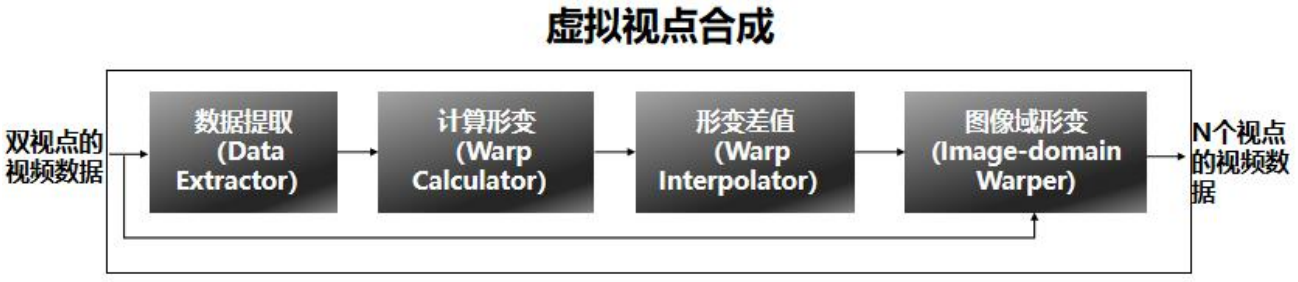


图 31 双视点合成多视点步骤

首先进行数据提取，获取输入图像的稀疏视差特征以及图像显著性特征。稀疏视差就是同一极线上匹配的像素点的横坐标之差，常用的方法有特征点匹配、光流追踪等，显著性特征则可以在后续步骤中减少合成误差。完成数据提取后，如下图所示，需要将输入图像被形式化为一个个规则的网格，然后定义一个非线性能量函数对网格进行畸变后得到新的图像。接着进行形变插值，一般选取两个最近的相机作为参考，并将其

warp 到中央相机，而其他位置的虚拟视图则可以通过左右真实相机以及中央的虚拟视图插值而得到。最后进行图像域形变，由于 warp 是连续的，因此虚拟视图中不会出现空洞现象，或者说通过对非显著性区域进行拉伸隐性地对闭塞区域进行了纹理修复。然而，仅仅使用一张图合成虚拟视点会造成边缘区域空洞，因此该区域再使用另一张图作为参考以补偿边缘空洞。

这种方法依赖于稀疏视差和图像显著性信息，约束合成的虚拟视图强行满足目标的视差估计，在没有深度图的情况下依然具有相对高质量的合成结果。

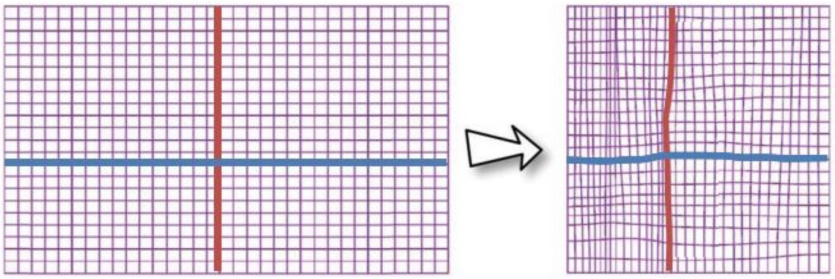


图 32 wraping 示例图

4.三维沉浸视频交互

三维沉浸视频交互是指用户通过直观的交互方式，沉浸在三维虚拟环境中并与其中的内容进行互动。用户由被动观看变为主动欣赏，可以在三维空间中自由探索，从而获得身临其境的互动体验。典型的交互方式包括：

头部追踪及人眼追踪：在头戴式显示器或增强现实眼镜中，用户可以通过转动头部自由改变视角，从而更自然地观察和探索三维场景。此外，在裸眼 3D 显示技术中，通过人眼追踪技术，系统可以实时检测用户的视线方向，从而动态调整显示内容，为用户提供更宽广的 3D 可视角度和更稳定的立体效果。

触摸屏操作：用户通过触摸屏幕或触摸板等设备直接与系统进行交互。例如，在裸眼 3D 设备上，用户可以通过滑动屏幕来切换视角，或者通过多指操作来放大、缩小、旋转、进入或退出场景。触摸交互的直观性和便捷性使其成为三维沉浸视频中常见的交互方式。

体感和手势交互：体感设备可以捕捉用户的身体动作，实现与沉浸视频内容的交互，如微软的 kinect 可以通过深度摄像头和红外传感器捕捉用户的全身动作。手势识别技术的进步使得手势操作更加丰富和精确，如 Apple Vision Pro 定义了互点手指、捏合并拖移、轻触等多种手势，使得交互过程更加自然和直观。

控制器交互：是虚拟现实体验中常见交互方式。虚拟现实控制器，如 SKYWORTH Pancake 1C 6DoF 手柄和 HTC Vive Controllers，通过内置传感器和触控面板实现精确的运动跟踪和便利的操作。这些控制器不仅提供高精度的空间定位，还支持振动反馈和力反馈，使得 VR 体验更加丰富和身临其境。

语音交互：随着大语言模型如 ChatGPT、文心一言等的广泛应用，以及语音识别技术的发展，语音交互技术赋予了虚拟环境更高的智能化和便利性。例如，用户可以通过语音命令轻松实现视角切换、播放控制、场景变换等操作。这种交互方式在无需手动操作的情况下提供了极大的便利，尤其是在需要专注于其他任务或复杂操作的场景中，语音交互成为了一种高效的交互手段。

2.6 三维沉浸视频终端显示

三维沉浸视频的终端设备包括支持高分辨率和高帧率的显示设备、3D 眼镜与 VR 头戴显示器、裸眼 3D 设备等。立体显示设备的种类繁多，基本原理都是相似的，通过为两只眼睛呈现不同的图像，以实现三维立体的效果，下面依次介绍几类三维沉浸终端显示设备。

1.平面视频显示器

平面视频显示器是一种用于显示 2D 视频内容的设备，通常采用阴极射线管（CRT）、液晶显示（LCD）、发光二极管（LED）、有机发光二极管（OLED）或其他类似技术。这些显示器广泛用于各种场景，包括电视、电脑显示器、移动设备、商场大屏等。

平面视频显示器具有一些显著的优势，例如高分辨率、良好的色彩表现。它们能够以高质量和高清晰度显示视频内容，使用户能够享受到更加逼真和清晰的视觉体验。此外，平面视频显示器通常具有较低的功耗，使其在节能环保方面具有优势，同时也更加轻薄便携，适用于各种场所和应用场景。

近年来，平面显示器的发展体现出以下趋势：

更高的分辨率和更高的像素密度：随着技术的进步，平面显示器的分辨率和像素密度不断提高，从 1080p 到 4K 再到 8K 甚至 12K，以实现更清晰、更逼真的图像显示。高分辨率和高像素密度的显示器可以呈现更多的细节和更精细的图像，提升用户的视觉体验。

10bit 色深、高动态范围、宽色域：随着显示技术的发展，平面显示器对色彩的还原能力也在不断提升。10bit 色深意味着显示器可以呈现更细腻的色彩渐变和更真实的色彩表现。高动态范围意味着显示器可以展示更广泛的亮度范围和更高的对比度。宽色域意味着显示器能够覆盖更广的色彩空间，显示出更多的颜色。

更薄更轻的设计：随着人们对便携性和美观性的需求不断增加，平面显示器的设计趋向更薄更轻。柔性屏和折叠屏的广泛使用使得屏幕变大的同时，体积和重量减少，便于收纳和携带。

更高的刷新率和更快的响应时间：对于游戏和多媒体应用来说，高刷新率和快速的响应时间是至关重要的。目前显示器支持的刷新频率从 60Hz 到 120Hz，一些高端的显示器还支持 144Hz 甚至 240Hz，以满足用户对于流畅游戏和视频播放的需求。

然而，与其他类型的显示技术相比，平面视频显示器也存在一些局限性，例如有限的观看角度，缺乏立体感等。

2.眼镜式 3D 显示

眼镜式 3D 显示的主要实现方法有三种，色分式、偏光式和时分式。

色分式 3D 眼镜又称为红蓝眼镜，左右眼分别看到的图像使用不同的颜色滤光片进行过滤，通常一个眼镜片是红色，另一个是蓝色。在观看时，一只眼睛只会接收到红色光，另一只眼睛只会接收到蓝色光，从而实现立体效果。然而，这种技术会导致颜色失真，并且观看时可能出现视觉疲劳，因而适用范围较小。

偏光式 3D 眼镜是一种广泛应用于电影院、电视和其他娱乐场所的 3D 眼镜。这种眼镜利用偏振光的特性，使观众的左右眼分别接收到对应偏振方向的光，从而实现立体效果。在观看 3D 影像时，屏幕上显示的图像采用交替的线性偏振或圆偏振方式。每只眼睛的眼镜偏振片过滤掉特定方向的光线，确保每只眼睛只能接收到对应的偏振图像。这样，左右眼看到的图像经过大脑的融合，产生了真实的立体效果。

时分式 3D 显示技术会在不同的时间段内切换显示不同的图像或图像信号。例如，在某一时刻，屏幕会显示左眼所需的图像，同时眼镜的滤光器或偏振器会使左眼只接收到这部分图像的光信号，而右眼则会被屏蔽或接收到不完整的图像光信号。然后，在接下来的时刻，屏幕会显示右眼所需的图像，并相应地调整眼镜的滤光器或偏振器，使右眼只接收到这部分图像的光信号，而左眼则被屏蔽或接收到不完整的图像光信号。这种方式为了保证能看到连续不闪烁的 3D 图像效果，一般会要求显示器的刷新率达到 120Hz，这样左右眼分别可以达到 60Hz 的刷新率。

3.头戴显示器

头戴显示设备（Head Mount Display，HMD）是一种可以佩戴在头部的装置，通过将显示屏置于用户的眼睛前方，实现沉浸式的观看体验。由于头戴式显示设备天然的在用户左右眼分别播放画面，所以可以方便的支持 3D 立体显示。这些设备通常包括显示屏、透镜、传感器和计算处理单元等组件，可以提供沉浸式的视听体验和交互功能。头戴显示设备可以分为虚拟现实（VR）头显、增强现实（AR）头显和混合现实（MR）头显三种类型：

虚拟现实头显（VR）：虚拟现实头显通过完全封闭用户的视野，并提供全景的虚拟环境，使用户完全沉浸在虚拟世界中。这些设备通常配备高分辨率的显示屏和透镜，可以呈现出逼真的虚拟场景，并通过头部追踪技术实现用户的视角变换，从而营造出身临其境的体验。

增强现实头显（AR）：增强现实头显通过透明的显示屏将虚拟内容叠加在现实世界中，使用户可以同时看到虚拟图像和真实环境。这些设备通常配备摄像头和传感器，可以实时捕捉用户的周围环境，并将虚拟图像与现实场景进行融合，为用户提供丰富的增强现实体验。

混合现实头显（MR）：混合现实头显结合了虚拟现实和增强现实的特点，既可以呈现出完全虚拟的环境，又可以将虚拟图像与现实环境进行交互和融合。这些设备通常具有更高级的传感器和计算处理能力，可以实现更复杂的虚实融合效果，为用户提供更加逼真的混合现实体验。

头戴显示设备可以应用于游戏娱乐、教育培训、医疗保健、工业设计等领域，为用户提供沉浸式的体验和全新的交互方式。随着技术的不断进步和成本的降低，头戴显示设备有望成为未来人机交互和娱乐体验的重要载体。

4.裸眼 3D 显示

裸眼 3D 显示设备可以分为三个大类，全息 3D 显示器（holographic 3D displays）、体积 3D 显示器（volumetric 3D displays）、和多视角立体 3D 显示器（autostereoscopic 3D displays）。

