北京大学

硕士研究生学位论文

题目：沉浸式自由视点视频通信关键技术研究

姓　　名：宁泓博

学　　号：2301212783

院　　系：深圳研究生院

专　　业：计算机应用技术

研究方向：智能媒体处理

导　　师：马思伟 教授

☑ 学术学位　　☐ 专业学位

二〇二六 年 六 月

摘要

随着多媒体通信、计算机视觉与立体显示技术的持续发展，视频通信正由传统二维视频向具备自由观看能力和空间临场感的沉浸式形态演进。不同于传统平面视频，沉浸式自由视点视频支持用户在一定空间范围内自由移动观看，并能够在裸眼观看条件下动态提供双目视差和运动视差。在有限带宽、有限算力以及严格时延约束下，实现高质量自由视点生成与高效沉浸式媒体数据传输，是推动沉浸式自由视点视频通信系统走向实际落地与广泛应用的关键问题。

本文围绕面向接收端实时显示的视点生成方法和面向多视点数据高效压缩与传输的视频编解码方法开展研究。现有方法在实际部署中仍面临显著挑战：一方面，接收端需要在稀疏视点输入和实时交互条件下快速生成高质量目标视图；另一方面，多视点视频数据规模庞大，传统全视点编码传输方式难以兼顾带宽效率、视频质量与系统实时性。针对上述问题，本文提出了快速视点合成方法和多视点视频编解码方法，并构建沉浸式通信系统对所提方法的有效性进行验证。

针对实时通信场景中接收端高效自由视点生成问题，本文提出了一种基于多尺度光流预测的快速视点合成方法。该方法构建了融合图像特征的多尺度光流预测网络，用于建立参考视图与目标视图之间的对应关系，并通过加权融合实现虚拟视图合成；同时，结合输入降采样与模型量化策略对推理过程进行加速。实验结果表明，所提方法在保证合成质量的同时显著降低了推理开销，能够满足沉浸式通信场景下的实时视点生成需求。

针对多视点视频数据冗余大、传输带宽高的问题，本文提出了一种基于三维高斯重建的多视点视频编解码方法。该方法采用稀疏视点编码传输、解码端三维重建和多视点渲染恢复的基本思路，在编码端通过内容自适应视点选择压缩传输数据，在解码端结合双目深度估计与三维高斯参数回归网络恢复未传输视点。实验结果表明，所提方法在保持重建质量的同时显著降低了带宽开销，表现出良好的率失真性能。

为验证上述关键技术在实际场景中的有效性，本文进一步构建了沉浸式自由视点视频通信系统，完成了采集、编码、传输、解码、渲染等模块的集成实现。

关键词：沉浸式通信，自由视点视频，视点合成，多视点视频编码

Research on Key Technologies of Immersive Free-Viewpoint Video Communication

Hongbo Ning (Computer Application Technology)
Directed by: Prof. Siwei Ma

ABSTRACT

With the continuous development of multimedia communication, computer vision, and stereoscopic display technologies, video communication is gradually evolving from conventional 2D video toward immersive forms that provide free-viewpoint interaction and spatial presence. Unlike traditional planar video, immersive free-viewpoint video allows users to move within a certain viewing range while perceiving both binocular disparity and motion parallax without head-mounted displays. Under practical constraints such as limited bandwidth, restricted computational resources, and stringent latency requirements, achieving high-quality free-viewpoint rendering together with efficient immersive media transmission remains a fundamental challenge for real-world deployment of immersive communication systems.

This thesis focuses on two key problems in immersive free-viewpoint video communication: real-time novel view synthesis at the receiver end and efficient compression and transmission of multiview video data. Existing approaches still face notable limitations in practical applications. On the one hand, the receiver is required to generate high-quality target views from sparse inputs under real-time interactive conditions. On the other hand, multiview video contains substantial inter-view redundancy, making conventional full-view transmission schemes difficult to balance bandwidth consumption, visual quality, and system latency simultaneously. To address these issues, this thesis proposes a fast novel view synthesis method and a multiview video codec framework, and further validates the proposed techniques through an immersive communication system implementation.

For efficient free-viewpoint generation in real-time communication scenarios, a fast novel view synthesis method based on multi-scale optical flow prediction is proposed. The method introduces a multi-scale optical flow estimation network with integrated image feature extraction to establish correspondences between reference views and target views, followed by weighted fusion for virtual view synthesis. In addition, input downsampling and model quan-

tization strategies are adopted to accelerate inference. Experimental results show that the proposed method substantially reduces computational overhead while maintaining synthesis quality, meeting the real-time requirements of immersive communication applications.

