



**中国移动**  
China Mobile

# **全息通信技术白皮书**

**(2022 年)**

**中国移动通信集团有限公司**

## 前 言

随着当前移动网络的带宽和时延等关键性能的进一步提高以及未来 6G 网络的前景越来越明朗，可以依托于高性能网络的业务研究也进入了一个新的阶段。全息通信以其无以伦比的高真实性和沉浸感以及应用于元宇宙的前景成为其中一个被广泛关注的领域。

白皮书介绍了全息技术的概念和发展历程，梳理了全息通信的应用场景，对全息通信技术链条中的内容采集、算法处理、网络传输、渲染和显示等各重要技术环节进行了分析研究。结合当前全息通信产业现状，分析了相关的技术案例，并展望了全息通信技术的发展前景。

本白皮书的版权归中国移动所有，未经授权，任何单位或个人不得复制或拷贝本文之部分或全部内容。

# 目 录

|                               |    |
|-------------------------------|----|
| 1. 全息技术概述 .....               | 2  |
| 1.1 全息概念 .....                | 2  |
| 1.2 发展历程 .....                | 2  |
| 2. 全息技术在通信中的应用 .....          | 3  |
| 2.1 一对多场景：全息演讲或教学 .....       | 3  |
| 2.2 一对一场景：私人交流 .....          | 4  |
| 2.3 多对多场景：会议 .....            | 5  |
| 3. 全息通信关键技术 .....             | 6  |
| 3.1 内容采集 .....                | 6  |
| 3.1.1 彩色相机 .....              | 6  |
| 3.1.2 深度相机+彩色相机阵列 .....       | 6  |
| 3.2 算法处理 .....                | 7  |
| 3.2.1 非三维重建 .....             | 8  |
| 3.2.2 传统三维重建算法 .....          | 8  |
| 3.2.3 基于深度学习的三维重建 .....       | 12 |
| 3.3 传输 .....                  | 13 |
| 3.3.1 高带宽 .....               | 14 |
| 3.3.2 低时延 .....               | 14 |
| 3.3.3 强安全 .....               | 14 |
| 3.3.4 大算力 .....               | 15 |
| 3.4 渲染技术 .....                | 15 |
| 3.4.1 多视图立体渲染技术 .....         | 15 |
| 3.4.2 超多视点的虚拟立体内容渲染技术 .....   | 16 |
| 3.4.3 多平面图像渲染技术 .....         | 19 |
| 3.5 显示技术 .....                | 20 |
| 3.5.1 穿戴式设备 .....             | 20 |
| 3.5.2 裸眼 3D 显示设备 .....        | 21 |
| 4. 全息通信技术案例 .....             | 22 |
| 4.1 微软（MICROSOFT） .....       | 23 |
| 4.2 谷歌（GOOGLE） .....          | 23 |
| 4.3 螳螂慧视（MANTIS VISION） ..... | 23 |
| 5. 总结和展望 .....                | 24 |
| 6. 编写单位和作者 .....              | 25 |
| 缩略语列表 .....                   | 26 |
| 参考文献 .....                    | 27 |

## 1. 全息技术概述

### 1.1 全息概念

“全息”（Holography）即“全部信息”，这一概念是在 1947 年由英国匈牙利裔物理学家丹尼斯·盖伯首次提出，他也因此获得了 1971 年的诺贝尔物理学奖。全息技术是一种利用干涉和衍射原理来记录物体的反射，透射光波中的振幅相位信息进而再现物体真实三维图像的技术。它与物理学、计算机科学、电子通信及人机交互等学科领域有着密切的联系。

### 1.2 发展历程

自 1947 年由丹尼斯·盖伯发明了全息术，全息技术总共经历了三个大的阶段，分别是：传统光学全息（Traditional optical holography），数字全息（Digital holography）与计算全息（Computational Holography）。在每一个时期内，存在着一些分支技术的发展，以及全息术与各个领域结合产生的新技术。

**传统光学全息：**光学全息的全部过程分为信息数据采集与信息图像重构两个阶段，采集阶段相当于照相机的拍摄过程，而信息图像重构阶段相当于洗照片的过程。

**数字全息：**由于全息图只是对物体的物光束和参考光波进行相干叠加时产生的一些列干涉条纹进行了记录，而要得到物体的再现像，就必须对全息图进行重新处理，数字全息是利用电荷耦合器件来代替传统的光学记录材料来记录全息图，将物体的物光信息数字化记录，便于存储、数字处理以及重现。它最早是由 Goodman 在 1967 年提出的。

**计算全息：**计算全息最早是由 Kozma 和 Kelly 提出，但是限于当时计算机技术水平的不足，计算全息一直没有发展起来，直到 21 世纪初期数码照相机的普及和计算机技术的发展成熟才又进入了发展时期。计算全息是一种数字全息领域的分支，这种新型的方法是利用计算机去模拟物体的光场分布，用算法去进行全息图的制作，该方法可以不依赖实物，而是基于该物体的数学描述进行全息图制作，实现了全息术从实际物体到虚拟物体的突破。计算全息三维显示技术是近

年来将全息术、光电技术及计算机高速计算技术相结合发展起来的最具潜力的三维显示技术，与传统光学全息术相比具有灵活、可重复性好的特点。

## 2. 全息技术在通信中的应用

广义上说，全息通信业务是高沉浸、多维度交互应用场景数据的采集、编码、传输、渲染及显示的整体应用方案，包含了从数据采集到多维度感官数据还原的整个端到端过程，是一种高沉浸式、高自然度交互的业务形态。结合 6G 技术，进行扩展与挖掘可获得包括数字孪生、高质量全息、沉浸 XR、新型智慧城市、全域应急通信抢险、智能工厂、网联机器人等相关全息通信场景与业务形态，体现“人-机-物-境”的完美协作。

本文所涉及的内容为狭义上的通信概念，即指人与人之间通过某种媒介进行的信息交流与传递。

目前，远程通信用户面临的痛点主要为：语音通话、视频通话存在着临场感差和交互通道单一等弊端；受限于通信网络性能，视频通话常存在网络波动影响通讯质量等问题；传输高质量的视觉通讯内容受制于传输带宽而难以实现的问题。其中全息通信主要解决第一个问题，而诸如 6G 等高性能网络主要解决后两个问题，赋能全息通信应用。

基于全息通信具有真实度高、参与感强和沉浸感佳的特点，全息通信可以应用于以下三类场景：一对多场景、一对一场景和多对多场景。

### 2.1 一对多场景：全息演讲或教学

当前，远程演讲或教学越来越多的应用于现实生活，重要信息的传播可以不受地域限制。相较于传统的通信方式，全息的高真实性特点使受众专注度大为提升，学习效果进一步贴近线下演讲或教学。