全息 3D 显示是一种能够记录并再现实物的振幅和相位信息的先进显示技术。它通过记录激光光束经过物体时的相位和振幅信息，然后再用这些信息通过特定介质（例如光折射聚合物）来进行再现的技术。此外，通过使用直接调制相干波的空间光调制器，可以通过数值模拟实现计算机生成的全息系统。

体积 3D 显示利用一些特殊的介质，如被困的颗粒或荧光屏幕，来产生空间中的光点（也称为体素）。这些光点通过在介质中激发光源，形成发光的图像点。通过控制光源的位置和强度，可以在空间中形成各种形状和图案，从而实现立体显示效果。体 3D 显示器还可以通过高速旋转的 2D 屏幕形成可供显示的 3 维空间，然后利用高速投影仪将 3 维内容各个角度的切片图像投影到 2D 屏幕上，这需要切片图像和 2D 旋转屏保持合适的频率。全息 3D 显示和体 3D 显示所需要的数据内容极其庞大，因而面临着数据处理和传输的挑战。

多视角立体 3D 显示与上述技术相比，通过将 3D 物体的连续光场分解成多个视图，大大降低了计算成本。典型的多视角立体 3D 显示仅需两个主要组件：光学元件和可刷新显示面板（如液晶显示、有机发光二极管显示、发光二极管显示）。这种设计紧凑、易于与平板显示设备集成、易于调制且成本较低，非常适用于便

携式电子设备。多视角立体 3D 显示中光学元件的作用是调制视图与视图之间的角度间隔，依照调制方式的不同可以分为以下 3 个类别：

1）基于视差壁障的 3D 显示：这种技术使用一层被称为视差障碍或者视差栅栏的遮光层，该层位于显示屏和观众之间。时差障碍层包含一系列微小的条纹或凹槽，通过这些结构来限制观看着左右眼看到的图像，从而在脑海中形成立体的图像。这种方式的缺点是随视角增加，分辨率和亮度均会降低。

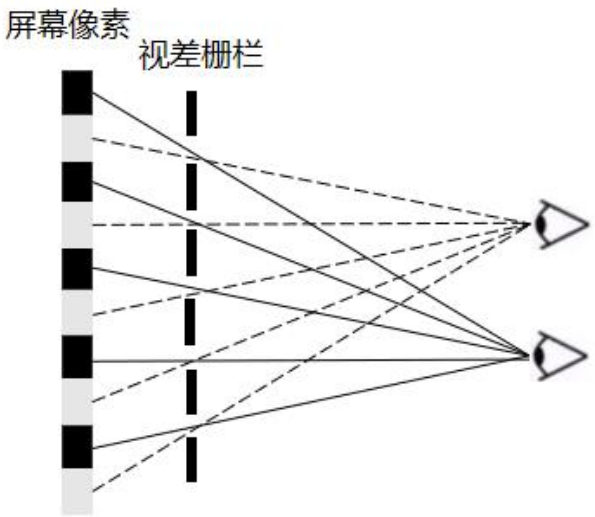


图 33 基于视差壁障的 3D 显示

2）基于柱状透镜的 3D 显示：这种技术使用柱状透镜，透镜表面有一系列纵向排列的微型柱状凸起。这些柱状透镜通过调整左右眼所看到的像素，使得左眼和右眼分别感知到不同的图像，从而产生立体效果。为实现多视角的显示，可以使用每个微透镜记录多个视角的子图像，每个微透镜的子图像都包含了若干个像素，此时各像素所记录的光线强度就来自于一个微透镜和一个镜头的子孔径区域之间所限制的细光束。这种技术同样会导致分辨率的损失。为了拓展可视角度，还可以添加眼球追踪系统获取人眼的所在的位置，通过这种方式实时调整显示屏显示图像的位置，从而扩大可视范围。

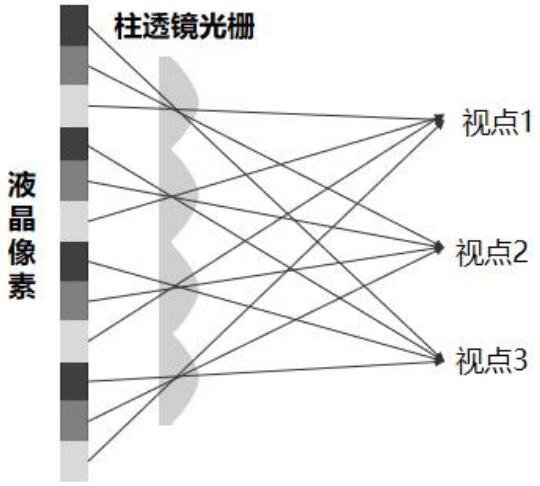
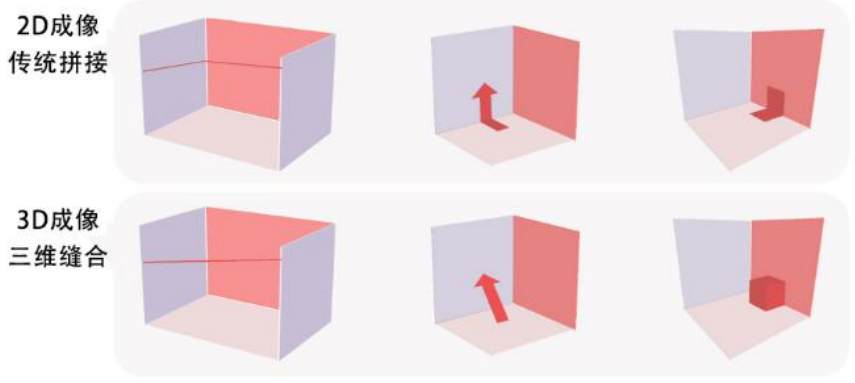


图 34 基于柱状透镜的 3D 显示

3）指向光源的裸眼 3D 显示：该技术搭配两组不同角度的 LED，配合快速反应的 LCD 面板和驱动方法，控制两组屏幕分别向左右眼投射视图，让 3D 内容以序列（sequential）的方式先后进入观看者的左右眼产生视差，进而让人感受到 3 维立体的显示效果，这种方式在分辨率和透光率方面能得到保证，不会影响现有的设计架构，但技术尚未成熟。

4）动态视点的裸眼 3D 透视显示，该技术使用多块不同角度的屏幕（至少是两个不同显示面，常见为 LED 屏幕）使用三维缝合拼接技术进行立体内容显示。真实 LED 屏幕的空间姿态和箱体参数与实时渲染三维虚拟空间进行了重建与匹配计算，使两个空间中屏幕参数精准匹配，显示内容基于最佳视点进实时计算渲染，显示内容会跟据最佳视点移动而变化，始终保持最佳视点看到的立体视频内容 3D 透视关系是正确的。



来源：七维视觉科技

图 35 动态裸眼 3D 显示

2.7 三维沉浸视频质量评价

针对普通的二维视频，其质量评价主要可以分为两大类：主观质量评价、客观质量评价。主观质量评价是由测试者按照规定的实验流程观看一组存在失真的视频，并对视频质量进行主观打分评价的方法。而客观质量评价通过设计数学模型来模拟人眼对图像质量的感知，以尽可能实现和主观评价一致的评价结果。与传统的二维视频不同，三维沉浸视频可以提供空间维度、沉浸感和临场感等观影体验。因其沉浸式的视觉体验，观众往往会有较强的临场感和包围感，这对视频质量的评价产生重大影响。受观影设备的影响，三维沉浸视频观影设备（如 VR 头显）的性能对视频体验影响巨大，分辨率、刷新率、FOV 和设备的舒适度都可能直接影响到质量评价。因此，除了二维视频的评估指标外，三维沉浸视频质量评价还需要考虑深度感知、视场角（FOV）、延迟、运动跟踪精度、渲染质量和实时性等影响。其主观评价也会包括更多心理层面的评价，如舒适度、沉浸感等。

1. 三维沉浸视频主观质量评价方法[6,7]

三维沉浸视频主观质量评价实验包括针对各种终端显示设备，如使用眼镜式 3D 显示（由于用途限制，使用较少），头戴式显示器（HMD），以及裸眼 3D 显示设备等观看的 360°视频。[ITU-T P.919]标准较为详细介绍了时长为 10s-30s 之间的较短三维沉浸视频的主观实验方法。

1）视频源选择：实验使用的 360°视频源应根据学术研究的具体目标进行选择，并记录在数字存储系统中。应保证原始视频的质量尽可能高，尽可能使用最大空间分辨率和帧率，并使用原始的、未压缩的视频。视频源应当具有足够充分的空间信息和时间信息，同时应当保证在测试中，源视频可以引起受试者各种不同类型的探索行为。

2）主观实验环境：应控制实验环境尽量保持安静，且环境场景中没有可能引起受试者注意力分散的因素。同时保证受试者可以合理地利用实验设备进行实验。为了保证沉浸视频的特点，并保证对全景视频质量的准确感知，应该保证使用的沉浸式视频终端显示设备符合商用设备的使用规范和要求，且需要保证显示设备有足够的分辨率和刷新率来显示要测试的内容。为了观看完整的 360 度视频，如果使用 HMD 显示设备，受试者应该坐在转椅上以便能够自由旋转身体，同时应保证受试者的头部和眼球可以自由活动，如果使用裸眼 3D 显示设备或是其他设备,应该保证受试者所处的位置其视场范围能够完全覆盖显示设备的显示范围。为了防止受试者突发不适并控制实验的正常进行，实验组织人应当在不影响正常测试的情况下与受试者同处一室或在隔壁房间全程监控实验的进行。实验环境的具体配置应当以文件形式记录。

3）主观实验方法：主要方法包括绝对类别评级法（Absolute Category Rating ，ACR）和损伤类别评级法（Degradation Category Rating，DCR）。ACR 方法是单刺激主观评价法，每次只呈现一个失真视频，

并在类别范围内独立评分。ACR 使用五级评分标准：5 优秀；4 良好；3 一般；2 较差；1 极差。DCR 方法属于双刺激损伤方法。使用此种方法时，参考视频和对应的损伤视频应先后成对出现，且应保证参考视频第一个出现。受试者被要求参考原始视频对失真视频进行评级。DCR 使用五级评分标准：5 难以分辨视频对之间的损伤差异；4 有可分辨的差异，但是不会引起观看体验下降；3 损伤差异会引起轻度的体验下降；2 损伤差异会引起中度的体验下降；1 损伤差异非常明显，并且极度影响体验。实验过程中，每轮观看时间不可超过 25 分钟，每轮观看之后应当要求受试者至少休息 15 分钟。

4）主观评分记录方法：若受试者使用裸眼 3D 显示设备等可以随时切换注视方向结束观看的终端显示设备，则可以在受试者面前摆放一台电脑，在观看完每段视频后让其使用合理的评分界面进行评分。若受试者观看视频时头戴 HMD 设备，无法在任意时间结束观看，因此不能用传统的纸笔或者电脑屏幕的滑块记录评分。合理的替代方法包括，使用 VR 程序在每段视频观看结束后，在 HMD 设备上显示一个评分栏，并且以受试者凝视交互或者使用手持式控制设备进行打分; 受试者也可以在观看完视频后口头描述评分情况，由实验组织人员实时记录。每段视频至少应该记录 28 名受试者的主观评分。另外，主观实验应当实时记录受试者的头动情况和头部位置，记录应由 HMD 内部的应用程序完成。

5）实验数据处理方法：对于主观测试实验的结果，应使用统计方法筛选符合规范的主观测试数据，剔除离群值。最终应给出每段视频的评估等级统计分布的均值，即平均意见分数（Mean Opinion Score, MOS）和标准差。这些统计值的计算方法见[ITU-R BT.500-14]，[ITU-T P.800.2]提供的有关信息。

2.三维沉浸视频客观质量评价方法

全景沉浸视频质量客观评价旨在设计合理的算法，准确预测沉浸视频的用户观看质量，使之达到与主观质量评价结果相近的结果。优良的客观评价算法或模型能够快速有效地预测各种场景下的失真沉浸视频质量，其研究成果能够用 于指导沉浸视频相关技术的设计和优化,进而提高沉浸视频应用的用户视觉体验质量。