To reduce the redundancy and transmission bandwidth of multiview video data, this thesis further proposes a multiview video coding method based on 3D Gaussian reconstruction. The proposed framework adopts sparse-view transmission together with decoder-side 3D reconstruction and multiview rendering recovery. At the encoder, content-adaptive viewpoint selection is employed to reduce transmitted data. At the decoder, untransmitted views are recovered through binocular depth estimation and a 3D Gaussian parameter regression network. Experimental results show that the proposed method significantly reduces bandwidth usage while preserving reconstruction quality, achieving favorable rate-distortion performance.

To further validate the effectiveness of the proposed techniques in practical applications, an immersive free-viewpoint video communication system is implemented, integrating modules for data acquisition, encoding, transmission, decoding, and rendering.

KEY WORDS: Immersive communication, Free-viewpoint video, Novel view synthesis, Multiview video coding

目录

第一章　引言.....1
1.1　研究背景与意义.....1
1.2　研究问题与挑战.....1
1.3　研究内容.....3
1.4　论文结构安排.....5
第二章　相关工作.....6
2.1　多视点视频编解码.....6
2.1.1　多视点视频编解码标准.....6
2.1.2　场景表征压缩.....7
2.2　视点合成.....9
2.2.1　深度图像渲染.....9
2.2.2　基于深度学习的视点合成.....10
2.2.3　隐式神经表示.....11
2.2.4　三维高斯重建.....12
2.3　沉浸式视频通信系统.....12
第三章　基于光流预测的快速视点合成方法.....16
3.1　引言.....16
3.2　问题描述与视点合成框架.....17
3.3　基于多尺度光流预测的快速视点合成.....19
3.3.1　参考视图预处理与几何校正.....19
3.3.2　多尺度光流预测网络结构.....21
3.4　网络训练策略与推理加速方法.....25
3.5　实验设置与结果分析.....27
3.5.1　数据集与评价指标.....28
3.5.2　视点合成质量分析.....29
3.5.3　推理效率与实时性分析.....29
3.5.4　消融实验.....30
3.6　小结.....32

第四章　基于三维高斯重建的多视点视频编解码.....34
4.1　引言.....34
4.2　问题描述与总体框架.....35
4.3　基于三维高斯重建的多视点视频解码.....37
4.3.1　多尺度特征提取与双目深度估计.....38
4.3.2　像素级三维高斯参数回归.....40
4.3.3　基于高斯溅射的多视点重建与渲染.....42
4.3.4　损失函数与训练策略.....43
4.4　内容自适应视点选择与稀疏视点编码.....44
4.5　实验设置与结果分析.....49
4.5.1　实验设置.....49
4.5.2　总体带宽与率失真性能分析.....50
4.5.3　重建质量主客观对比分析.....52
4.5.4　消融实验.....53
4.5.5　关键参数影响分析.....54
4.6　小结.....55
第五章　沉浸式自由视点视频通信系统.....58
5.1　系统设计.....58
5.1.1　系统总体架构.....58
5.1.2　运行模式设计.....59
5.2　系统实现.....61
5.2.1　系统硬件平台配置.....61
5.2.2　发送端实现.....62
5.2.3　接收端实现.....64
5.3　系统测试与分析.....65
5.3.1　端到端时延与实时性.....65
5.3.2　带宽占用.....66
5.4　小结.....67
第六章　总结与展望.....68
6.1　本文总结.....68
6.2　研究展望.....68
参考文献.....70

攻读硕士期间发表的论文及其他成果 75

致谢 76

表格索引

表 2.1　典型沉浸式通信系统对比 ……14

表 3.1　实时采集场景下的视点合成质量主客观评价结果 ……29
表 3.2　视点合成网络不同部署方案的配置与效率对比……30
表 3.3　模型量化对视点合成质量的影响 ……31
表 3.4　低分辨率光流估计对视点合成质量的影响 ……31

表 5.1　沉浸式视频通信系统的两类运行模式设计 ……60
表 5.2　单用户眼动驱动模式系统端到端延迟……65
表 5.3　多用户多视点光场显示模式系统端到端延迟 ……66