此应用场景具有如下特点：信息流重要程度通常不对等，下行流重要性（演讲者或授课者的信息）大于上行流（受众的反馈），信息流呈现辐射状。

基于以上特点，初期的业务端到端解决方案可采用下行全息显示、上行高清显示的方式，利于在全息技术和 6G 等高速网络技术发展的初期部署。业务场景

如图 1 所示：

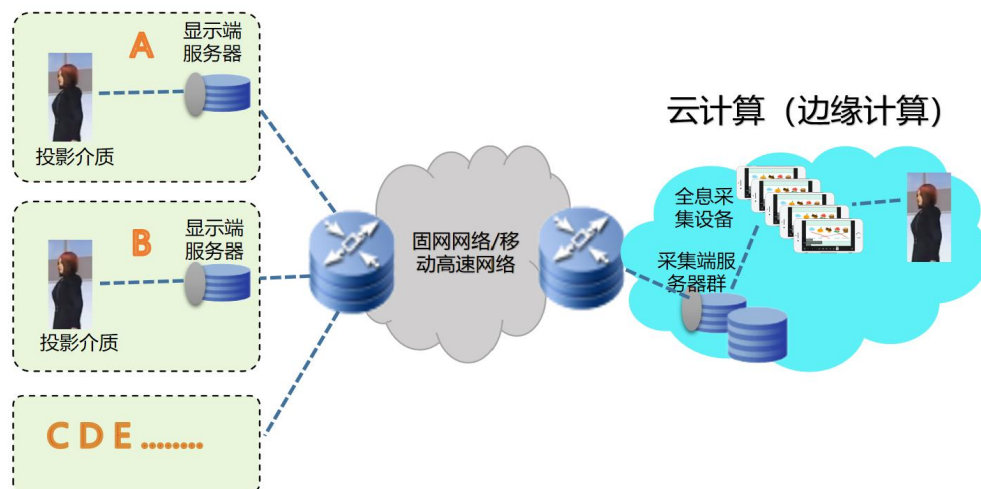


图 1 一对多场景

## 2.2 一对一场景：私人交流

随着网络的发展，从语音交流到视频交流，人们对于通信中的沉浸感要求越来越高，远在异地的亲朋好友可以通过全息通信获得近似面对面交流的体验，真正做到天涯若比邻。

此应用场景具有如下特点：信息流重要程度对等，双向都需要全息显示，通信模式为点对点通信。

基于以上特点，业务端到端解决方案需要采用对称的双向全息模式，每个用户既是被采集者同时也是接受者，此种模式对网络带宽的需求较高。业务场景如图 2 所示：

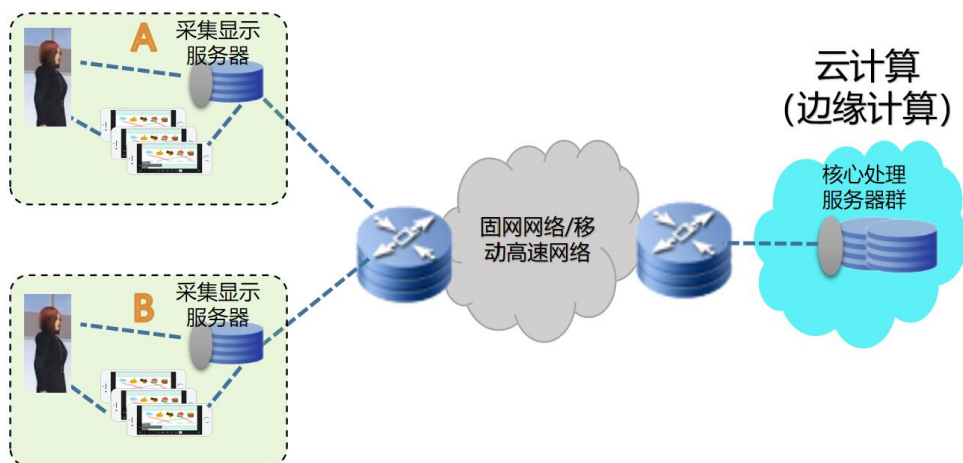


图2 一对一场景

### 2.3 多对多场景：会议

当全息技术和大带宽网络技术发展到一定阶段，可以构筑高质量的互动通信。在视频会议这一场景中，线上参会人员的人物数据将会被实时采集，通过全息显示技术构建高真实度的参会场景，实现身临其境般的线上会议感受。

此应用场景具有如下特点：信息流重要程度对等，每个人的面前都需要显示所有其他人的全息影像和声音，是一对一场景的复杂形式。

基于以上特点，业务端到端解决方案中每个用户既是被采集者也是接受者，作为接受者时，同时获取来自其他用户的全息影像和声音。此种模式对网络带宽的需求很高。业务场景如图3所示：

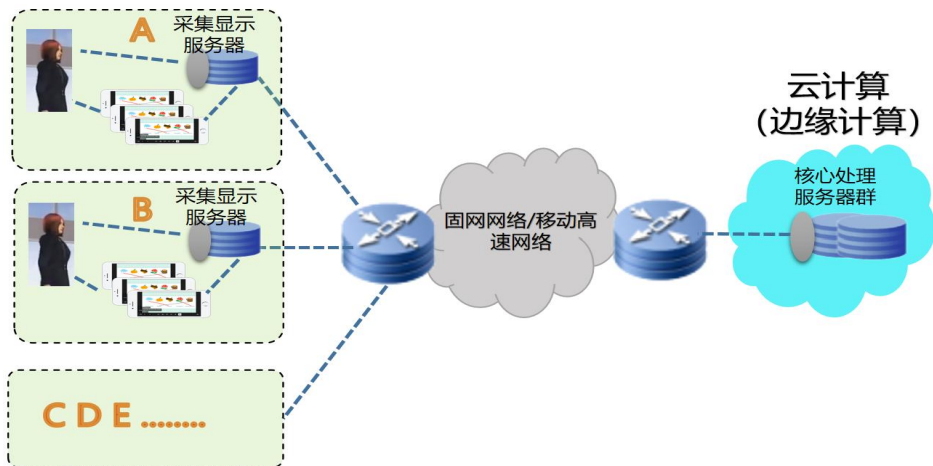


图3 多对多场景

### 3. 全息通信关键技术

全息通信的关键技术包括内容采集、算法处理、传输、渲染和显示。

#### 3.1 内容采集

全息通信所需的动态三维内容又称作“体积视频”（Volumetric Video），其采集方式可以分为纯彩色相机阵列采集和深度相机+彩色相机阵列采集。

##### 3.1.1 彩色相机

用几十甚至上百个彩色相机从多个角度捕捉人像和其动作，为了后期方便数据提取，通常会在周围布置绿幕。拍摄时，通过时间控制器控制相机阵列同步启动拍摄。

根据应用场景等不同，彩色相机阵列又可分为局部围绕式和 360° 围绕式。

当仅需采集单面人体时，可以搭建小于 180° 的相机阵列，仅用单反相机围成半圈甚至更小的范围。如果要采集人体 360° 全方位的数据，需要将相机阵列围成一圈，做成影棚的形态，这样可以同时采集人体各个角度的影像。

##### 3.1.2 深度相机+彩色相机阵列

相较于纯彩色相机阵列，目前市场上的主流做法是通过深度相机搭载彩色相机阵列来完成。和单纯用彩色相机相比，加上深度相机后，生成的人物三维数据更加精细，细节表现会更好。例如脸部的三维效果更明显，可以清晰看到鼻梁的高度、嘴唇的轮廓等细节。