目前绝大多数现存的关于全景沉浸式视频的质量评价方法集中于二维沉浸视频的质量评价方面。二维沉浸视频的质量评价主要关注投影变形对图像质量的影响，特别是在极地区域的失真（如 ERP 投影）。此外，由于用户可以自由选择视角，全景 2D 视频的质量评估需要综合考虑用户在不同视角下的感知体验。由于二维沉浸视频通常是通过特定的投影方式将 360 度视角压缩到一个 2D 平面上，因此可以通过对投影结果的 2D 视频质量进行分别评估，再回归合成对二维全景沉浸视频的质量的整体预测。

对于三维全景沉浸视频的客观评价算法，由于三维全景沉浸视频不仅需要考虑视角范围和投影变形，还必须处理双目视觉带来的额外复杂性，处理双眼之间的协调性和舒适度。全景 3D 图像包含两个视角（左视

图和右视图），在用户佩戴头戴显示设备（HMD）时，这些图像会被分别呈现在用户的左右眼中，形成立体视觉效果。因此，在三维全景沉浸视频中，除了传统的 2D 图像变形问题外，还需要处理双目视觉中的深度感知、双眼融合、竞争或抑制等问题。这些额外的挑战与难点使得现存的 2D 与 3D 视频质量评价方法难以被有效迁移至三维全景视频客观质量评价方法中，这使得目前有效的三维全景沉浸视频的客观质量评价方法数量较为有限。下面将分别简单介绍现存的二维和三维全景沉浸视频的客观质量评价方法。

1）二维沉浸视频全参考客观质量评价方法：早期的二维沉浸视频（图像）的全参考质量评价方法主要基于对 2D 视频（图像）的经典全参考评价方法，整体属于基于经验和手工提取特征的方法。一些学者直接对参考和失真全景二维视频（图像）ERP 投影格式的 2D 形式计算峰值信噪比（PSNR）和结构相似度(SSIM)，从而得出简单的全参考质量估计。然而，ERP 投影在极点附近存在严重的拉伸失真，这些方法没有考虑到 ERP 投影引起的此类明显的视觉失真。针对这类问题，更多学者进一步提出了对传统 PSNR 等方法的的增强版本，以适应用于 360°全景内容的评价。球形加权峰值信噪比(WS-PSNR[8])通过引入反映像素投影失真的权重调整原始 PSNR 的计算。克拉斯特抛物线投影峰值信噪比(CPP-PSNR[9])计算克拉斯特抛物线投影的 PSNR, 由于这种投影方式可以最大限度地减少极点扭曲，因此比原始方法有更优良的性能。此外，球面峰值信噪比(S-PSNR[10])使用均匀分布在球面上的 655362 个采样点来评估峰值信噪比。

在此之后，出现了一些基于数据驱动的机器学习方法，一些模型采用特征提取—质量评价模型训练的思路,首先设计了两组特征来描述拼接失真(模糊、重影和几何失真等)在二维沉浸视频中引起的结构和空间一致性的变化,并分别从失真视频和原始无失真视频中进行提取。然后,计算失真和无失真图像之间的特征差值,并将它们作为支持向量回归器的输入训练质量评价模型。

近年，一些基于深度神经网络的深度学习全参考质量评价方法已被提出。一种较新的全参考质量评价方法综合考虑了三维沉浸视频观看过程中视窗的选取和视窗投影图像的显著性检测。视窗（实际观看空间）指用户观看视频过程中某个时刻视野内呈现的内容范围，是一幅 2D 图像。视窗的大小与用户观看时设备的的视场角(FoV)密切相关。这种方法将失真视频 t 时刻的 360°帧图像与之前Δt 时刻的 360°帧图像取时间差值，之后将该差值与该时刻的 360°帧图像合并输入基于球面 CNN 的视窗选取网络，输出待选择的视窗和该视窗的重要性权重。之后使用软非极大值抑制(Softer NMS)算法合并重合度较高的视窗，并剔除不重要的视窗。最后选取得到用于全参考评价的视窗。之后将 t 时刻的每个选取视窗与参考二维沉浸视频对应的视窗计算误差图，并与该视窗本身合并输入到质量评价网络中，使用 CNN 与计算显著性图结合的方式预测出该视窗的全参考质量预测分数，之后将 t 时刻所有视窗的质量分数加权平均得到 t 时刻 360°失真图像帧的预测质量分数，最后取所

有帧的平均作为整段视频的全参考质量预测分数。这种方案综合考虑了用户观看二维沉浸视频时对不同时刻不同视窗内容的敏感差异以及对于单个视窗内容关注的显著性差异，是一种较全面的二维沉浸视频全参考质量评价方法。

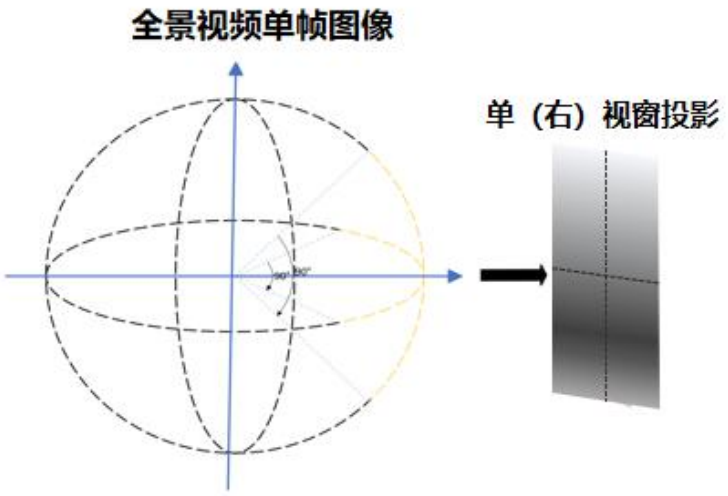


图 36 全景沉浸式视频单帧投影到右视窗示意图（FoV 为 90°）

2）二维沉浸视频无参考客观质量评价方法：主要包含基于 ERP 投影空间的方法以及基于视窗的方法。

基于 ERP 投影空间的方法的主要思想是直接通过对 ERP 投影形式图像进行特征提取和特征融合得到质量分数。由于 ERP 投影图像为二维平面形式,所以该类方法也较为简单和易于理解。然而, ERP 投影空间中的图像存在明显的拉伸形变,尤其是越靠近图像的两极区域,拉伸形变越明显,这种拉伸效应造成图像在投影空间和实际观看空间中的差异，因而造成客观评价结果与主观评价结果的不一致, 这降低了该类方法的评价准确性。

基于视窗（实际观看空间）的方法主要是通过模拟人类在现实中观看二维沉浸式内容时的真实过程或特性, 以获得与人类主观评价更接近的客观质量评价结果。这类方法中视窗的投影和选取非常重要。一种基于多通道 CNN 的无参考质量评价方法[11]引入了六面视窗投影预处理，得到二维沉浸式视频前，后，左，右，上，下六个视窗的视窗投影序列。之后使用改进的 ResNet34 结构对输入的六面投影视频帧序列进行特征提取。考虑到 ResNet 网络各个阶段的输出特征表示由底层到高层的视觉信息，为了充分利用视觉信息的等级性，在之前 ResNet34 结构的基础上，使用 HyperResNet 结构，融合网络中间层特征。最后通过全局平均池化和全局标准差池化得到每帧视窗图像的特征向量，使用全连接层输出每帧图像的质量分数，之后进行平均得到整段视频的预测质量分数。一种面向视窗的图卷积网络模型，建立了一个全景图像中视口之间的相互依赖关系模型。图形节点首先由被可视概率较高的选定视窗定义，然后通过空间关系将这些节点连接起来，捕获它们之间的交

互。 最后通过图卷积网络对获得的图像进行推理。一种基于多视窗特征融合的无参考质量评价方法，在前面所述基于多通道 CNN 的无参考质量评价方法的基础上，引入了 SlowFast 运动特征提取预训练网络和时序回归池化部分，有效建模了观看视频过程中的时序长期依赖和人记忆功能的时序滞后效应，从而得到针对二维沉浸视频更准确的质量预测分数。

3）三维沉浸视频无参考客观质量评价方法：由于三维沉浸视频的质量评价需要考虑输入图像的深度信息以及双目差异，一种较为直接的方法[12]是将左眼视图和右眼视图图像之间的差异帧作为输入，这些差异帧可以反映失真和深度信息。之后使用 3D 卷积神经网络来无参考预测 3D 全景视频的质量。这种方法概念简单易于理解，然而，上述模型是在 ERP 格式的 2D 图像块上训练的，这与实际的观看体验相冲突。

另一种较为复杂的且在生物学上合理的三维沉浸视频无参考客观质量评价方法基于预测编码理论[13]。具体而言，双目竞争被模拟为高层次图案之间的竞争，而非低层次信号之间的竞争，因为人类视觉系统（HVS）的处理原则是将自下而上的视觉刺激与自上而下的预测进行匹配。在基于预测编码的双目竞争模块（Predictive Coding-Based Binocular Rivalry Module，PC-BRM）中，左视图和右视图的假设将根据竞争优势进行竞争。该模块由预测编码过程中的先验和似然组成。因此这种方法还开发了一个多视图融合模块（Multi-View Fusion Module，MvFM），通过位置权重和内容权重方案来整合视窗图像的质量分数。双目竞争模块和多视图融合模块可以分别应用于 3D 图像和 2D 全景图像。

双目竞争模块基于人类视觉系统（HVS）的预测编码理论。传统的双目竞争模型通常认为竞争发生在早期视觉皮层的低层次信号之间，而预测编码理论则强调高层次图案之间的竞争。PC-BRM 模拟左视图和右视图的假设之间的竞争，根据竞争优势来生成视口图像的质量评分。在此模块中，预测编码模型用于计算每个视窗图像块的相似性和竞争优势。通过对视窗图像的左（眼）视图和右（眼）视图分别进行预测编码，作者得到了代表输入图像的编码系数和基向量。这些信息被用于计算视口图像的相似性和双目竞争优势，从而生成视窗图像的质量评分。

多视图融合模块用于整合视口图像的质量评分，并计算出整个立体全景图像的最终质量评分。多视图融合模块引入了内容权重和位置权重，以反映用户对场景内容和观看方向的偏好。具体而言，视窗图像的内容权重由其空间信息（Spatial Information SI）反映，高 SI 值表示视口图像中包含更多细节，因此应分配更大的权重。位置权重则基于视口中心点的纬度，使用 Laplace 分布模型计算其观看概率，从而反映用户更倾向于观看赤道区域的习惯。

最终，这两个模块组成了三维全景图像/视频质量评估器，该评估器能够准确预测三维全景图像/视频的视觉质量。同时，该模型是一个不需要回归学习的参数模型，且模型中的每个参数都对应着明确的物理意义。因此在计算复杂度方面具有明显的优势。