插图索引

图 1.1　多媒体技术发展趋势……2
图 1.2　本文的内容组织结构示意图……4

图 2.1　沉浸式通信系统发展……13

图 3.1　视点合成流程示意图……21
图 3.2　多尺度级联细化的光流预测网络结构……22
图 3.3　视点合成视觉质量定性对比……32

图 4.1　基于三维高斯重建的多视点视频编解码总体框架……36
图 4.2　特征提取、深度估计与高斯参数回归网络结构图……37
图 4.3　不同目标距离条件下候选视点对与重建质量之间的关系……45
图 4.4　多视点视频编解码总体带宽与视觉质量评估……51
图 4.5　基于三维高斯重建的多视点编解码率失真曲线……52
图 4.6　基于三维高斯重建的多视点编解码及其变体的消融实验……54
图 4.7　不同视点数量下的带宽和视觉质量评估……55
图 4.8　多视点解码视觉质量定性对比 1……56
图 4.9　多视点解码视觉质量定性对比 2……57

图 5.1　沉浸式自由视点视频通信系统总体架构……59
图 5.2　沉浸式自由视点通信系统双端实现图……61

第一章　引言

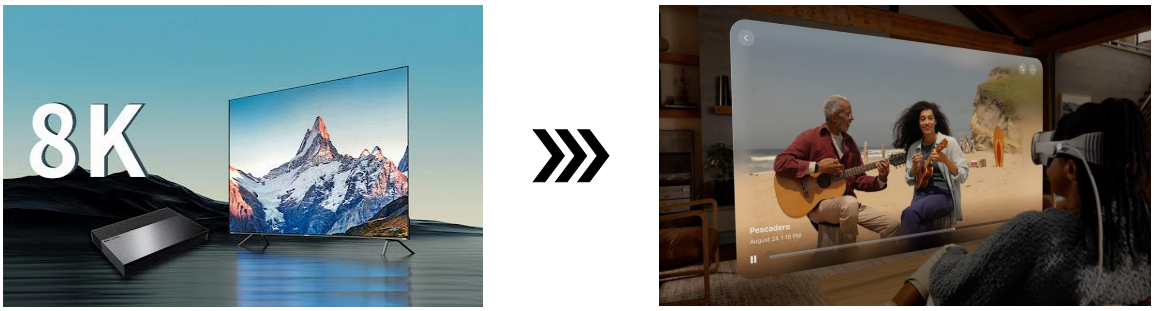
1.1　研究背景与意义

随着多媒体通信、计算机视觉以及立体显示技术的持续发展，视频通信正由传统的二维平面视频形态，逐步向具备自由视点观看能力和空间临场感的沉浸式视频形态演进。传统视频通信通常依赖固定摄像机视角，用户只能被动接收单一视点的图像信息，难以感知真实场景中的空间结构及获得观察自由度。相比之下，沉浸式视频通过对远端场景的三维结构与空间关系进行建模与重建，并结合多视点立体显示、光场显示等三维显示技术进行呈现，形成立体视觉。用户能够在远程通信过程过获得双目视差和一定程度的运动视差，进而提升视觉真实感与空间临场感。自由视点视频（Free-Viewpoint Video, FVV）通信是在上述技术背景下发展起来的重要研究方向。其核心在于根据用户的观察位置或交互意图，在接收端动态生成与之对应的目标视图。通过该机制，用户可以在一定空间范围内连续改变观察视角，实现从“固定观看”向“可交互观看”的转变，并获得更加自然的视线交流体验以及更强的空间沉浸感。

从理论研究角度看，沉浸式自由视点视频通信涉及多视点数据传输、视点合成与三维重建等多个方向。其核心问题在于如何在有限采集条件下实现高质量的空间信息重建与表达。围绕这一问题开展研究，有助于推动视点合成方法在保证重建质量的同时向低复杂度与实时性能方向发展，并为多视点视频的高效压缩与传输提供新的思路。从应用层面来看，沉浸式自由视点视频通信在远程会议、远程医疗、在线教育以及虚拟社交等场景中具有较高的应用价值。例如，在远程会议场景中，立体显示与自由视点显示能够提升人与人之间的交互真实感；在医学教学或远程诊疗场景中，多视角观察方式有助于理解复杂三维结构；在虚拟现实与数字人交互等应用中，沉浸式视频通信也为高真实感虚拟环境的构建提供了基础支撑。因此，针对沉浸式自由视点视频通信相关关键技术开展研究，对于推动视频通信由“看得清”向“身临其境”演进，并促进沉浸式视频通信技术的实际落地，具有重要意义。