另外，如图 4 所示，深度相机+彩色相机阵列的采集方式无需布置绿幕，对场地要求也比较灵活。

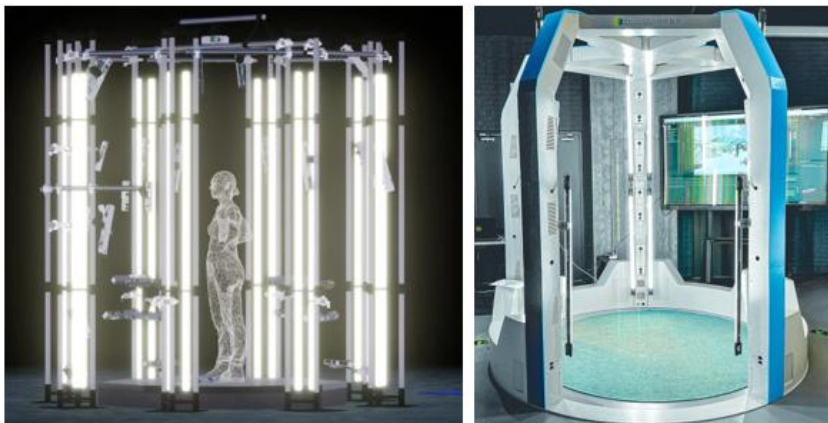


图 4 深度相机+彩色相机阵列

掩膜编码结构光深度相机包括一组深度相机（红外发射器+红外相机）和两个彩色相机，可以同时采集人物的深度信息和彩色信息，如图 5 所示。



图 5 掩膜编码结构光深度相机

### 3.2 算法处理

非三维重建处理主要指自由视点技术，自由视点技术对于不同的视角显示不同的图像，是一种相对“粗糙”的处理方式。

三维重建处理包括基于深度学习的三维重建和传统的三维重建。近年来，基于深度学习的三维重建算法的发展有雨后春笋之势，在某种程度上，它们预示着未来全息通信技术的发展方向——实时重建+减少对多相机的依赖，更加“轻便”、“快捷”。而传统三维重建方式比基于深度学习的三维重建更加稳定成熟，但也更依赖于硬件结构，如相机阵列等。当然，将深度学习与传统三维重建算法相结合，可以提高其性能和效果，这也是未来发展的可能方向之一。

### 3.2.1 非三维重建

自由视点技术一般采用此种方式处理，可以理解成多相机之间的“切换”，也就是切换成观看者想要看到的视角。当然，也会通过生成“虚拟视角”的方式以弥补相机的密集度不足。

“虚拟视角合成”是指利用已知的参考相机拍摄的图像合成出参考相机之间的虚拟相机位置拍摄的图像，这样能够获取更多视角下的图片，是让自由视点观看方式变得“自由”的关键。其合成方式为利用相邻两个相机成像上的差异——即视差图，在同一行上平移虚拟相机位置，从而生成新的视角图像。

假设相邻两个相机拍摄的图像像素点的视差值为 1，我们要生成两个相机正中间虚拟相机的视角，则可以将左边相机拍摄图像的像素点均向右移 0.5，或者将右边相机拍摄图像的像素点向左移动 0.5。以此类推。

合成虚拟视图既可以利用左参考图像和对应的左视差图，也可以利用右参考图像和对应的右视差图，更好的是都利用上得到两幅虚拟视点图像，然后做图像融合，比如基于距离的线性融合等。

### 3.2.2 传统三维重建算法

传统三维重建算法可分为两大类：纯彩色相机阵列的被动式和深度相机加彩色相机的主动式。

被动式三维重建算法是直接根据 2D 图片信息，不依靠发射信号，对物体进行重建。传统的被动式三维重建算法，如 SFM 主要是通过还原点云进行三维重建。SFM 是一种全自动相机标定离线算法，以一系列无序的图像集作为输入，估计出的相机参数矩阵和稀疏点云为输出。由于 SFM 算法得到的点云是稀疏的，因此需要再进行 MVS 算法对稀疏点云进行处理，转换为稠密点云。

主动式三维重建算法需要通过传感器对物体发射信号，然后通过解析返回的信号对物体进行重建。代表性的算法有结构光、TOF 等。其中，以红外结构光为例，依靠红外投射器将编码的红外光投射到被拍摄物体上，然后由红外相机进行拍摄，获取被拍摄物体上编码红外光的变化，将其转换为深度信息，进而获取物体三维轮廓；TOF 法通过投射器向目标连续发送光脉冲，然后依据传感器接收到返回光的时间或相位差来计算距离目标的距离。主动式算法如结构光法和 TOF

法能够精准构建 3D 模型，但二者都需要较为精密的传感器。

图 6 为一组三维重建的过程：先是生成密集点云，再由密集点云生成面片网格三维数据，最后贴上彩色照片。



图 6 三维重建过程

### 3.2.2.1 被动式三维重建算法 SFM

SFM, Structure from Motion, 顾名思义，用于从“动作”中重建 3D 结构，也就是从时间系列的 2D 图像中推算 3D 信息。

人的大脑可以从动的物体中取得其三维的信息，是因为大脑在动的 2D 图像中找到了匹配的地方，即重叠区域。然后通过匹配点之间的视差得到相对的深度信息，在这一点上，原理和基于双目视觉的三维重建相同。

SFM 的输入是一段动作或者一时间系列的 2D 图群，然后通过 2D 图之间的匹配可以推断出相机的各项参数。重叠点可以用 SIFT, SURF 来匹配，也可以用最新的 AKAZE（SIFT 的改进版）来匹配。

在 SFM 中，误匹配会造成较大的错误，所以要对匹配进行筛选，目前流行的方法是 RANSAC（Random Sample Consensus）。2D 的误匹配点可以应用 3D 的几何特征来进行排除。Bundler [2] 就是一种 SFM 的方法，Bundler 使用了基于 SIFT 的匹配算法，并且对匹配进行了过滤去噪处理。

### 3.2.2.2 被动式三维重建算法 MVS

SFM 的重建成果是稀疏三维点云，为了得到更好的深度结果，需要使用多视

角立体视觉（Multiple View Stereo, MVS）算法。某种意义上讲，SFM 其实和 MVS 是类似的，只是前者是摄像头运动，后者是多个摄像头视角。也可以说，前者可以在环境里面“穿行”，而后者更像在环境外“旁观”。

SFM 中我们用来做重建的点是由特征匹配提供的，这些图像特征的代表为图像中的一个小区域（即一堆相邻像素）。而 MVS 则几乎对照片中的每个像素点都进行匹配，几乎重建每一个像素点的三维坐标，这样得到的点的密集程度可以较接近图像为我们展示出的清晰度。

其实现的理论依据在于，多视图照片间对于拍摄到的相同的三维几何结构部分存在极线几何约束。

### 3.2.2.3 主动式三维重建算法 结构光算法

结构光（Structured Light）三维成像的硬件主要由相机和投射器组成，结构光就是通过投射器投射到被测物体表面的主动结构信息，如激光条纹、格雷码、正弦条纹等，然后通过单个或多个相机拍摄被测表面即得结构光图像，最后基于三角测量原理经过图像三维解析计算从而实现三维重建。

结构光技术就是使用提前设计好的具有特殊结构的图案（比如离散光斑、条纹光、编码结构光等），将图案投影到三维空间物体表面上，使用另外一个相机观察在三维物理表面成像的畸变情况。如果结构光图案投影在该物体表面是一个平面，那么观察到的成像中结构光的图案就和投影的图案类似，没有变形，只是根据距离远近产生一定的尺度变化。但是，如果物体表面不是平面，那么观察到的结构光图案就会因为物体表面不同的几何形状而产生不同的扭曲变形，而且根据距离的不同而不同，根据已知的结构光图案及观察到的变形，就能根据算法计算被测物的三维形状及深度信息。