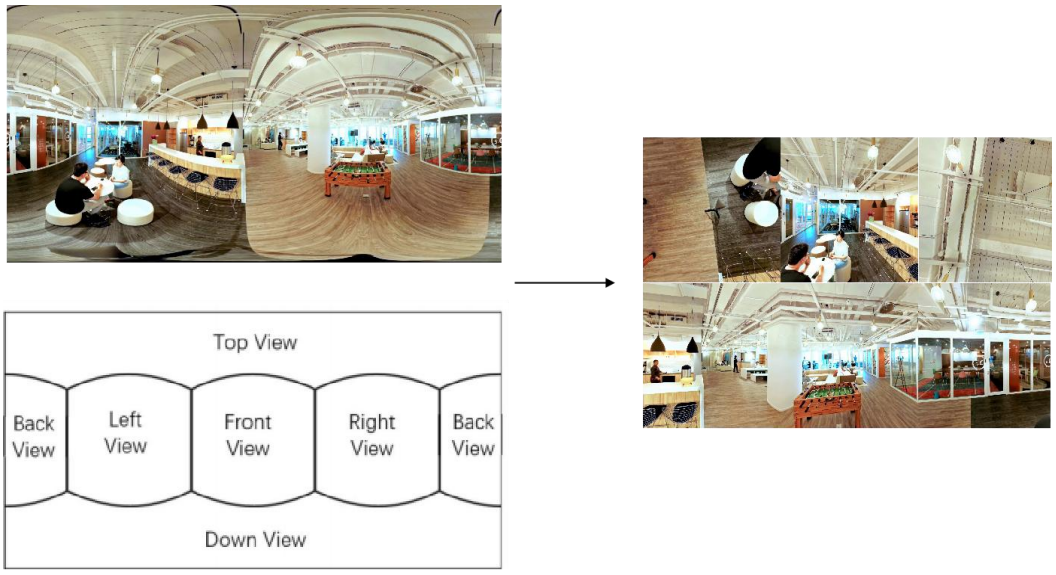


图 37 二维全景视频帧和其对应的视窗的图像(右图从上至下从左至右分别为下、后、上、左、前、右、视图)

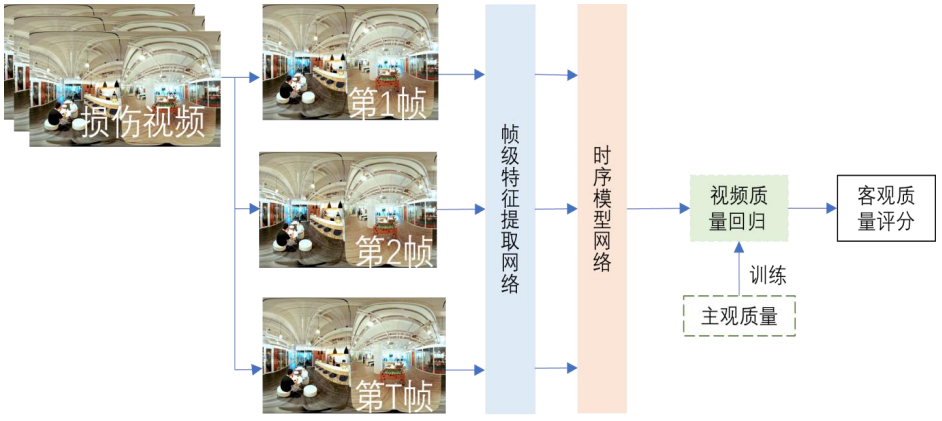


图 38 基于多视点特征融合的全景二维视频无参考质量评价模型

3.三维沉浸视频客观评价算法性能衡量

根据视频质量专家组(VQEG)的建议，使用以下指标进行**全景沉浸视频（包括二维和三维沉浸视频）的**客观评价算法性能评价。

1） SRCC(Spearman Rank-order Correlation Coefficient) 斯皮尔曼秩相关系数：

$$\mathrm{SRCC} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

其中，d_i为一段视频的主观质量评分与客观质量评分的排序之差，n 是视频总数。SRCC 衡量预测值与真实值之间的单调性。SRCC 值越接近 1，说明客观模型的预测分数与主观质量评分之间有更好的单调性。

2）PLCC(Pearson Linear Correlation Coefficient) 皮尔逊线性相关系数：

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{n}(q_i-\overline{q})(s_i-\overline{s})}{\sqrt{\sum_{i=1}^{n}(q_i-\overline{q})^2}\sqrt{\sum_{i=1}^{n}(s_i-\overline{s})^2}}$$

其中，q_i和s_i分别为第 i 段视频的客观分数和对应的主观评分，$\overline{q}$和$\overline{s}$分别是q_i和s_i的均值。PLCC 衡量预测值与真实值之间的线性相关性。PLCC 越接近 1，说明客观模型的预测分数与主观质量评分之间有更好的线性相关性。

3）RMSE(Root Mean Squared Error) 均方根误差：

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(s_i-q_i)^2}{n}}$$

RMSE 衡量预测值的预测准确度。显然，RMSE 越接近 0，预测准确度越高。

需要说明的是,在计算 PLCC 和 RMSE 两个指标之前, 需要完成五参数的非线性逻辑映射, 目的是将所有的客观质量评价方法的质量评价分数统一到同一范围内：

$$q = \beta_1\left(\frac{1}{2}-\frac{1}{1+e^{\beta_2(Q-\beta_3)}}\right)+\beta_4 Q+\beta_5$$

其中，$\{\beta_i | i = 1,2,\ldots,5\}$是拟合参数，$q, Q$ 分别为映射后和映射前的分数。最终的拟合参数实质上是能够使映射后的客观分数和主观分数之间的误差平方和最小的参数组合。

4.三维沉浸视频质量评价数据集

IVC (Images and Video Communications) [14]数据库是由法国南特大学创建的一个广泛应用于图像和视频处理领域的数据库。该数据库包括多种图像和视频质量评估数据集，常用于研究图像压缩、视频编码、图像和视频质量评价等方面。Waterloo IVC 3D Image Quality Assessment Database[15]是专门用于评估 3D 立体图像质量的数据库，Waterloo IVC 3D Video Quality Database[16] 提供了用于评估立体视频质量的数据集。IVQAD (Immersive Video Quality Assessment Database) [17]是由上海交通大学创建的一个专门用于评估 VR 和 360 度视频质量的数据库，它包含了多个 360 度视频和 VR 场景，以及通过主观实验收集的用户评分，涵盖了用户对不同视觉失真的主观感知。LIVE 3D Image Quality Database[18]这是由美国奥斯汀大学德州分校的 LIVE 实验室创建，分为阶段一和阶段二数据库，该数据库同时提供了对称和非对称失真的 3 维图

像。MCL-3D [19] 数据集是一个专门用于评估 3D 视频质量的公开数据集。该数据集由美国南加州大学的 Ming C. Lin 实验室开发，旨在支持立体 3D 视频的主观和客观质量评价研究。3D60 Dataset[20]该数据集包含了一组室内场景的 360 度立体图像和深度图，用于研究三维重建、立体视觉和沉浸式环境下的计算机视觉任务。Facebook360 Dataset 由 Facebook 发布的一个 360 度视频数据集，包含了全景视频的多种场景，用于研究视频编码、质量评估和用户体验优化，其相关工具和地址参考[21]；SUN360 Dataset[22]360 度全景图像数据集，涵盖了各种室内和室外场景，适用于沉浸式场景的研究。

3.三维沉浸视频发展趋势

3.1 当前存在的问题

内容和形式的丰富多样是三维沉浸视频的一个显著特点。随着各方面的技术进步，三维沉浸视频从早期的双目立体发展到如今包含多视裸眼 3D、全景 3D、自由视点、体积视频等多样的形式。无论是终端显示方式还是技术路径，内容表达形式还是交互方法，均差异显著，成为创新和创意表达的新前沿。这种多样化不仅丰富了用户的体验，也带来了更加沉浸式的观看体验。然而，三维沉浸视频也面临着以下一些挑战：

现有的三维沉浸视频技术难以在用户体验中兼顾高交互自由度与照片级真实的渲染效果。基于模型方法的三维沉浸视频，提供了很高的交互自由度。用户能够个性化他们的体验，例如随意切换视角、改变观看角度，甚至动态调整场景元素，以更好地符合个人需求和偏好。这种基于模型的方法虽然在交互性方面表现出色，但在渲染真实感方面却不尽如人意。这是因为其渲染效果高度依赖于三维模型的精度和质量，如果模型不够精细，渲染结果可能会显得不真实或有明显的几何误差。基于模型的渲染适用于静态或预定义的场景，对于动态变化或实时生成的内容，其适应性较差，模型的质量和计算资源的限制会影响动态场景的真实感和沉浸感。

相比之下，基于图像的方法通常用于高质量的视觉效果制作，如电影和电视广告中的三维场景。通过使用高分辨率的相机和先进的图像处理技术，可以采集到非常细致的纹理和光影效果，从而生成高度逼真的场景。但是，基于图像的方法在交互自由度上受到限制。一旦场景被拍摄和渲染，观众的视角和互动（交互自由度受限）就受到了固定镜头和视角的限制，缺乏个性化和动态探索的可能性。

三维沉浸视频的应用形态多样且碎片化，没有形成规模化应用。例如，不同厂商在拍摄多路自由视点视频时会使用自己的私有格式标记视频流，编码和压缩时会采用不同的格式，播放时必须使用专用的设备或者 APP 才能达到预定的观看效果。这种格式和技术上的差异不仅限制了内容的普及和分享，也增加了内容创作者

和开发者的负担。他们需要为不同的平台和设备开发特定的解决方案，这无疑增加了工作量和复杂性。对于消费者而言，这意味着他们可能需要购买多种设备或订阅不同的服务来体验不同类型的三维沉浸视频内容，这不仅增加了经济负担，也可能导致用户体验的碎片化。

内容匮乏也是制约三维沉浸视频发展的一个重要问题。尽管技术不断进步，但高质量的三维沉浸内容仍然不足。这主要是因为创建高质量的沉浸式内容需要大量的时间和资源。例如，创建一个复杂的虚拟现实环境不仅需要高分辨率的相机多角度采集，还需要经过复杂的三维重建以及庞大的数据存储和处理，这对开发者和内容创建者来说是一个巨大的挑战。内容的缺乏限制了技术的普及和应用的扩展，影响了用户对三维沉浸视频的接受度和需求。

总体来说，三维沉浸视频技术在用户体验、高度碎片化的应用形态以及内容匮乏等方面面临挑战。

3.2 技术发展趋势

由于现有的三维沉浸视频技术上面临的挑战。行业急需寻找一种新的 3D 视频表示方法，既能提供高自由度的交互，又能保持照片级真实的渲染效果，这种需求推动了对新技术的探索和发展。

近年来，人工智能（AI）飞速发展，为三维沉浸视频技术的演进和快速发展提供了新的机遇。神经辐射场（Neural Radiance Fields，NeRF）和 3D 高斯溅射（3D Gaussian Splatting）是两个重要的技术创新，它们在三维沉浸视频的交互和渲染方面展示了强大的潜力。NeRF 能够生成高质量的视角合成图像，提升了三维场景的真实感；而 3D 高斯溅射则通过高效的点云表示和简化计算，降低了数据处理成本。这些技术不仅可以实现照片级真实的渲染效果，提供高交互自由度，也使得三维场景的渲染变得更加实时，适应了不同的应用需求。

1.神经辐射场 NeRF

NeRF 是一种基于神经网络的渲染技术，旨在实现高度逼真的三维场景渲染。相比传统的基于几何和纹理的渲染方法，NeRF 不需要事先建立场景的几何模型或纹理映射，而是通过训练神经网络来学习场景的体密度和颜色信息，最终合成新的视角下的图像。

NeRF 的核心思想是将场景表示为一个神经辐射场，它是一个可学习的函数，接受三维空间中点的坐标和视线观察方向作为输入，并输出观察到的颜色和体密度。为了训练这个模型，需要收集场景的多个角度的观察图像，通过优化神经网络的参数，使其能够准确地从任意点生成对应的颜色和体密度。在渲染阶段，当给定相机的位置和方向时，nerf 就可以生成照片级逼真的新视图，从而实现多视角的沉浸体验。