1.2　研究问题与挑战

近年来，面向沉浸式远程通信的原型系统不断涌现，推动了沉浸式自由视点视频通信从概念研究走向系统验证。例如，Project Starline[1] 结合多视点采集、低时延传输与裸眼立体显示，实现了具有空间临场感和目光接触体验的远程交互；VirtualCube[2] 面向多人远程会议场景，利用多路 RGBD 采集与大尺寸立体显示设备构建了近似等身



高清多媒体：“看得清”　　沉浸式视频：“身临其境”

图 1.1　多媒体技术发展趋势

比例的沉浸式通信环境；Tele-Aloha[3] 则进一步展示了基于稀疏 RGB 采集、低时延视频传输和前馈式三维高斯重建的实时系统方案。这些工作表明，沉浸式自由视点视频通信需要联合考虑采集、编码、传输、重建、渲染与显示等多个环节。其中任何一个模块的性能瓶颈，都可能影响整体系统的实时性与用户体验。

从技术链路上看，沉浸式自由视点视频通信需要同时解决两个关键问题。其一，接收端需要根据用户位置、显示方式或交互需求，快速生成对应目标视点的图像内容。其二，发送端和传输链路需要控制多视点数据规模，避免由于视点数量增加而带来过高的码率、编码复杂度和端到端时延。因此，视点合成与多视点视频编码分别构成了沉浸式自由视点视频通信中的两个关键支撑环节：前者决定了系统是否能在有限算力条件下灵活地响应用户视角变化并生成高质量目标视图，后者决定了系统是否能在有限带宽条件下以可接受的码率和复杂度支撑自由视点交互。

视点合成是实现自由视点交互的基础能力，其本质是在给定一个或多个参考视图的条件下，重建目标视点处的图像内容。早期视点合成方法大多依赖显式几何建模，典型代表是基于深度图像渲染（Depth-Image-Based Rendering, DIBR）的视图生成框架。这类方法通常先通过立体匹配或多视图几何估计场景深度，再利用投影变换将参考视图映射到目标视图。然而，在实际动态场景中，深度不连续区域、遮挡边界、非刚性运动和纹理稀疏区域都会增加深度估计与投影渲染的难度，容易产生空洞、错位、边界撕裂等伪影。随着深度学习方法的发展，基于神经网络的深度估计和视图生成方法在重建质量方面取得了明显进展，但其模型规模和计算开销通常较大，难以直接满足沉浸式通信中对低时延、高帧率和高分辨率输出的要求。

隐式神经表示的发展进一步推动了视点合成研究范式的演化。NeRF[4] 通过学习连续空间中的体密度与有向辐射，实现了高保真场景重建与新视点合成，在真实感和几何一致性方面表现出显著优势。然而，NeRF 类方法通常依赖沿像素射线的密集采样和体渲染计算，训练与推理成本较高，其端到端计算开销难以完全适应实时交互式通信

场景。

近年来，以三维高斯溅射（3D Gaussian Splatting, 3DGS）[5] 为代表的显式场景表示方法，在渲染效率和场景表达能力之间取得了较好的平衡。该类方法利用具有位置、尺度、旋转、颜色和不透明度等属性的高斯基元描述三维场景，并通过高效光栅化实现实时新视点渲染。相比 NeRF 类隐式表示，3DGS 在推理速度上具有明显优势。然而，从通信系统角度来看，直接采用高精度三维高斯模型仍存在模型参数规模大、构建成本高以及对输入视角覆盖程度依赖较强等问题。特别是在实时视频传输场景中，若缺乏与编码传输过程相适配的轻量化表示与重建机制，其实时渲染的优势难以充分转化为系统层面的通信效率。

与单目视频相比，多视点视频同时包含时域冗余和跨视点冗余，数据规模随视点数量增加而显著增长。MVC[6]、MV-HEVC[7] 等多视点视频编码技术主要通过跨视点预测降低冗余，在立体视频和多视点视频压缩方面取得了重要进展。MPEG-I Immersive Video（MIV）[8] 则进一步面向沉浸式视频应用，通过有限视图、深度信息和视点合成机制支持更大范围的自由观看。然而，在沉浸式自由视点视频通信场景中，随着视点数量、视频分辨率和交互自由度的提升，传统多视点编码传输方法在带宽占用、编码时延和系统扩展性方面逐渐显现不足。尤其对于裸眼 3D 显示和自由视角切换场景，仅依赖传统多视点视频编码难以同时满足高压缩效率、低时延和高交互性的综合要求。