结构光 3D 成像技术主要由 4 大部分组成：

1) 不可见光红外线发射模组（IR Projector）：用于发射经过特殊调制的不可见红外光至被拍摄物体；

2) 不可见光红外线接收模组（IR）：接收由被拍摄物体反射回来的不可见红外光；

3) 彩色相机模组（RGB）：采用普通彩色镜头模组，用于 2D 彩色图片拍摄；

4) 图像处理芯片（非必须，有些结构光供应商提供的解决方案可利用主机 CPU，如手机 AP 处理）：将红外相机拍摄得到的红外照片通过计算，得到被拍物体的深度信息。

以 iPhoneX 为例，用的是以色列 PrimeSense 公司的光编码技术（散斑结构光），如图 7 所示。这种结构光方案，通过投射人眼不可见的伪随机散斑红外光点到物体上。这些散斑投影在被拍摄物体上的大小和形状根据物体和相机的距离和方向而不同，由此计算得到被拍摄物体深度信息。

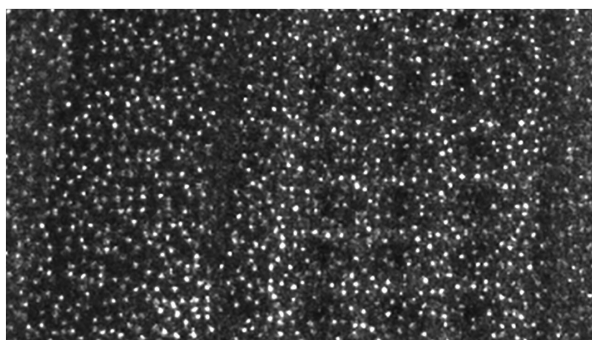


图 7 散斑结构光

#### 3.2.2.4 主动式三维重建算法 TOF 算法

TOF (Time of Flight) (光) 飞行时间，字面理解就是通过光的飞行时间来计算距离。

TOF 的基本原理是通过红外发射器发射调制过的光脉冲，遇到物体反射后，用接收器接收反射回来的光脉冲，并根据光脉冲的往返时间计算与物体之间的距离。由于光的速度快，这种调制方式对发射器和接收器的要求较高，对于时间的测量有极高的精度要求。

直接测量光飞行时间的 TOF 算法又叫 DTOF (Direct TOF)。在实际应用中，通常调制成脉冲波（一般是正弦波），当遇到障碍物发生漫反射，再通过特制的 CMOS 传感器接收反射的正弦波，这时波形已经产生了相位偏移，通过相位偏移可以计算物体到深度相机的距离。这种 TOF 算法又叫做 ITOF (Indirect TOF)。

如上所述，结构光和 TOF 均属主动式三维成像方式，但两者在工作距离、成像精度、应用场景等方面又有差异，比较如下：

|          | 工作距离                                        | 精度                                                 | 功耗 | 应用场景                                               |
|----------|---------------------------------------------|----------------------------------------------------|----|----------------------------------------------------|
| TOF      | 几米至十几米<br>（和光源功率成正比）                        | DTOF 在近<br>距离可达毫米<br>级（1-2 毫米）；<br>中远距离精度<br>比结构光好 | 高  | 无人机、机器人<br>避障；游戏类动作识<br>别、动作捕捉等                    |
| 结 构<br>光 | 消费级模组工<br>作距离 2 米以内；<br>工业级模组工作距<br>离可达 7 米 | 最高可达<br>亚毫米级（0.5<br>毫米）；近距离<br>精度比 TOF 好           | 中等 | 金融刷脸支付；<br>人体、物体 3D 建模；<br>空间 3D 建模；动作<br>识别、动作捕捉等 |

### 3.2.3 基于深度学习的三维重建

除了上述传统的被动和主动三维重建，利用深度学习模型对数据集的学习获取先验知识，再在少量图片的基础上进行重建，相比原先传统算法，可以大大减少对图片的依赖。

早期 Saxena 等提出了利用监督学习的办法去预测照片的像素对应的深度[3]。同样，ECCV2022 收录的来自 Niantic 和 UCL 等机构的研究者关于“没有 3D 卷积的 3D 重建方法[4]”则是基于前者的提升，无论从效果到性能均显著优于前者。

近期，华盛顿大学计算机学院的 GRAIL 图形和成像实验室发布了一项基于 NeRF 合成的新技术 HumanNeRF，该方案的最大特点就是利用 AI 算法将 2D 视频合成高保真 3D 全身模型[5]，该论文被收录在 CVPR2022。

#### 3.2.3.1 SimpleRecon：无 3D 卷积实时三维重建

从姿态图像重建 3D 室内场景通常分为两个阶段：图像深度估计，深度合并（Depth Merging）和表面重建（Surface Reconstruction）。过去的研究依赖于昂贵的 3D 卷积层，限制了其在资源受限环境中的应用。

来自 Niantic 和 UCL 等机构的研究者利用强大的图像先验以及平面扫描特征

量和几何损失，设计了一个 2D CNN。所提方法（SimpleRecon）在深度估计方面效果显著，更重要的是允许在线实时低内存重建，每帧仅用约 70ms。而实时三维重建正是全息通信的关键技术之一。

该研究的关键是将现有的元数据与典型的深度图像特征一起注入到代价体积（Cost Volume）中，以允许网络访问有用的信息，如几何和相对相机姿态信息。通过整合这些之前未开发的信息，该研究的模型能够在深度预测方面显著优于之前的方法，而无需昂贵的 3D 卷积层、复杂的时间融合以及高斯过程。

### 3.2.3.2 HumanNeRF：从 2D 视频提取动态人像，并转换为 3D 模型

NeRF 方法是 2020 年 ECCV 的论文提出的。仅仅过去不到 2 年，关于 NeRF 的论文数量已经十分可观。NeRF 是 Neural Radiance Fields 的缩写，中文译作神经辐射场，它是一种小型神经网络，可通过 2D 图片来学习 3D 建模和渲染。把 GRAIL 实验室的研究 HumanNeRF 提出来，是因为它和全息通信息息相关——人物三维重建。

HumanNeRF 解决了 3D 人像渲染的两大难题：神经网络渲染动态对象和对于多摄像头方案的依赖。此外还可学习人体 T 型姿态，并通过运动场来学习刚性骨骼运动和非刚性运动。运动场和姿态预测学习信息可根据 2D 视频中的姿态去修改 3D 模型，并在 NeRF 中渲染。当然，目前该技术还需继续优化，譬如环境光变化对结果的影响等。

HumanNeRF 方法将稀疏图像作为输入，在大型人类数据集上使用预先训练的网络，然后就可以从一个新的视角有效地合成一个照片级的真实感图像。通过一小时对特定数据的微调，即可生成改进后的结果。

## 3.3 传输

全息通信本身并不带来新的传输技术，但是由于三维显示带来的高真实性和沉浸感以及对实时性的需求，导致了对网络提出了更高的要求，主要表现为以下四个方面：高带宽、低时延、强安全和大算力。