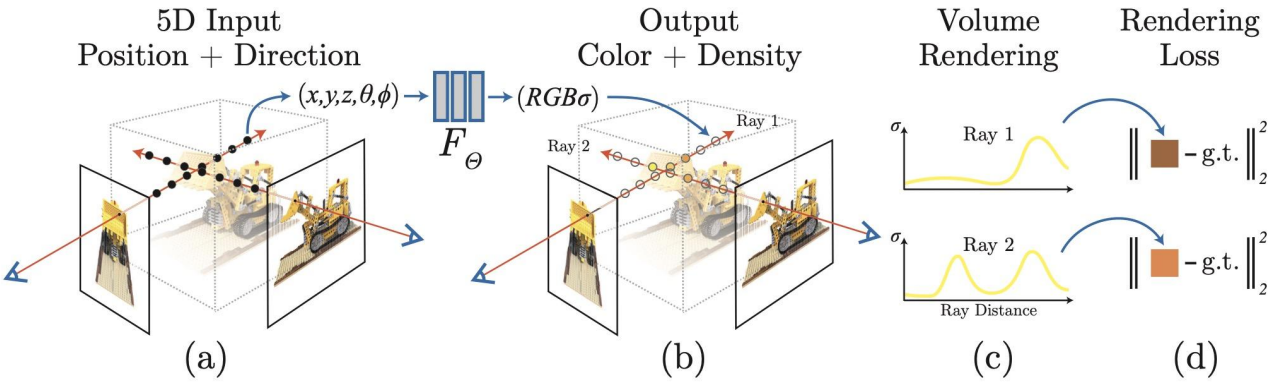
Nerf 将静态场景表示为一个连续的 5D 函数，该函数在空间中任意一点（x，y，z）的每个方向（θ，φ）上产生辐射，并且任意一点都有一个密度，该密度可以控制通过该点的射线的累积辐射。这种方法通过优化一个不包含任何卷积层的深度全连接神经网络（通常称为多层感知机 MLP）来表示这个函数，从而将一个 5D 坐标映射为一个体密度以及与视角相关的 RGB 颜色值的函数。渲染某个特定视角的 nerf 需要经过以下几步：

1）通过场景中的相机射线，生成一组 3D 点的采样集合。

2) 使用这些点及其对应的二维视点方向作为输入，产生颜色和体密度的输出集合。

3) 使用经典的体渲染技术将这些颜色和密度累积到一张二维图像中。

由于以上过程是可微的，因此可以通过最小化渲染图像与真实图像之间的距离，使用梯度下降的方式优化网络模型。通过在多个视图上最小化这个距离，使得网络将高体密度以及准确的颜色信息分配到隐含真实场景内容的位置，预测出连续的场景模型。如下图所示：



（来源：文献[24]）

图 39 nerf 表示与可微渲染过程的描述

NeRF 的优点在于其高质量的任意视角图像生成。然而，NeRF 也存很多问题，如速度慢，泛化性差，训练需要过多视角、渲染高分辨率图像很耗时等。因此，关于 nerf 的优化和相关研究仍然有很高的热度。

2. 3D Gaussian Splatting

3D Gaussian Splatting，其相关论文《3D Gaussian Splatting for Real-Time Radiance Field Rendering》获得了 2023 年 SIGGRAPH 大会的最佳论文奖。这项技术在短短数月内便在三维视觉和 SLAM（同步定位与地图构建）等领域掀起新的研究热潮。以其高质量的实时渲染能力，迅速成为学术界和工业界的热点话题。

3D Gaussian Splatting 的核心思想是：基于稀疏点云初始化一组三维高斯分布，由高斯球位置（三维高斯分布均值）、大小和朝向（三维高斯协方差）、球谐系数以及不透明度来定义。辐射场的不同方向的观测

颜色通过球谐系数（Spherical harmonic，SH）表示。优化过程中，将这组三维高斯分布通过点云抛溅算法渲染到各个视角上来计算渲染损失，并通过反向传播优化三维高斯参数。

3D Gaussian 高效的关键在于 tile-based 的光栅化器，它基于深度对高斯球快速排序，基于不透明度混合算法得到二维平面像素点颜色。该光栅化器通过累积不透明度值跟踪，支持快速反向传播。此外，三维高斯优化过程中支持根据梯度状态、透明度属性、高斯球大小做自适应高斯克隆和裁剪。

3D 高斯的输入为一组图像和使用 SFM 对该组图像估计的点云数据。将每个点初始化为一个 3D 高斯，借助 SFM 估计每个 3D 高斯的初始位置和颜色。更高质量的表征需要基于输入的多视角图像对 3D 高斯进行训练优化，以推理出更精细的位置、协方差、球谐系数和不透明度。训练步骤如下：

1）用当前所有可微高斯函数渲染出图像；

2）根据渲染图像和真实图像之间的差异计算损失；

3）根据损失调整每个高斯分布的参数；

4）对当前高斯分布进行自适应密度控制。

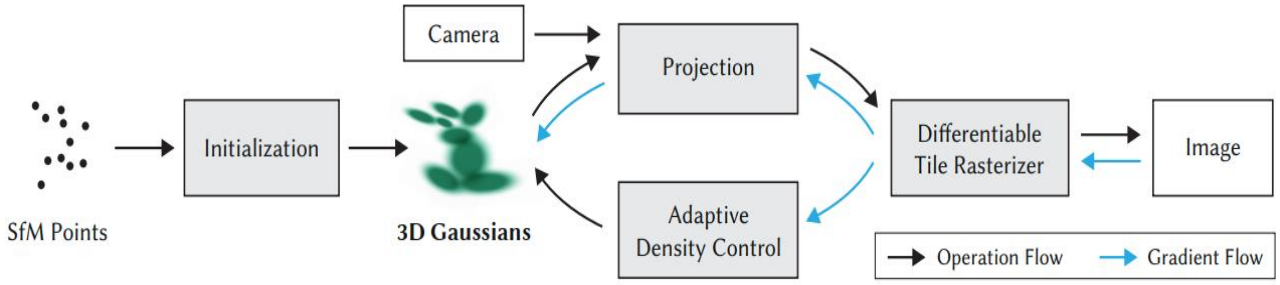
渲染步骤如下：

1）针对给定相机视角，把每个 3D 高斯投影到 2D；

2）按深度对高斯进行排序；

3）对每个像素，从前到后计算每个高斯在该像素点的值，并将所有值混合以得到最终像素值。

由于以上过程是可微的，因此可以通过最小化渲染图像与真实图像之间的距离，使用梯度下降的方式优化网络模型。通过在多个视图上最小化这个距离，使得网络将高体密度以及准确的颜色信息分配到隐含真实场景内容的位置，预测出连续的场景模型。如下图所示：



（来源：文献[25]）

图 40 3D Gaussian 的优化与渲染流程

3D 高斯的优点为：支持高质量、逼真的场景；快速实时的渲染和更快的训练速度。但是 3D 高斯也存在以下缺点：模型优化中的破碎的高斯分布（点太大、太长、冗余等）；更高的显存使用率；更大的磁盘占用；与现有渲染管线不兼容等。

3.3 3D全真视频

考虑到传统三维沉浸视频技术面临显著的体验瓶颈，即无法同时满足照片级渲染和六自由度交互的要求。现有的解决方案，不论是依赖基于模型的方法还是基于图像的渲染方法，在实现过程中往往会在三维场景的真实空间感和互动性之间做出妥协，导致用户体验不足。除此之外，沉浸视频的体验往往受到硬件的限制，对设备的依赖程度高，需要依赖复杂的观影设备，如 VR、AR 设备等。而当前的 VR/AR 设备往往较为笨重，佩戴时间过长可能会导致疲劳或不适。

为此，需要发展三维沉浸视频的一些新形态。三维全真视频（即 3D 全真视频）旨在采用多视点拍摄采集真实场景的数据，重建真实场景的全真动态三维对象，以高效的数据表达和压缩，结合先进的渲染技术，使得观众可享受在任意终端实现场内场外自由穿梭的观影体验。

3D 全真视频作为一种面向未来视频的代表形态之一，不仅追求照片级真实、超高分辨率的高质量立体视觉效果，还力求在渲染在跨平台、交互性、实时性等方面达到新的高度。随着技术的不断进步，尤其是人工智能的飞速发展，相信 3D 全真视频终将实现，并展示出广阔的应用前景。

1. 3D 全真视频的特点

相对于传统的三维沉浸视频，3D 全真视频具有以下特点：

照片级真实的渲染：3D 全真视频通过高精度摄像机和先进的渲染技术实现照片级渲染。视频中的每个细节都被真实地呈现，使观众仿佛亲临其境，获得一种极为逼真的观看体验。

6DoF 交互：3D 全真视频支持六自由度（6DoF）交互，允许观众视角在三维空间内自由移动和旋转，从而全方位探索视频内容，定制化观看体验。6DoF 交互大大增强了沉浸感，使观众能够以更加自然和灵活的方式与虚拟环境进行互动，提升了观看体验的沉浸度和互动性，也为教育、培训、游戏等领域的应用提供了更多可能性。

高效压缩与实时处理：为了应对高分辨率、高帧率和高细节的要求，应当考虑为 3D 全真视频制定更先进的压缩标准。这些技术可以在保证图像质量的同时，降低数据传输和存储的负担，并能够实现处理和播放。高效的压缩算法和优化的实时渲染技术确保了流畅的用户体验，同时保持了高水平的视觉效果。

易于编辑：提供便捷的编辑功能，在创建和调整三维沉浸视频内容时，用户能够高效灵活的进行修改，满足不同的创作和定制化需求。

跨平台和互操作性：可以实现更多的跨平台整合，使用户能够在不同设备上无缝享受沉浸的观影体验无论是 PC、游戏主机、移动终端还是 VR 头显。由此，行业需要制定更多的标准和规范，以促进不同设备之间的互操作性和兼容性，从而为用户提供更加统一和一致的体验。

2. 3D 全真视频应用效果

3D 全真视频展现出广泛的应用前景。在娱乐与媒体领域，它可以为用户提供了沉浸式的观影体验，例如虚拟现实电影和互动游戏。在教育与培训中领域，它可以通过逼真的 3D 场景提升学习效果，特别适合医学和工程等专业培训。在旅游与文化遗产领域，它能够让观众足不出户探索名胜古迹和历史场景，增强文化体验等。

下图展示了 3D 全真视频在赛事直播中初步应用的实例。在该应用中，观众可以通过触摸滑动或者手势识别操作，360 度观看体育赛事，并且可以自主选择任意位置任意视角，享受全方位的沉浸观赛体验。

图 41 3D 全真视频不同视点下观看同一场景

该应用具备如下特征：

画面真实生动流畅：在 3D 全真视频中，画面不仅具有高度的真实感，还保持了生动和流畅的视觉效果。通过高精度摄像机采集和 AI 渲染技术，每一个细节都被真实呈现，提供身临其境的视觉体验。

自由选择视点：在 3D 全真视频中，观众可以享受 6DoF 的沉浸观赛体验。用户可以随意切换视角，无论是环绕视角还是特定角度，以获得更加个性化和沉浸的观看体验。这种多视角的互动性使得观众能够全方位感受视频内容。

热点区域智能导播：在 3D 全真视频中，系统能够根据观众的关注点和热点区域进行智能导播。当某一区域的关注度较高时，视频会自动调整视角和焦点，使得观众无需手动操作即可观看到最有趣或最重要的部分。这种智能导播功能提升了观看的便捷性和体验感。

由此，3D 全真视频不仅提供了更高质量的视觉享受，还可以通过交互性和智能化的功能提升了用户体验，使其在各类场景中都具有广泛的应用潜力。未来，随着技术的成熟和普及，3D 全真视频有望在教育、娱乐、医疗、建筑等领域得到更广泛的应用，为人们带来更加丰富和身临其境的体验。

4 标准化建议

4.1 三维沉浸视频标准

1.国内标准

为发挥标准在产业发展中的推动作用，国内标准化技术组织机构在三维沉浸视频的相关领域和方向不断推动标准的制定与发布。采集重建方面，国内主要有全国信息技术标委会的 SC24（计算机图形图像和环境数据表示分委会），编码渲染等相关标准则主要通过 SC29（多媒体编码分委会），终端显示和质量评价主要通过 SAC TC242（音频、视频及多媒体系统与设备标委会）。AVS(数字音视频编解码技术标准工作组）、UWA（世界超高清视频产业联盟）、虚拟现实产业联盟、虚拟现实产业推进会等组织也在大力推进标准制定。