综上所述，当前沉浸式自由视点视频通信仍面临两方面突出挑战。其一，在视点合成层面，需要在稀疏输入、动态场景和严格实时性约束下，高效生成几何一致、边界自然且细节完整的目标视图，避免显式几何方法中深度误差传播带来的伪影，同时避免隐式渲染方法带来的过高计算代价。其二，在多视点视频编码与传输层面，需要突破传统多视点传输方式的数据冗余限制，探索能够兼顾压缩效率与解码重建质量的多视点视频传输方式，从而为自由视点交互提供可接受的码率和复杂度支撑。

1.3 研究内容

面向沉浸式自由视点视频通信在实时视点合成与多视点视频高效传输方面面临的关键问题，本文围绕快速视点合成与基于三维高斯的多视点视频编解码两条技术主线开展研究，目标是在保证重建质量与空间一致性的前提下，降低系统的计算复杂度、传输开销以及端到端时延。具体研究内容如下。

（1）针对沉浸式通信场景中接收端需要根据用户观察位置实时生成目标视图的问题，本文研究了一种基于多尺度光流预测的快速视点合成方法。该方法从参考视图之间的像素对应关系出发，将目标视图的生成过程表示为由参考视图到目标视图的映射问题，在不依赖显式深度的条件下实现目标视图的直接生成。针对沉浸式场景中稀疏

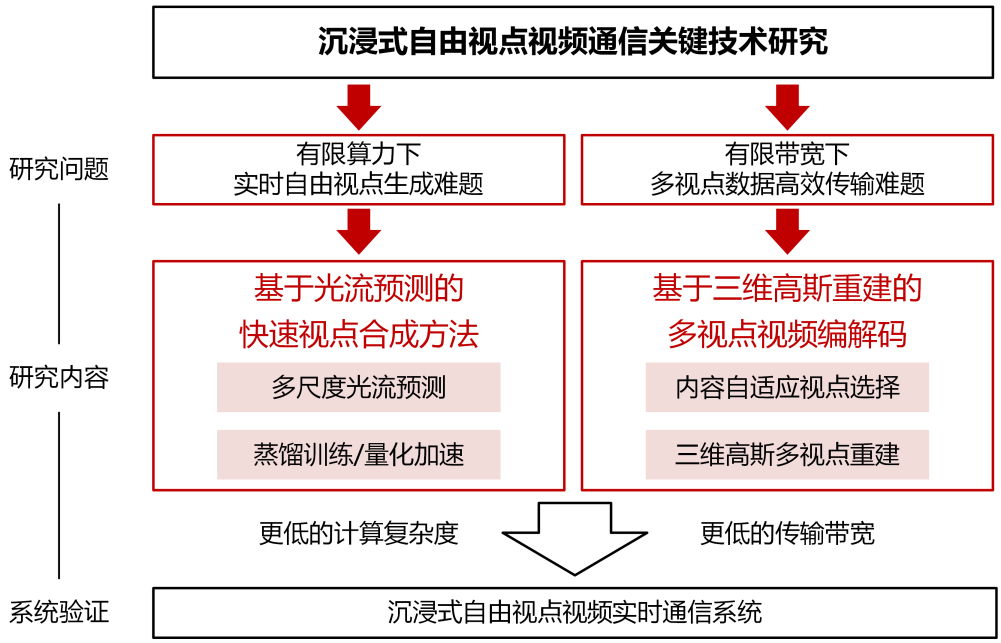


图 1.2　本文的内容组织结构示意图

视点输入、视差变化较大以及动态内容带来的挑战，本文引入多尺度金字塔对不同尺度的像素运动进行统一刻画，以提升对复杂视图变化的适应能力。在此基础上，进一步面向实际系统部署对算法的推理过程进行优化，以降低计算开销并提升运行效率，使其能够在有限算力条件下实现稳定的实时推理性能，支持自由视点交互场景下连续视点变化的实时生成需求。