### 3.3.1 高带宽

与传统高清或双目立体视频相比，全息通信传输的流媒体对网络带宽的需求将达数百 Mbps。例如一个包含 10 个摄像头传感器的全息通信系统，每个摄像头输出 1080P 彩色图像，每个像素有 32 位的彩色数据，输出分辨率为 512dpi × 424dpi 的深度图像，每个像素有 16 位的深度数据。按照 60 的帧率和 100 倍的压缩率计算，需要上行带宽约为 420Mbps。随着对图像精度的提升，传感器数量、视点数量和帧率也会随之增加，对网络带宽的要求将更高。目前实现全息采集传输显示的技术路线有多条，不同的技术方案所需要的网络带宽也不同，从几百 M 到几个 G。

使用更高效的图像压缩技术和编解码方案（例如 H.266），在一定程度上可以缓解全息通信的带宽需求，但仍需未来网络具有超高的带宽。对毫米波、太赫兹、可见光等更高工作频段的研究表明，未来网络可提供的用户体验速率可以有效的满足全息通信的带宽需求。

### 3.3.2 低时延

全息通信中的时延可以分为数据处理时延和网络传输时延。为了减少整体时延，需要处理节点具有高算力，并进一步缩减网络本身的传输时延。

全息通信的过程可描述为，首先通过采集端设备获取对象信息，计算处理后，经过编码压缩进行网络传输，在终端侧解码渲染并显示全息图像。获取真实度高的全息图像往往需要很高的算力，当前的主要矛盾集中在处理带来的时延往往直接带来了非实时性的感受，而实时性稍好的处理方式又往往导致真实感偏差。因此对于处理算法的优化研究是当前的热门方向。对于网络本身的传输时延，5G 端到端传输时延可以控制在 20ms 以内，随着未来网络的研究和部署，6G 网络的传输时延会进一步减少。

### 3.3.3 强安全

通过全息通信传输的数据中含有大量的信息数据，包括人脸特征、声音等敏感信息，需要网络提供绝对安全的保障，而现有安全技术的使用会增加端到端时

延。对时延和安全性的折中考虑是未来网络需要面对的难题之一。

### 3.3.4 大算力

由于全息通信包含的信息和数据量巨大，计算时间过长，除了会带来极大的带宽负担外，还会造成很大的 MTP 时延。随着云计算和 MEC 技术的快速发展，未来网络可通过云端和边缘端的快速部署解决全息通信的算力需求。

## 3.4 渲染技术

通过采集设备获取的图像数据经过算法处理后，生成的数据模型使用渲染技术在显示设备上展示。目前，全息技术涉及的渲染方法主要有多视图立体渲染技术、超多视点的虚拟立体内容渲染技术和多平面图像技术。

在以上三类渲染技术中，多视图立体渲染技术作为已经成熟的技术被广泛应用于 VR 商业市场，超多视点的虚拟立体内容渲染技术和多平面图像技术多应用于裸眼 3D 显示设备。基于全息通信的特点，人们更倾向于使用裸眼 3D 设备构成解决方案。

### 3.4.1 多视图立体渲染技术

多视图立体渲染技术主要用于虚拟现实（VR）设备的图像渲染。当图像通过虚拟现实眼镜等设备呈现在人眼前，设备呈现的画面质量直接决定用户的观看感受。在该类设备上，图形硬件厂商在提升画面视野，降低图形畸变，提高图形质量等方面不断努力，并推出一系列技术与解决方案。

#### 3.4.1.1 虚拟现实图形管道原理

图形应用程序为显示设备渲染一个 3D 场景时将在 3D 空间中创建一个虚拟摄像机并根据摄像机的位置对场景中的几何图形执行计算。渲染引擎执行像素阴影，并将单帧投影到显示设备上。虚拟现实的图形管道则不同，它需要渲染多个视图。一个典型的 VR 设备有两个镜头。每个镜头都会在观看者的左右眼中投射出一个单独的视图，即 3D 应用程序需要执行立体渲染——从轻微偏移的摄像机

位置渲染同一场景的两个视图。一种常见的立体渲染方法是，在将图像呈现给终端之前，要简单地按顺序一起执行两种绘制操作。这种方法对处理设备有一定的性能要求。

#### 3.4.1.2 单通道立体 (SPS) 渲染技术

NVIDIA 的 Pascal 架构引入了一种称为单通道立体 (SPS) 渲染的技术来帮助加速 VR 的几何处理。SPS 使 GPU 在一次渲染过程中最多同时绘制两个视图，这些视图只在 x 方向上变化。设备视场 (FOV) 越大，沉浸感越强。Pascal 的立体渲染技术完全适用于有限视场的共面显示设备，针对  $180^\circ + \text{FOV}$  的设备可以更逼真地展示虚拟现实效果。

#### 3.4.1.3 图灵多视图渲染技术

两个视角对于具有超侧视场的 VR 设备是不够的。由 NVIDIA 的图灵 GPU 支持的多视图渲染扩展了单通道立体渲染。MVR 支持在一次传递中最多渲染 4 个视图，支持每个视图顶点位置的不同组件，支持将其他通用属性设置为独立于视图的能力，使开发人员能够利用和扩展同时多投影算法到超宽的 FOV 设备与倾斜的显示器，使用最多 4 个视图。

#### 3.4.2 超多视点的虚拟立体内容渲染技术

目前市场上除了已经大规模普及的虚拟现实显示设备具备全息显示效果，还有一种无需人体佩戴的特殊设备——可直接观看的显示器，这类显示器一般具有特殊的光学结构，可以实现全息显示效果。针对该类显示器，同样从不同的角度设计了不同的渲染方式以提升设备的画面性能。

##### 3.4.2.1 逐视点渲染法

逐视点的渲染方法是指在虚拟空间中以一定规则摆放虚拟摄影机阵列，并逐个渲染出视点图像，最终合成全息编码图像，或者将每一个视点图像输出到相应的投影机上，多角度投影在相应的显示设备屏幕上。本方法既可以采用栅格化的

方法，也可以采用光线跟踪的方法。

前期的逐视点渲染法需要串行的执行。串行逐视点渲染法中间过程中生成的每个视点图像存在大量冗余，生成过程随着视点数目的增多而线性增长。一般应用在 100 个视点以内的，单视点图像分辨率小于  $640 \times 480$ ，场景面片数小于 10M 场景的实时渲染。

基于几何着色器的逐视点渲染法可一次性地在一个纹理上渲染出多个视点的图像，这样可以极大地节省多视点渲染时间。

基于 GPU instancing 的逐视点渲染法是一种用来提高渲染大量物体效率的技术[6]，随着场景品质需求的提升，需要在场景里绘制越来越多的物体，CPU 和 GPU 的压力都会上升。在场景中有大量重复物体需要绘制时，使用 GPU Instancing 技术只需要设置一次原始数据缓冲区，调用一次 drawcall 绘制出来。与基于几何着色器的逐视点渲染算法相比，它的显存需求不会随着视点数目的增加而增加。

#### 3.4.2.2 光线跟踪法

光线跟踪算法可以生成质量很高的全息图像，直接加载至二维显示面板或投影机即可显示。光线跟踪算法一般由三部分构成[10]，即光线的生成、光线的碰撞和像素的着色。传统显示设备的光线跟踪与全息显示设备的光线跟踪区别在于光线的生成部分。如图 8 所示，光线从虚拟摄影机出发经透镜到达基元图像像素，像素和光线满足一一对应的关系。通过显示器的像素生成光线，之后是碰撞检测，最后通过着色程序就可以生成相应的光场图像。