2024 年 5 月 28 日，国家市场监督管理总局（国家标准化管理委员会）批准了 AVS 虚拟现实国家标准《信息技术 虚拟现实内容表达 第 2 部分：视频》。标准规定了虚拟现实全景视频和自由视角视频的编码表示与重建方法，包括压缩域的语法、语义以及重建过程，以及与平面视频编码标准的接口。适用于虚拟现实视频内容制作、播出、传输等应用。其余相关标准如下：

序号	标准号	标准名称	技术组织
1	GB/T 38665.1-2020	信息技术 手势交互系统 第 1 部分：通用技术要求	SAC/TC28/SC24
2	GB/T 38665.2-2020	信息技术 手势交互系统 第 2 部分：系统外部接口	
3	GB/T 38247-2019	信息技术 增强现实术语	
4	GB/T 38258-2019	信息技术 虚拟现实应用软件基本要求和测试方法	
5	GB/T 36341.1-2018	信息技术 形状建模信息表示 第 1 部分：框架和基本组件	

6	GB/T 36341.2-2018	信息技术 形状建模信息表示 第 2 部分：特征约束	
7	GB/T 36341.3-2018	信息技术 形状建模信息表示 第 3 部分：流式传输	
8	GB/T 36341.4-2018	信息技术 形状建模信息表示 第 4 部分：存储格式	
9	GB/T 28170.1-2011	信息技术 计算机图形和图像处理 可扩展三维组件（X3D） 第 1 部分：体系结构和基础组件	
10	GB/T 28170.2-2021	信息技术 计算机图形和图像处理 可扩展三维图形（X3D） 第 2 部分：场景访问接口	
11	20214280-T-469	信息技术 增强现实 软件构件规范	
12	GB/T 38259-2019	信息技术 虚拟现实头戴式显示设备通用规范	
13	20220593-T-469	信息技术 计算机图形图像处理和环境数据表示 混合与增强现实中实时人物肖像和实体的表示	
14	20190776-T-469	信息技术 虚拟现实内容表达 第 1 部分：系统	SAC/TC28/SC29
15	GB/T 44115.2-2024	信息技术 虚拟现实内容表达 第 2 部分：视频	
16	20214282-T-469	信息技术 虚拟现实内容表达 第 3 部分：音频	
17	2019-0205T-SJ	显示系统视觉舒适度 第 3-1 部分：头戴式显示 蓝光测量方法	SAC/TC242
18	20213183-T-339	虚拟/增强现实内容制作流程规范	
19	2017-0279T-SJ	虚拟现实音频主观评价	
20	2019-1104T-SJ	超高清虚拟现实显示设备通用规范	
21	GB/T 44220-2024	虚拟现实设备接口 定位设备	

2.国外标准

1）2012 年 7 月，ISO/IEC MPEG 和视频编码专家组（Video Coding Experts Group，VCEG）成立了一个新的小组 JCT-3V，负责开发下一代 3D 编码标准。JCT-3V 开发了两个 HEVC 的扩展，一个是 MV-HEVC，它被纳入到 HEVC 的第二个版本，该版本在 2014 年 10 月份完成。为了提升多视点编码的性能并支持更先进的 3D 显示设备，JCT-3V 推出了 3D-HEVC。3D-HEVC 被纳入 HEVC 标准的第三个版本，该版本于 2015 年 2 月最终确定。MV-HEVC 标准仅包含高层级语法（HLS）添加，因此可以使用现有的 HEVC 的解码核心进行实现。而 3D-HEVC 通过引入新的压缩工具，能够更有效的对视频加深度的格式进行压缩。

2）在第 142 届 MPEG 会议上，MPEG 工作组（WG 04）将其 MPEG 沉浸式视频（MIV）符合性测试和参考软件标准（ISO/IEC 23090-23）推进至最终草案国际标准（FDIS）阶段，这是标准批准过程的最终阶段。该文档规定了如何进行符合性测试，并提供了针对 ISO/IEC 23090-12 MPEG 沉浸式视频的参考编码器和解码器软件。草案中包含了 23 个经过验证的符合性比特流，以及基于 MPEG 沉浸式视频测试模型（TMIV）15.1.1 版本的编码和解码参考软件。

3）点云编码： 随着点云使用场景的增加，MPEG 3D 图形编码组（3DG）在 2017 年发布了征集提案（CFP），旨在向学术界和工业界寻求高效的点云压缩解决方案。根据对该 CFP 的回应，选择了两种不同的压缩技术用于点云压缩（PCC）标准化活动：基于视频的点云压缩（V-PCC）和基于几何的点云压缩（G-PCC）。前者主要针对动态物体点云的压缩，适用于密集点云，后者针对静态场景点云的压缩，适用于稀疏点云。这些标准和平台的研发将为点云数据的压缩和传输提供标准化和高效的解决方案，推动三维点云技术在各领域的应用和发展。

4）网格编码：MPEG 的 3D 图形和触觉（3DGH）编码小组于 2021 年 10 月发布了一项关于新的动态网格编码标准的提案征集。动态网格是网络连接性频繁变化的网格序列，因而如何减少其庞大的数据量仍是一个极大的挑战。苹果、Interdigital、诺基亚、腾讯和索尼等对这一提案做出了响应，而苹果的解决方案被采纳为 V-DMC 标准的基础版本。2023 年，基于该提案已经发布了基于视频的动态网格压缩测试模型（TMM）。TMM 的解决方案与编码器无关，但目前采用了 HEVC 标准的 HM 编码器用于视频编码，采用 Edgebraeker 的实现用于几何编码。

4.2 标准化建议

上述标准主要是面向传统的三维沉浸视频形态构建的，在用户体验、压缩效率和编解码复杂度等方面仍然不够理想，且对于 3D 全真视频并不适用。为了突破传统三维沉浸视频无法兼顾交互自由度和渲染真实性的瓶颈问题，需要探索面向新兴 3D 全真视频的紧凑表示和编解码标准。近年来以 NeRF 和 3DGS 为代表的隐式可微表示受到广泛关注，特别是在表征三维数据和场景重建方面，具有表示紧凑、交互自由度高、视点渲染真实性高等优势。基于隐式可微表示的各类研究成果不断涌现，例如高效静态场景三维重建、动态场景三维重建、3D AIGC、三维场景编辑与理解等等。这些成果为制定下一代新型三维沉浸视频编解码标准提供了新的思路和研究基础。

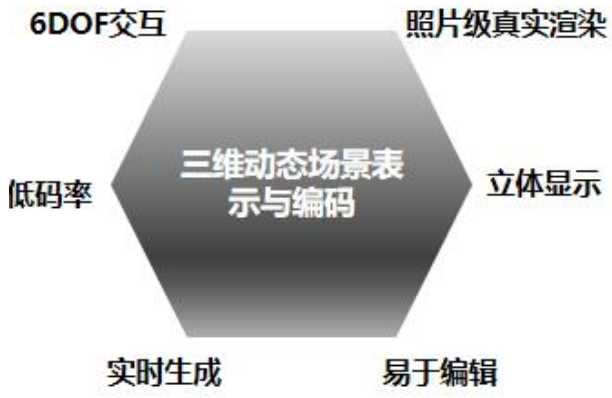


图 42 下一代三维沉浸视频编码标准需求

建议我国尽早在下一代三维沉浸视频编解码标准方面开展布局，新的标准应该具有以下六大特性：

1）支持六自由度用户交互

2）支持照片级真实的虚拟视点渲染

3）支持在各类移动终端上实现实时渲染和自由切换

4）编码效率明显高于现有三维视频编码标准

5）面向直播和实时通信场景，支持低延时实时编解码

6）视频内容易于编辑和处理

在这方面，我国 AVS 工作组 VRU 专题组于 2023 年 3 月开始了三维体视频智能编码的标准化需求探讨和技术证据征集工作，目前已经取得阶段性成果。

此外，有必要为三维沉浸视频从内容采集到终端显示整体路线制定对应标准，来为产业提供标准参考依据，一方面，帮助上下游企业打通链路，提高产业水平及发展速度；另一方面帮助跨平台企业达成一致，统一各项参数，提高三维沉浸视频在不同平台和设备之间的兼容性和互操作性。在各个阶段制定合适的质量评价标准也有利于提升三维沉浸视频内容整体质量，避免出现少数劣品破坏整个产业声誉的情况。

随着硬件技术不断突破，人工智能技术（AI）的不断发展，为三维沉浸视频的诞生与发展提供了良好的技术基础，并且作为元宇宙等未来高新技术产业的支柱技术，三维沉浸视频技术的发展重要性毋庸置疑，我国在多个政策性文件中也强调了要大力发展沉浸视频的决心。下一阶段需要统筹推进三维沉浸视频的标准化工作，各标准技术组织从内容采集、视频编码、终端显示以及质量评价的整体路线上共同推进三维沉浸视频相关标准的制定，并在关键技术节点加强标准化布局。

5. 附录

5.1 缩略语

下列术语和定义适用于本文件：

三维 （3D，3 dimensions）

人工智能 （AI，artificial intelligence）

中国篮球协会 （CBA，Chinese Basketball Association）

全方向立体投影 (ODS，Omnidirectional Stereo Projection)

飞行时间 （TOF，Time-of-Flight）

符号距离函数 （SDF，Signed Distance Funciton）

神经辐射场 （NeRF，Neural Radiance Field）

稀疏光束平差 （SBA，Sparse Bundle Adjustment）

运动恢复结构 （SFM，Structure From Motion）

尺度不变特征变换 （SIFT，Scale Invariant Feature Transform）

随机采样一致性方法 （RANSAC，Randam Sample Consensus）

多视立体 （MVS，Multiple View Stereo）

目前最优 （SOTA，State-of-the-Art）

基于动态直接线性变化法的拼接技术 （APAP，As-Projective–As-Possible with Moving DLT）

动态直接线性变换法 （Moving DLT，Moving Direct Linear Transformation）

国际标准化组织 （ISO，International Organization for Standardization）

国际电工委员会 （IEC，International Electrotechnical Commission）

国际电信联盟 （ITU，International Telecommunication Union）

数字音视频编解码技术标准工作组 （AVS，Audio Video coding Standard Workgroup of China）

开放媒体联盟 （AOM，Alliance for Open Media）

视频编码专家组 （VCEG，Video Coding Experts Group）

高性能视频编码 （HEVC，High Efficiency Video Coding）

3D 视频编码扩展开发 （JCT-3V，Joint Collaborative Team on 3D Video Coding Extension Development）

动态图像专家组 （MPEG，Moving Pictures Experts Group）

MPEG 沉浸视频标准 （MIV，The MPEG Immersive Video Standard）

自由度 （DOF，Degrees of Freedom）

通用视频编码标准 （VVC，Versatile Video Coding）

等距圆柱投影 （ERP，Equirectangular Projection）

基于视频的动态网格编码 （V-DMC，Video-based Dynamic Mesh Coding）

数字内容生成 （DCC，Digital Content Creation）

增强现实 （AR，Augmented Reality）

基于模型的渲染 （MBR，Model Based Rendering）

基于图像的渲染 （IBR，Image Based Rendering）

基于深度图的渲染 （DIBR，Depth Image Based Rendering）

反距离权重 （IDW，inverse Distance Weighting）

虚拟现实（VR，Virtual Reality）

混合现实 （MR，Mixed Reality）

视场角 （FOV，Field of View）

头戴式显示器 （HMD，Head-Mounted Display）

阴极射线管 （CRT，Cathode Ray Tube）

液晶显示器 （LCD，liquid-crystal display ）

发光二极管 （LED，Light-emitting Diode）

有机发光二极管 （OLED，Organic Light-Emitting Diode）

5.2三维沉浸视频应用

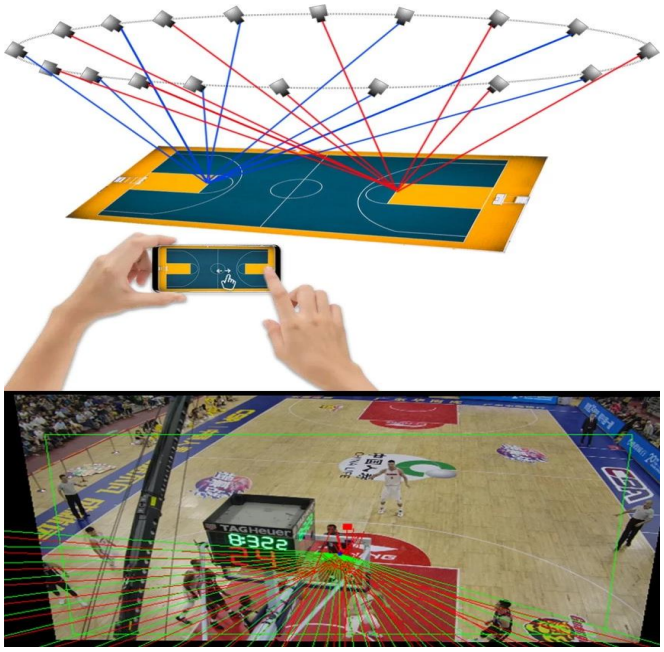
三维沉浸视频可以在虚拟环境下复刻真实场景的视觉信息，目前已初步应用在赛事直播、展示陈列、文旅场景、医疗场景、教育场景、工业场景等多个领域。随着技术的不断进步，三维沉浸视频将会获得更广泛和更深入的应用。