（2）针对多视点视频数据规模大、传输冗余高以及传统全视点编码方式难以兼顾压缩效率与交互灵活性的问题，本文研究了一种基于三维高斯重建的多视点视频编解码方法。面对多视点数据存在空间冗余的特性，该方法采用视点选择与解码端重建相结合的思路，在编码端对原始多视点视频进行内容自适应的视点筛选，仅传输具有代表性的少量参考视图，从而降低传输数据规模。针对稀疏视点条件下多视点恢复困难的问题，本文在解码端引入基于三维高斯的多视点重建方法。该方法不依赖传统三维高斯方法中针对单一场景的逐场景优化过程，而是根据解码端输入的稀疏参考视图直接重建场景的三维高斯表示，并进一步生成未传输视点。通过这种方式，使得在减少多视点编码冗余的同时，保持对动态多视点视频内容的重建能力，满足沉浸式自由视点交互对码率与计算复杂度的约束需求。

围绕上述方法，本文进一步构建了沉浸式自由视点视频通信系统，并对所提方法在实际通信场景中的性能进行了实验验证。从重建质量、运行效率以及带宽开销等方面对系统进行综合评估，分析所提方法在真实采集与传输环境下的适用性与性能表现。

1.4　论文结构安排

全文围绕沉浸式自由视点视频通信中的关键技术问题展开，内容安排如下。

第一章为引言，主要介绍沉浸式自由视点视频通信的研究背景与意义，分析当前方法在实时性与传输效率方面面临的主要问题与挑战，并给出本文的研究内容与整体技术路线。

第二章为相关工作，重点从视点合成、多视点视频编解码以及沉浸式通信系统三个方面总结现有研究进展，并分析其在实时性、压缩效率与系统部署方面的不足，为后续方法设计提供依据。

第三章研究基于光流预测的快速视点合成方法。该方法围绕稀疏视点输入条件下的视图生成过程进行建模，并结合推理过程优化提升方法实时运行能力。通过实验验证所提方法在重建质量和运行效率方面的有效性。

第四章研究基于三维高斯重建的多视点视频编解码。该方法通过视点选择与解码端三维重建相结合，实现由少量参考视图到多视点内容的恢复，从而降低数据传输规模并支撑自由视点交互。通过实验验证所提方法在重建质量和带宽效率方面的有效性。

第五章在上述方法基础上构建沉浸式自由视点视频通信系统，对视点合成与多视点编解码方法进行集成实现，并从运行效率与带宽占用等方面对系统性能进行实验分析与验证。

第六章对全文工作进行总结，并对沉浸式自由视点视频通信未来的研究方向进行展望。

第二章　相关工作

随着沉浸式媒体与通信技术的发展，沉浸式自由视点视频通信逐渐成为学术界与工业界关注的核心方向。为了在远程场景中真实再现人物与环境，并实现自然的视线交流、空间临场感以及实时交互体验，沉浸式通信系统需要在采集、编码、传输、重建与渲染等多个环节形成紧密协同。其中，多视点视频编解码与视点合成是支撑沉浸式通信系统的两项关键底层技术：多视点视频编解码面向多路视频数据的高效压缩与传输，是保证系统在带宽受限和时延敏感条件下稳定运行的前提；视点合成则决定了系统能否以有限视角生成高质量、几何一致的新视图，是实现自由移动观看与三维呈现的基础。而沉浸式视频通信系统本质上是对这些底层能力的系统级集成，通过低时延网络传输与实时三维渲染构建完整端到端管线，为用户提供近似面对面的三维临场体验。基于此，本章将分别从多视点视频编解码、视点合成与沉浸式通信系统三个层面，综述国内外相关工作的技术演进、典型方法与现有挑战，为后续系统设计与技术创新奠定背景基础。

2.1　多视点视频编解码

2.1.1　多视点视频编解码标准

多视点视频编码（Multiview Video Coding, MVC）旨在对来自同一场景的多路同步视频进行联合压缩，以支持立体显示、自由视点交互以及沉浸式媒体呈现。相较于传统单目视频，多视点数据同时存在时间、空间以及视角维度上的高度冗余，因此有效利用跨视角的几何相关性一直是多视点编码研究的核心。早期的研究主要在既有编码框架基础上进行扩展，其中最具代表性的工作是 H.264/AVC[6] 的 MVC 扩展，它在传统编解码流程中加入了跨视图预测结构，使得多路视频序列可以在保持与单视图编码兼容的同时进行联合压缩，大幅提升了对视角冗余的利用效率，对多摄像机系统和立体视频的普及起到了重要推动作用。