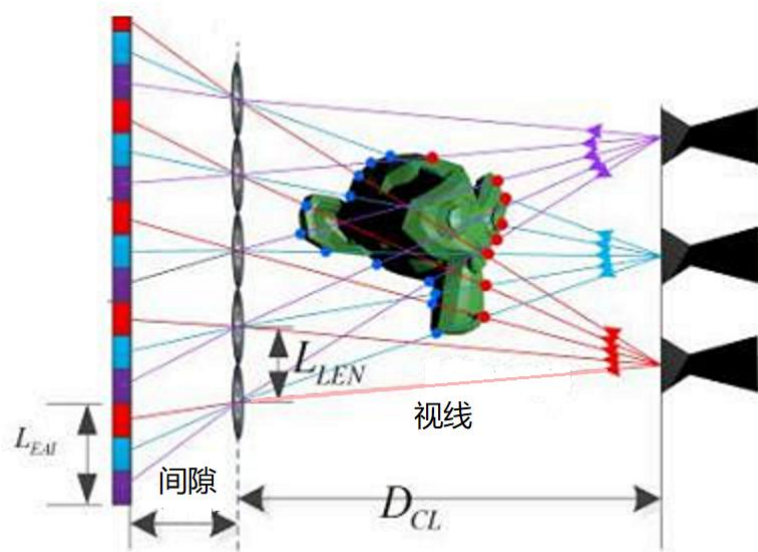


图 8 基于光线跟踪的集成成像算法

光线跟踪具有天然的并行性，可以很方便地提高光线跟踪的效率。光线跟踪的计算复杂度与屏幕的分辨率大小正相关，现有的实时光线跟踪硬件管线基本面向 2K 以下的显示设备且可绘制的场景有限[7]。

#### 3.4.2.3 基于深度信息的超多视点渲染

基于深度信息的渲染（Depth-Image-Based Rendering, DIBR）在虚拟场景的渲染过程中十分常用[8]。DIBR 算法是利用深度信息和其他附加信息通过插值产生其他视点的图像。它有效地降低了图形渲染的复杂度，渲染速度大大加快，缺点是造成渲染质量的下降。

根据参考视点的数目，可分两类：一类为单参考视点的 DIBR，一类为多参考视点的 DIBR[9]。单参考视点的 DIBR 可以只使用一幅深度参考图像和彩色图像就可以生成场景所需要的全部视点，当视角较大时空洞较大，填补困难，适用于  $10^\circ$  以内观看视角的光场显示设备。多参考视点的 DIBR 需要多个深度参考视点，能够有效地增大视角，消除空洞[10]。多参考视点一般使用左右两个视点来插值出中间视点。DIBR 技术具有带宽需求小、输入图像数量少和绘制速度快的优点。单参考视点的 DIBR 技术映射速度快，双参考视点的 DIBR 技术能够利用左右视图实现对遮挡区域的信息互补。

### 3.4.2.4 基于几何相关性的超多视点渲染

假设由三个点场景组成的三维场景，虚拟摄影机阵列为错切式排列，如图9所示，其对应的 EPI 图像为三条直线[11]。

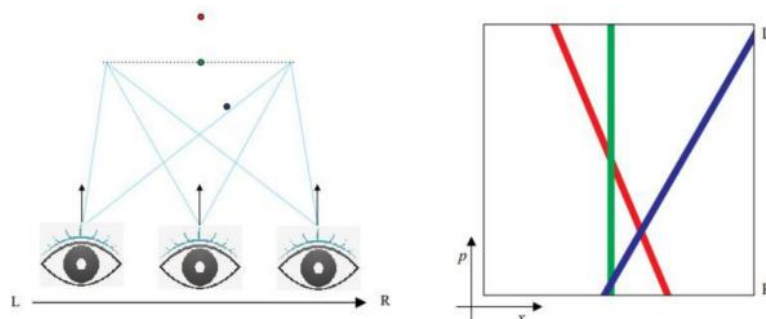


图9 虚拟点在错切排列下的采集以及对应的 EPI 图像

当点位于零平面时，所对应的 EPI 图像为竖直的直线；若点位于零平面内，则对应的 EPI 的图像是斜率为正的直线；若点位于零平面外，则对应的 EPI 图像是斜率为负的直线。因此，点的渲染就可以转化最左侧虚拟相机和最右侧相机对这一点的渲染，并在 EPI 图像上由这两视点生成相应的 EPI 直线，最终再转换为视点图像，这样就会大大简化渲染的流程。

### 3.4.3 多平面图像渲染技术

多平面图像渲染技术是一种基于图像渲染环境复杂真实场景的技术。例如在渲染具有遮挡或镜面反射等具有挑战性的复杂场景时，这种表示比传统的 3D 网格渲染更有效。多平面图像（multi-plane image, MPI）能够表示几何体和纹理（包括遮挡元素），并且使用 alpha 通道可以处理部分反射或透明对象以及处理柔软边界。增加平面数可以使 MPI 表示更宽的深度范围，并允许更大程度的相机移动。此外，从 MPI 渲染生成新视点非常高效，并且可以支持实时应用程序。

MPI 是一种分层表示法，使用一组沿参考视锥放置的平行半透明平面来近似场景的光场。如图 10 所示，MPI 由一组平行平面组成，位于参考相机坐标系的固定深度处，其中每个平面 d 编码 RGBA 图像，包括 RGB 通道和一个 alpha（透明度）通道，以捕获相应深度处的场景外观。MPI 以输入的视图作为监督由卷积神经网络预测得到。从 MPI 渲染新视图时，采用从后向前根据 alpha 权重累加 RGB 值的前向投影的方式进行。

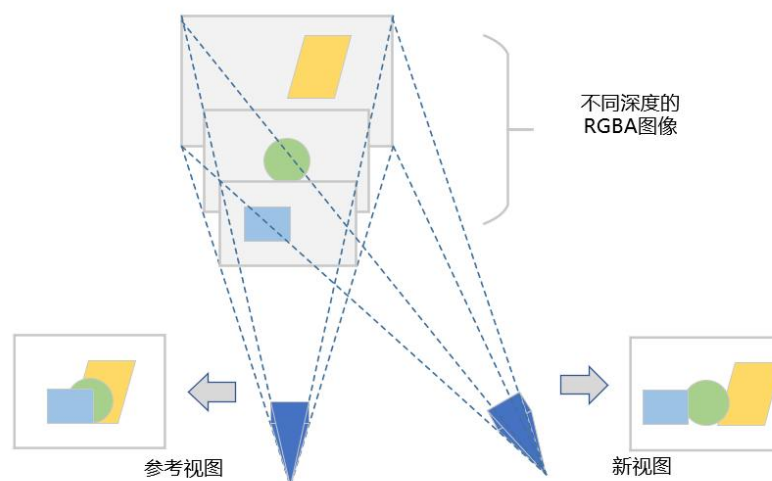


图 10 多平面图像（MPI）示意图

基于 SFM 或基于 RGB-D 相机的新视图生成方法完全依赖于精确的几何估计，然后从附近的视图重投影到新视图并混合纹理。这些方法侧重于输入视图之间的差值而不能预测场景中被遮挡的内容。基于光场渲染的方法通常需要使用数十个相机来对场景进行非常密集的采样。MPI 具有精度高，渲染速度快，输入视角少等特点。相比于其他基于深度学习的新视图生成方法，例如 DeepStereo、NeRF 等，MPI 具有更强的泛化性和更快的训练速度以及渲染速度，以满足实时应用的需求。NeRF 渲染一帧需要耗费 30s，而基于 MPI 的 LLFF[12]、NeX[13]等方法渲染速度可以达到 3-5ms 左右，渲染帧率可以达到 200-300fps。