1. 赛事直播

传统赛事直播使用的画面都是单一视角的，是对平面视觉效果的表达，不能还原真实世界的立体性和多维性，且视角选择由导播决定，观众无法实时自由选择自己想要观看的视角。新一代人工智能技术的发展和云计算的普及，为视频处理技术的革新创造了条件。基于深度智能三维重建的自由视角视频系统，满足了多维度视觉体验，能够接近真实地再现时空影像，完全突破了传统视频单一视角的局限性，赋予了视频任意视角、自由交互等特点，取得了视觉效果的革命性进步。例如，咪咕在 CBA 联赛直播中推出的原画级自由视角直播应用，用户可以通过触屏滑动，更自由地选取自己想要的观赛视角，并能够基于屏幕焦点实现 360 度的观赛体验。

此外，通过对采用自由视角拍摄方式的视频进行特定的加工处理，如暂停、旋转切换视角、慢动作、放大等生成的特效回放视频。子弹时间特效视频可以为精彩动作提供更加细致、生动、多视角的观看效果，突

出体现了拍摄目标的惊、奇、特等震撼效果，极大提高了视频的观赏性。目前，子弹时间特效已经应用到北京冬奥会、巴黎奥运会等多个大型体育赛事中。



来源：咪咕文化科技有限公司解决方案

图 43 赛事直播自由视角技术

2.家庭场景

尽管传统的 3D 电影已经进入千家万户，但需要佩戴特殊眼镜观看仍然是一些观众的负担。在家庭观影中，有限的 3D 片源和单一的观看视角也限制了 3D 视频的体验。

随着计算机视觉和人工智能技术的不断发展，2D 转 3D 视频技术迅速成熟。借助传统方法或者 AI 模型，我们能够快速、大规模地将 2D 视频转换为 3D，从一定程度上解决了 3D 视频片源不足的问题。而 VR 头显的更新换代和裸眼 3D 设备小型化则为家庭场景下观看 3 维沉浸视频提供了终端设备，典型的产品如苹果的 Vision Pro 以及中兴的 nubia Pad 3D 平板。

2024 年 2 月 2 日，苹果正式发布了 Apple Vision Pro 这一创新性的产品，受到了市场的热烈欢迎，引起了全球的广泛关注。这是苹果在空间计算领域的一次重大尝试，将数字内容与现实世界融合，为用户带来全新的沉浸式体验。Vision Pro 的无边际画布突破了传统显示屏的限制，利用眼睛、手势和语音等交互方式实现了 3D 交互，使用户能够身临其境地体验数字内容。随着元宇宙概念的兴起和虚拟现实技术的不断发展，市场对这类设备的需求日益增加，Vision Pro 有望满足人们对沉浸式数字体验的追求。这一产品的成功推出也预示着苹果在空间计算领域的未来发展潜力，并为行业带来新的探索方向。

2023 年 6 月 27 日，中兴通讯的全球首款 AI 裸眼 3D 平板电脑 nubia Pad 3D 在国内预售。这款平板采用独特的 3D 光场技术带来沉浸式的裸眼 3D 体验。通过 AI 人眼追踪技术实现 86 度超宽可视角度，通过高精度纳米薄膜提供 8 视场角方案，观看距离灵活，3D 分辨率更高。通过咪咕视频 3D 内容，实现了电影、电视剧、综艺、电台、体育、游戏直播等全方面内容的 3D 播放。为用户带来全新的、超越想象的沉浸式视频体验。

来源：中兴官网

图 44 裸眼 3D 平板

3. 展示陈列

三维沉浸视频在展示陈列场馆的应用主要涵盖艺术馆、展览馆、博物馆、企业展厅等场所，通过不同的应用方式为参观者提供更加沉浸式和生动的展示体验。

全息透明方柜/柱柜是一款互动透明展示的高科技展柜。这款透明展示柜采用结合实物共同展示的互动形式，透明屏带有触摸互动功能，通过屏上的画面配合后方的实物，两者相结合，以图文及视频方式展示，极大的丰富了现场产品的展示内容和展览效果。

来源： 洲明科技公众号 / 洲明科技解决方案

https://www.hangjianet.com/topic/1713249686519?key=%E5%85%A8%E6%81%AF

图 45 全息透明柜产品展示

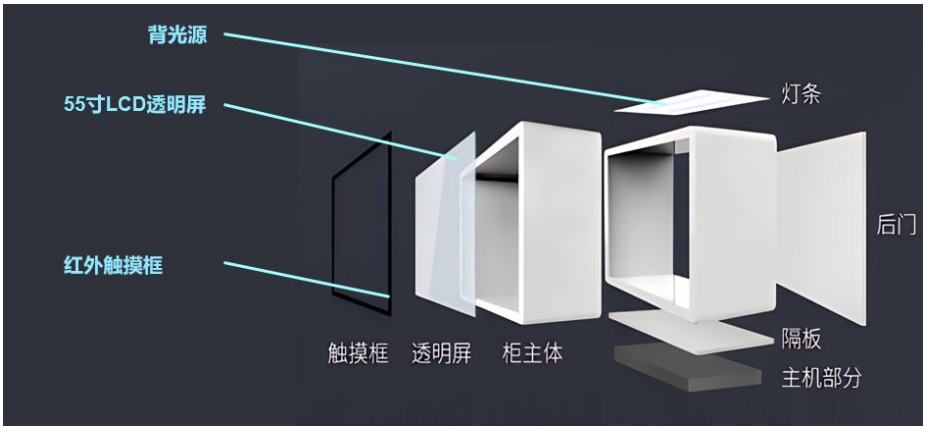
全息透明方柜/柱柜提供了一系列标准尺寸选项，满足不同场合的需求。客户可以根据液晶屏的标准尺寸，从多种尺寸规格中选择满足自己需求的产品。展示柜拥有高度的定制性，无论带触摸功能的互动型，还是偏好简洁的非触摸类型，都有对应的产品选型。

在柜体结构设计上，全息透明方柜/柱柜拥有了前开门和后开门两种方式，适应不同的安装和使用环境。一部分全息柜供应商额外提供附加定制服务，包括用户界面设计、主框架结构搭建、动画特效制作、素材修改美化及素材录入封装测试等，确保产品完全符合客户的个性化需求。

来源： 洲明科技解决方案

图 46 全息柜三维沉浸视频应用

透明互动展示柜不需要外部投影的协助即可实现直接独立显像，省电环保，其耗电量大约只有普通液晶显示屏的十分之一，能够在柜体内展示实物，同时在屏幕上发布显示物品的相关信息，如工艺结构、艺术价值和功能特性等，使得参照物品的信息更为清晰，为用户提供了一个直观、互动的展示平台。使得该产品很好的适配展厅展陈以及商超展示等场景。

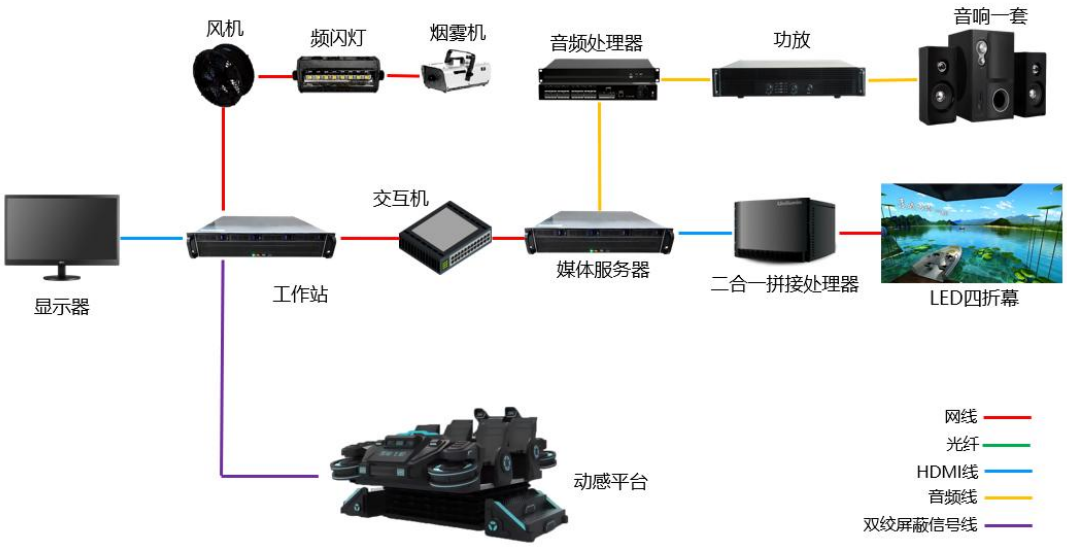


来源： 洲明科技解决方案

图 47 某全息柜硬件结构组成

4. 文旅场景

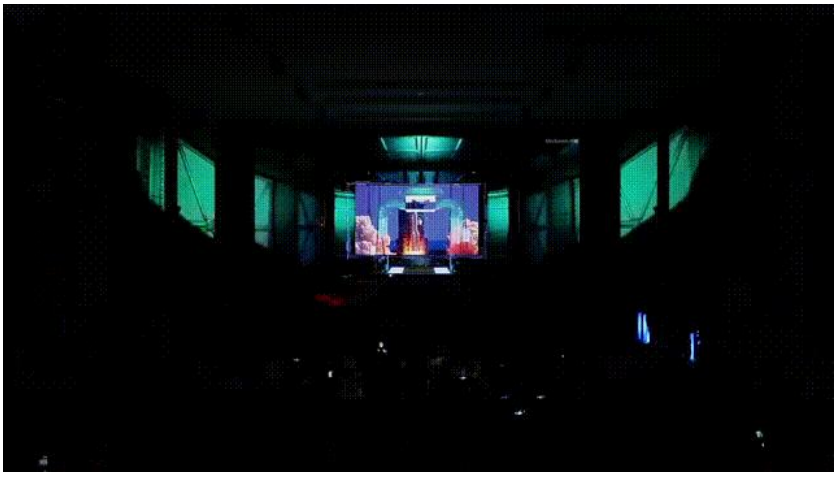
LED CAVE 沉浸式 5D 影院颠覆了传统的观影方式，它是一种结合了三维立体电影和多维环境效果的先进影院形式，让观众从观影者变成影片中的角色参与者。它拥有强烈的沉浸感，超现实的视觉效果与逼真特效完美同步，使观众全身心融入剧情；配备精确的动感技术和高精度模拟控制，动感平台动作细腻柔和，与影片无缝融合；设备兼容性强，支持主流放映设备和影片格式；同时具备严格的安全措施，如座椅占用和安全带识别功能，确保安全；此外，其系统结构和设备选型保证了系统的稳定性和可靠性。