随着视频分辨率和场景复杂度不断提升，MVC 在压缩效率和参考结构上的局限逐渐显现，H.265/HEVC[9] 标准下的 MV-HEVC[7] 由此提出。MV-HEVC[7] 在 HEVC 强大的编码工具基础上设计了更灵活的跨视点参考结构，在视角预测性能、复杂场景适应性以及硬件实现友好性方面均显著优于 MVC，同时继承了 HEVC 成熟的工程生态，使其成为多视点视频压缩的重要里程碑。在此基础上，3D-HEVC[7] 进一步针对三维显示和自由视点应用的需求引入了纹理—深度联合编码方式，通过深度建模、几何校正与

深度特定工具（如深度边缘保真、视差补偿优化等）实现高质量的合成视图生成，使得基于深度图像渲染的自由视点视频在标准层面得到系统支持。

近年来，在虚拟现实、增强现实和全景交互式视频等沉浸式媒体的推动下，传统的视点扩展编码已难以满足六自由度（6DoF）场景中对大范围视场与自由移动的需求。MPEG 因此制定了 MPEG Immersive Video（MIV）标准[8]，其核心理念是在不改变底层视频编码工具的前提下，通过挑选一组关键视图进行传输，再在解码端结合深度信息与视点合成技术重建完整场景。MIV[8] 的设计兼顾压缩效率、渲染质量与带宽约束，使得利用有限的观测视图即可实现自由移动和沉浸式观看，成为新一代沉浸式视频标准体系中的重要组成部分。

与此同时，我国的 AVS[10] 标准体系亦在多视点编码方面持续推进。从 AVS2[11] 到 AVS3[12]，标准中逐渐加入对立体视频、多视点视频和基于深度的自由视点渲染的支持，形成了较完整的国产三维与多视点视频编码方案。这些扩展不仅面向国内产业生态提出了工程可落地的解决方案，也为未来在沉浸式媒体领域的自主可控技术体系奠定了基础。整体来看，从 MVC、MV-HEVC 到 3D-HEVC 再到 MIV 和 AVS 系列，多视点视频编码逐步从帧间预测扩展至视角预测、从纹理编码扩展至纹理—深度联合编码，在标准体系上逐渐向场景层级的信息组织与复原能力演进。

2.1.2　场景表征压缩

随着三维传感器和三维建模技术的发展，点云（point cloud）、体素（voxel grid）、网格（mesh）等显式三维表示逐渐成为沉浸式媒体研究的重点。为应对点云数据规模巨大、结构稀疏等问题，ISO/IEC MPEG 推出了基于几何与属性的点云压缩（Point Cloud Compression, PCC）标准[13]，包括基于体素投影的 V-PCC（Video-based PCC）和基于几何树结构的 G-PCC（Geometry-based PCC）。V-PCC 将点云投影视图编码为二维视频帧，利用现有视频编码器实现压缩；G-PCC 则直接在三维空间中进行熵编码。尽管 PCC 标准在几何压缩方面取得良好效果，但点云表示对细节与外观的表达能力有限，难以满足高保真沉浸式通信需求。此外，PCC 依旧延续压缩、传输、重建的范式，与实时交互式通信对低延时、动态一致性和高质量纹理的要求存在差距。

近年来，随着隐式神经场表示与可微分渲染技术的发展，多视点视频编码逐渐从传统的基于图像序列的压缩模式转向对场景三维结构与外观的直接编码。与 MVC、3D-HEVC 或 MIV 等标准不同，这类方法不再把每个视频帧视作独立编码单位，而是将整个场景表示为紧致的神经函数或显式点基元，从而在根本上消除跨视点、跨时间的冗余，使多视点视频传输从视频流压缩迈向场景表征压缩。其中，两条发展线最具代表性：基于 NeRF[4] 的隐式辐射场编码，以及基于三维高斯泼溅[5] 的可学习点云编码。