### 3.5 显示技术

全息视频的展示方式分为穿戴式设备和裸眼 3D 显示设备两种。基于全息通信的特点，人们更倾向于使用裸眼 3D 设备构成解决方案。

#### 3.5.1 穿戴式设备

VR 头显是“虚拟现实头戴式显示器设备”的简称，VR 头显不是通过过滤来自外部屏幕的内容来工作的，而是生成自己的双眼图像，并直接呈现给相应的眼睛。VR 头显通常包含两个微型显示器（左眼一个，右眼一个），经过光学元件的放大和调整，显示在用户眼前的特定位置上。

AR 眼镜，又称作“增强现实头显”。当前增强现实头显变得越来越普遍，

增强现实技术可以把数字世界和现实世界融合在一起。为了确保真实感，增强现实系统不仅需要追踪用户在真实世界的头部运动，同时也要考虑自己所在的现实3D环境。现实世界的光线从不同的方向进入瞳孔之中，这样我们双眼可以看到真实的世界。

### 3.5.2 裸眼3D显示设备

如果不想利用这些穿戴式设备，又想同时以多个视角看到全息影像，则需要用到裸眼全息屏。目前主流的裸眼全息屏技术有基于双目视差和视觉暂留效应的狭缝光栅技术、柱状透镜技术和人眼追踪技术，以及基于空间中三维物体光场重构的体三维技术和光场立体显示技术。

狭缝光栅技术的原理是在屏幕前加了一个狭缝式光栅，应该由左眼看到的图像显示在液晶屏上时，不透明的条纹会遮挡右眼；同理，应该由右眼看到的图像显示在液晶屏上时，不透明的条纹会遮挡左眼，通过将左眼和右眼的可视画面分开，使观者看到3D影像。

柱状透镜技术的原理是通过透镜的折射，将左右眼对应的像素点分别投射在左右眼中，实现图像分离。对比狭缝光栅技术，其最大的优点是透镜不会遮挡光线，所以亮度有了很大改善[14]。

传统的狭缝光栅及柱状透镜全息屏技术只在空间中形成有限的最佳视点，当用户头部移动到最佳视点之外时，双眼会看到串扰的立体图像，影响了立体视觉体验。针对这种问题，通过人眼追踪技术实时定位人眼的空间坐标，再由人眼坐标对图像像素进行重新排布改变最佳视点的区域，很好的扩展了全息视野。不过，由人眼追踪的技术原理可知，目前带有人眼追踪技术的裸眼全息屏只能支持单人观看，即使多人同时看，也只能追踪到一人的视线。而传统的狭缝光栅等技术实现的全息屏则在可视范围内可以多人多视点观看。

体三维显示是一种全新的三维图像显示技术，通过适当方式激励点亮位于显示空间内的物质，利用可见辐射的产生、吸收或散射形成大量的体像素，从而构建出三维图像[15]。体三维显示技术呈现的图像就像是一个真实的三维物体一样，符合人类观察普通三维图像的任何特点，几乎能满足所有的生理和心理深度暗示，可实现多人、多角度、同一时间裸眼观察。

光场三维显示技术如图11所示，这种技术的原理是利用带有方向的光束来构建空间三维物体的光场。空间中任意一个三维物体都可以看作是由无数个发光点组成，任意一个点能够主动或者被动地向空间中各个方向发出携带自身特性的光线[16]。通过设计控光单元的结构、对2D显示设备上加载图像进行有规律的编码等方式，调制有控光单元出射的携带三维场景信息的方向光，使其能够在空间中会聚并构建出向不同方向投射不同空间信息的体像素，用这些体像素来模拟真实物体的发光点，从而实现裸眼观看真三维的显示技术，使人眼获得更真实，自然的3D影像。

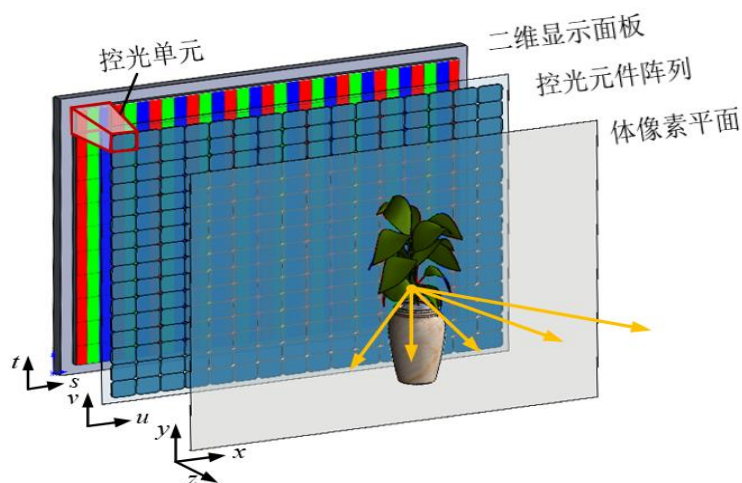


图 11 光场立体显示技术示意图

#### 4. 全息通信技术案例

1849 年，安东尼奥·乌梅奇第一次实现电话通信；1877 年，贝尔将电话技术普及并创立电话公司（AT&T 前身）；1973 年，摩托罗拉公司的工程技术员马丁·库帕发明了世界上第一部民用手机；21 世纪初移动通信在全世界普及；2010 年，iPhone4 面世，人类真正进入智能手机+移动互联网时代，其自带的 FaceTime 功能让“视频通信”成为普通消费者亦触手可及的功能；2014 年，微信视频通话功能上线，让“视频通信”成为了普通中国消费者的日常交流手段。

从固定电话通信到移动通信，花了 100 年；从移动语音通信到视频通信，花了 10 年。现在，我们处在移动互联、5G 和视频通信的时代，也是即将推出“全息通信”的时代。我们有理由相信，在下一个 10 年内，随着 5G 甚至 6G 通信技

术的普及，“全息通信”将会融入我们的生活。

以下是几例走在世界前列的“全息通信”技术案例。

#### 4.1 微软（Microsoft）

微软于 SIGGRAPH 2015 推出了人体三维捕捉原型系统[17]，通过一组 RGB 及 IR 相机采集人体并生成纹理数据，该系统在当时的创新性在于：

- 1) 多模态多视图立体融合 RGB、IR 和轮廓信息；
- 2) 由视觉上重要（敏感）区域的自动检测引导的自适应网格划分；
- 3) 网格跟踪以创建时间上相关的子序列；
- 4) 将纹理网格编码为 MPEG 视频流。

#### 4.2 谷歌（Google）

在今年 5 月举行的谷歌 I/O 大会上，谷歌公布了其多年研究成果——全息视频聊天技术 Project Starline。Project Starline 是一个 3D 视频聊天室，和普通的视频电话会议相比，用户感觉就像坐在真人面前一样。通过 Starline 相互交流的人，不需要佩戴任何眼镜或者头盔，就像坐在对面聊天一样[18]。