来源： 洲明科技解决方案

图 48 5D 影院系统拓扑图

在 CAVE 沉浸式 5D 影院的环境下，艺术的展示已经远远超越了传统画廊空间的限制。借此可以探索艺术作品如何藉由全方位的视听体验，在沉浸式环境中与观众建立更深层的情感连接。通过将艺术作品的多维感官元素与高度沉浸的空间设计相结合，观众被引导进入一个多感官的艺术世界，其中包括但不限于：绘画中丰富的色彩层次、雕塑的立体质感，以及摄影作品中那些精致而永恒的瞬间。



来源： 洲明科技 公众号

https://www.hangjianet.com/topic/1703572655105

图 49 CAVE 沉浸式 5D 影院案例

裸眼 3D 户外大屏多应用于商圈和文旅场景。近年来，国家出台相关政策拉动内需消费，促进夜间经济发展，文旅夜游是行业内热度很高的关注点。夜游项目作为文旅系统中的一个板块，也赋予了文旅项目的属性特点，就是需要考虑商业运营和投资回报。裸眼 3D 作为文旅夜游系统解决方案中的局部板块，其投资回报性也在逐渐增强。

传统 3D 户外屏因需佩戴专业 3D 眼镜，常引起视觉疲劳和不适，并增加硬件和人工成本，难以吸引观众。裸眼 3D 显示屏无需眼镜观看，提供更自然的视觉体验，具备多重商业价值：如降低广告位空置率、打造地标品牌案例、推动品牌传播，同时吸引人气、带动线下消费，并响应政策推动地方经济发展。

目前市面上裸眼 3D 显示屏主要有 2 类：

1）弧形方案：屏体拐角处采用弧形过渡，显示效果整体性强，3D 感更真实，过渡自然。

来源： 洲明科技公众号 九江·新旅浔阳里 1723 项目
https://www.163.com/dy/article/GIARJ15U0538VQMO.html

图 50 模组弧形过渡

2）直角方案：屏体拐角处采用箱体切角，直角拼接，一般会有直观的拼缝。

来源： 洲明科技解决方案

图 51 常规箱体切角

5. 医疗场景

在医疗教育与培训方面，三维沉浸视频技术为医学生和医生们提供了一种全新的学习方式。相比于传统的平面视频，观看三维沉浸视频可以多角度、更全面、更深入的了解人体解剖结构、疾病发展过程等重要医学知识，提高对医学知识的理解和记忆。如下图，使用创维 MR 头显，在真实环境下沉浸式学习生物医学知识。

图 52 VR 医学知识教学

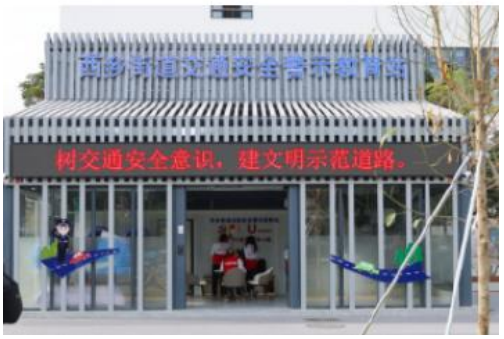
在远程诊断方面，传统的远程医疗借助平面视频对患者进行观察和会诊，存在视角和信息传递有限的问题。三维沉浸视频可以采集和传递更丰富的视觉信息，医生通过多角度观察患者的身体部位和症状，更清晰的查看皮肤病变、肢体运动等细节，医患沟通更加自然，可以给予患者更好的治疗建议。

对于医学手术，医生可以利用三维沉浸视频进行手术规划、风险评估和实时指导，从而提高手术的准确性和安全性。这种虚拟手术技术为医生提供了一个先进的工具，使他们能够在真实手术之前进行充分准备。在远程手术过程中，通过观察手术过程的多角度立体图像，经验丰富的医生可以给予手术团队实时指导和建议，确保手术的安全和成功。

目前，全国已有多所医院建立了远程医疗中心，有力的满足了偏远地区对优质医疗资源的需求，以及基层患者和医疗机构对于上级医院医疗支援的需求，三维沉浸视频在医疗领域的运用具有巨大的空间和前景。

6. 教育场景

当前，全国多地已有 3D 沉浸式安全教育体验馆落地，例如消防安全体验馆、消防安全教育基地、交通安全体验馆等。通过真实的 3D 模拟环境，观看者可以身临其境地体验各种安全情境，从而提高对安全知识的理解和记忆，大大增强警示作用。例如在 2024 年 1 月，深圳西乡街道已有首个 VR 交通安全警示教育站落地，运用 VR 等技术，直观的让参观者体验交通事故场景，深刻感受交通安全的重要性。



来源：深圳宝安区西乡街道办事处工作动态

图 53 交通安全体验站

此外，三维沉浸视频在教育培训领域还有许多潜在的应用，例如：

通过三维沉浸视频，学生可以无需真实的实验设备，直观的观察和学习复杂的实验过程。这可以提供更安全、更经济的实验环境。

教师可以利用三维沉浸视频拍摄的各种场景，如历史场景、地理环境、自然生态系统等，让学生身临其境地体验地理特征或自然现象，从而提升学习的效果和吸引力。

在职业培训中，三维沉浸视频可以用于展示各种实际场景，例如医疗手术、飞行驾驶、工程施工等。相较于平面视频，学员不仅可以自由选择观看视角，还可以观察到立体效果，创造出和真实环境相近的学习的体验。

教师使用三维沉浸视频进行远程教育，学生利用三维沉浸视频进行团队合作和沟通，通过不同角度观察聊天对象，能够产生一种面对面沟通的效果，使得互动方式更加真实自然。

7. 工业场景

工业设备的装配和维修过程需要操作人员具备一定的技能和经验。通过三维沉浸视频技术，可以制作装配培训和维修指导视频，多角度立体的展示设备的组装过程、零件结构以及常见故障的识别和排除方法，帮助操作人员快速掌握操作技巧并提高工作效率。如下图，使用创维 MR 头显，在真实环境中进行工业机械的拆解安装教学。通过三维体视频，专业人员不仅可以进行远程指导，也可以通过机械臂远程操作，直接进行设备装配和维修。

工业生产中的设备监控和维护是确保生产线正常运行的关键环节。通过三维沉浸视频技术，可以实现设备的远程监控和远程维护，工程师可以通过虚拟现实技术远程查看设备状态、诊断故障并进行维修操作，减少现场维护的人力成本和时间成本。

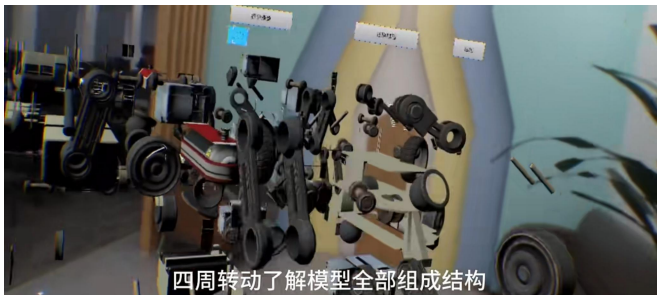


图 54 VR 工业机械拆解安装教学

8. 影视场景

因为三维沉浸视频多视点高自由度的特性，在影视制作中有许多创新应用，让影视虚拟制作得以快速高效进行内容生产，创造出更为震撼的视觉效果。三维沉浸视频作为影视拍摄的虚拟背景，不仅画面效果真实调整灵活，还能让演员自己置身于虚拟场景之中，提供一种身临其境的表演环境和即见即所得的拍摄方式。三维沉浸式视频在虚拟制作中应用场景有电影、电视、短剧、新闻报道、音乐剧、情景剧、纪录片拍摄等。

三维沉浸视频在影视制作应用中降低了成本， 减少外景拍摄，减少实体场景搭建、道具制作和运输等方面的费用。实时的视觉反馈让拍摄团队能够及时发现和解决问题，保证画面质量，减少后期制作时间。虚拟环境不受物理限制，创作者可以自由地构建任何想象中的场景。对于涉及危险动作或难以实现的场景，可以安全地在虚拟环境完成拍摄，减少危险拍摄， 减少人群聚集风险。

图 55 XR 虚拟制作

5.3 引用

[1] Xie J, Girshick R, Farhadi A. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks[C]//Computer Vision–ECCV 2016: 14th European Conference,

Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016: 842-857.

[2] Shih M L, Su S Y, Kopf J, et al. 3d photography using context-aware layered depth inpainting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8028-8038.

[3] Schonberger J L, Frahm J M. Structure-from-motion revisited[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4104-4113.

[4] Furukawa Y, Hernández C. Multi-view stereo: A tutorial[J]. Foundations and Trends® in Computer Graphics and Vision, 2015, 9(1-2): 1-148.

[5] Yao Y, Luo Z, Li S, et al. Mvsnet: Depth inference for unstructured multi-view stereo[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 767-783.

[6] Gutierrez J, Perez P, Orduna M, et al. Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360 Videos: ITU-T Rec. P. 919[J]. IEEE transactions on multimedia, 2021, 24: 3087-3100.

[7] Xu J, Lin C, Zhou W, et al. Subjective quality assessment of stereoscopic omnidirectional image[C]//Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19. Springer International Publishing, 2018: 589-599.

[8] Sun Y, Lu A, Yu L. Weighted-to-spherically-uniform quality evaluation for omnidirectional video[J]. IEEE signal processing letters, 2017, 24(9): 1408-1412.

[9] Zakharchenko V, Choi K P, Park J H. Quality metric for spherical panoramic video[C]//Optics and Photonics for Information Processing X. SPIE, 2016, 9970: 57-65.

[10] Yu M, Lakshman H, Girod B. A framework to evaluate omnidirectional video coding schemes[C]//2015 IEEE international symposium on mixed and augmented reality. IEEE, 2015: 31-36

[11] Sun W, Min X, Zhai G, et al. MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 14(1): 64-77.

[12] Yang J, Liu T, Jiang B, et al. 3D panoramic virtual reality video quality assessment based on 3D convolutional neural networks[J]. IEEE Access, 2018, 6: 38669-38682.

[13] Chen Z, Xu J, Lin C, et al. Stereoscopic omnidirectional image quality assessment based on predictive coding theory[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(1): 103-117.

[14]Qualinet. "IRCCyN/IVC Image Quality Database. "https://qualinet.github.io/databases/image/irccynivc_image_quality_database/

[15]University of Waterloo. "Waterloo IVC 3D Image Quality Database." https://ivc.uwaterloo.ca/database/3DIQA.html

[16]Wang, Jiheng. "Waterloo IVC 3D Video Quality Database." https://sites.google.com/view/wangjiheng/databases/waterloo-ivc-3d-video-quality-database

[17]Duan, Huiyu, et al. "IVQAD 2017: An immersive video quality assessment database." 2017 International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, 2017.

[18]University of Texas. "LIVE 3D Image Quality Database." http://live.ece.utexas.edu/research/quality/live_3dimage.html

[19]University of Southern California. "MCL 3D Database."https://mcl.usc.edu/mcl-3d-database/

[20]VCL. "3D60: A Dataset for 360° Images in 3D."https://vcl3d.github.io/3D60/

[21]Facebook. "Facebook 360 Depth Estimation."https://github.com/facebook/facebook360_dep

[22]Princeton University. "SUN360: A High-Quality 360° Database." https://3dvision.princeton.edu/projects/2012/SUN360/

[23]Schonberger J L, Frahm J M. Structure-from-motion revisited[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4104-4113.

[24] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[25]Kerbl, Bernhard, et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering." ACM Trans. Graph. 42.4 (2023): 139-1.

UHD World Association
世界超高清视频产业联盟

联系我们：
UWA联盟邮箱：support@theuwa.com
UWA联盟官网：www.theuwa.com