该系统三维采集部分仍采用了深度相机+彩色相机的组合（RGBD），系统包括三组 RGBD 相机、跟踪相机等。

该系统虽然只是谷歌早期的实验结果，目前也未有更多的信息表示谷歌会将其投入商用，但其颠覆传统视频通话的浸入感让所有体验者都眼前一亮。

#### 4.3 螳螂慧视（Mantis Vision）

螳螂慧视科技有限公司于 2019 年推出全息视频采集系统原型，并在其以色列研发中心第一次实现了全息直播。如图 12 所示，螳螂慧视的全息影棚采用的多组深度相机+彩色相机（RGBD）的方式，解决了主动式深度相机互相干扰的问题，其标准版全息影棚可配置多达 21 组 RGBD 相机（每组相机由一组深度相机和两组彩色相机组成）。



图 12 螳螂慧视标准全息影棚

## 5. 总结和展望

在全息技术问世以来的几十年间，各机构对于其内涵、技术原理和路线、应用场景等一直在摸索研究，近年来伴随着移动网络的发展，特别是元宇宙概念的引入，全息在通信中的应用受到了格外的重视。全息通信的高沉浸感和真实性等特点使其在人与人的交流中独具优势，而当前 5G 和未来 6G 等网络传输技术的高带宽、低时延、安全可靠和算网结合等特点又极大地拉近了全息通信技术从研究到商用的距离。

对于全息通信技术所涉及的端到端链条上的各个部分，国内外研究机构和相关厂商投入了大量的资源，也提出了各种各样的解决方案，不但相关论文不断见诸于各专业期刊，而且有些产品也已经对外发布。但当前全息通信仍然存在如下问题：

- 1) 采集设备集成度较低，设备占用空间较大；
- 2) 在图像质量和处理时长两方面无法兼顾。好的图像质量往往导致实时性不佳，而可接受的实时性又会以牺牲图像质量为代价；
- 3) 高速移动网络与全息通信技术的融合缺乏架构设计，算网结合等网络特性未纳入当前解决方案的考虑；
- 4) 缺乏能提供整体技术方案的厂商，产业链需要整合。

中国移动希望利用自身在高速移动网络方面的技术优势和广泛经验，大力推

动基于 5G 和 6G 网络的全息通信的研究和产业化，同相关机构和产业界共同解决当前全息通信中存在的问题，促进全息通信产业的发展。

当前全息通信技术正方兴未艾，虽然很多解决方案和产品还有种种不足和限制，但我们相信作为元宇宙概念中的一环，依赖网络技术的成熟和其他相关技术的发展，全息通信在不远的将来会为人类世界带来全新的交流体验。

## 6. 编写单位和作者

中国移动通信有限公司研究院：孙宇平、杨本植、郭漫雪、李征、喻炜、刘堃、赵璐、杨晓伟

螳螂慧视科技有限公司：张小锋、黄诗文、章海波

北京邮电大学：于迅博、邢树军、高鑫、王一平

深圳臻像科技有限公司：黄辉

## 缩略语列表

| 缩略语  | 英文全名                          | 中文解释      |
|------|-------------------------------|-----------|
| 6G   | 6th Generation                | 第六代移动通信技术 |
| XR   | Extended Reality              | 扩展现实      |
| MTP  | Motion to Photons             | 运动到成像     |
| MEC  | Mobile Edge Computing         | 移动边缘计算    |
| SFM  | Structure from Motion         | 运动恢复结构    |
| MVS  | Multi-View Stereo             | 多视角立体视觉   |
| TOF  | Time of Flight                | 飞行时间      |
| NeRF | Neural Radiance Fields        | 神经辐射场     |
| CNN  | Convolutional Neural Networks | 卷积神经网络    |
| FOV  | Field of View                 | 视场角       |
| MVR  | Multi-View Rendering          | 多视角渲染     |
| DIBR | Depth-Image-Based Rendering   | 基于深度图像的渲染 |
| EPI  | Epipolar Plane Image          | 核极线平面图    |
| MPI  | Multi-Plane Image             | 多平面图像     |

## 参考文献

- [1] 満上育久 “Structure from Motion – Osaka University” 映像情報メディア学会誌 Vol.65, No.4, pp.479-482, 2011.
- [2] N.Snaveley, S.M. Seitz, R.Szeliski, “Modeling the World from Internet Photo Collections”, International Journal of Computer Vision, vol.80, no.2, 2008.
- [3] Saxena, A; Chung, SH; Ng, AY 3-d depth reconstruction from a single still image[J]. INTERNATIONAL JOURNAL OF COMPUTER VISION, 2008:76-1.
- [4] Mohamed Sayed, John Gibson, Jamie Watson, Victor Adrian Prisacariu, Michael Firman, Clément Godard: SimpleRecon: 3D Reconstruction without 3D Convolutions. ECCV 2022
- [5] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, Ira Kemelmacher-Shlizerman: HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video. CVPR 2022
- [6]W. Zhou, H. Tang, and Z. Ji, GPU instancing method for crowd animation based on motion capture data[J], J. Comput. Inf. Syst., vol. 9, no. 11, pp. 4459 - 4467, 2013.
- [7]S. G. Parker et al., “OptiX: A general purpose ray tracing engine,” ACM SIGGRAPH 2010 Pap. SIGGRAPH 2010, 2010, doi: 10.1145/1778765.1778803.
- [8] W. Sun, L. Xu, O. C. Au, S. H. Chui, and C. W. Kwok, An overview of free viewpoint Depth-Image-Based Rendering (DIBR) [J],no. December, pp. 14 - 17, 2010, [Online].  
Available: [http://www.apsipa.org/proceedings\\_2010/pdf/APSIPA197.pdf](http://www.apsipa.org/proceedings_2010/pdf/APSIPA197.pdf).
- [9]S. Li, C. Zhu, and M. T. Sun, Hole Filling with Multiple Reference Views in DIBR View Synthesis[J], IEEE Trans. Multimed., vol. 20, no. 8, pp. 1948 - 1959, 2018, doi: 10.1109/TMM.2018.2791810.
- [10] J. P. Lin, W. X. Wang, J. M. Yao, T. L. Guo, E. Chen, and Q. F. Yan, Fast multi-view image rendering method based on reverse search for matching[J], Optik (Stuttg)., vol. 180, no. December 2018, pp. 953 - 961, 2019, doi: 10.1016/j.ijleo.2018.12.003.
- [11]L. T. Muftuler and O. Nalcioglu, Improvement of temporal resolution in fMRI using slice phase encode reordered 3D EPI [J], Magn. Reson. Med., vol. 44, no. 3, pp. 485 - 490, 2015.
- [12]Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics ,2019
- [13]Wizadwongsa, S., Phongthawee, P., Yenphraphai, J., Suwajanakorn, S.: Nex: Real-time view synthesis with neural basis expansion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8534 - 8543 ,2021
- [14]王书路. 基于人眼视觉特性的三维显示研究[D]. 中国科学技术大学,

2016.

[15] 于迅博. 3D 显示技术及其优化方法[M]. 北京:北京邮电大学出版社, 2022.

[16] Ng R. Digital light field photography[M]. Stanford: Stanford university, 2006.

[17] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, Steve Sullivan, Microsoft Corporation: High-Quality Streamable Free-Viewpoint Video, 2015

[18] Jason Lawrence, Danb Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, Kevin Tong: Project starline: a high-fidelity telepresence system, 2021



---

## 全息通信技术白皮书