基于 NeRF[4] 的多视点编码方法将场景表示为连续辐射场，并通过压缩网络参数、体素特征或密度信息来降低多视点数据的传输开销。早期工作中，Hedman 等[14] 采用 baking 策略，将训练完成的 NeRF 从连续 MLP 转换为稀疏特征网格，从而提升渲染效率并减少存储成本。随着相关研究的发展，NeRF 逐渐被视作一种可压缩的三维场景表示。Li 等[15] 结合体素稀疏化、可学习量化和熵编码，将完整 NeRF 模型压缩至约 1 MB，同时仍能保持较好的新视点渲染质量。NeRFCodec[16] 提出了面向神经特征的压缩框架，通过表示分解与端到端优化提高特征网格的编码效率。除静态模型压缩外，也有研究关注 NeRF 在流式传输场景中的应用。Li 等[17] 将动态场景表示为可逐步更新的辐射场，使接收端能够根据网络带宽变化实时重建场景。Wang 等[18] 对相邻视图之间的辐射场残差进行建模，以减少动态场景中的冗余信息，使其编码结构更接近传统视频压缩框架。Zhang 等[19] 则在隐式场表示中引入码率控制机制，使动态 NeRF 的特征存储与体渲染过程能够在带宽约束下进行自适应优化。总体来看，这类方法在视图合成质量和场景表达能力方面具有明显优势，但较大的模型规模和较高的推理开销仍限制了其在低时延沉浸式通信中的实际应用。

相比 NeRF，基于三维高斯的显式表示更适合实时多视点渲染与传输。3DGS[5] 将场景表示为一组带有位置、协方差、颜色和不透明度等属性的三维高斯集合，通过光栅化方式完成新视点渲染，相比传统 NeRF 显著降低了渲染阶段的计算复杂度，因此在自由视点视频与沉浸式通信场景中受到广泛关注。不过，原始 3DGS 往往包含大量高斯基元，模型规模通常达到数百 MB 甚至更高，在传输与终端部署过程中会带来较大压力。围绕这一问题，近年来出现了大量面向 3DGS 的压缩研究，其思路大致可以分为两类：一类通过剪枝、结构重构等方式减少高斯数量；另一类则在保持整体结构基本不变的情况下，对高斯属性进行量化与熵编码，以降低存储与传输开销。在具体实现上，早期方法主要采用高斯剪枝和向量量化策略。例如，Scaffold-GS[20] 利用锚点结构减少需要显式存储的高斯参数，并结合量化与熵编码实现较高压缩率。HAC[21] 通过上下文建模实现 3DGS 压缩，利用结构化 hash-grid 特征为无序高斯提供上下文信息，并结合条件概率建模与算术编码完成高效熵编码，从而进一步降低码率开销。

除了模型级压缩之外，近年出现的另一方向是基于 3DGS 的自由视点视频流式编码，其目标是在动态场景中实现可实时更新、可实时渲染、可实时传输的场景表示。3DGStream[22] 将动态序列的每一帧表示为在全局 3DGS 参考模型上的轻量变换，通过 Neural Transformation Cache（NTC）对高斯的平移与旋转进行紧致建模；Dai 等[23] 从运动结构进行建模，通过静态层与动态层构建可流式的高斯表示，每个动点只需要存储其属性的时间偏移量，并通过 H.265 对这些偏移量进行编码；GIFStream[24] 采用基于规范空间的四维高斯场景表示，通过随时间变化的特征流建模复杂运动，并整合时

间和空间压缩网络实现端到端压缩。这些工作共同展示了 3DGS 在多视点视频编码中的潜在优势：其显式结构便于剪枝、量化和上下文建模；其参数可自然映射到二维结构，易于与现有视频标准结合；其渲染效率高，解码侧几乎没有 MLP 推理开销，非常适合构建低延迟的沉浸式通信系统。然而，目前大多数方法仍停留在文件级模型压缩或离线编码阶段，如何将 3DGS 的压缩、更新、传输和显示整合成端到端低时延系统，仍是未来研究的重要方向。

总体来看，多视点视频编码正经历从传统视频预测压缩向神经场景表示压缩的范式演进。随着 NeRF、3DGS 等场景表征方法在实时性和质量方面不断提升，未来的多视点编码可能不再依赖逐帧视频序列，而是以轻量、可压缩、可渲染的三维场景模型为核心，实现高效、灵活、低延迟的自由视点传输，为沉浸式通信系统奠定更加坚实的技术基础。

2.2　视点合成

视点合成（Novel View Synthesis, NVS）旨在在有限数量的输入视角下生成未观测到的虚拟视点图像，是自由视点视频、虚拟现实、增强现实和沉浸式通信系统的核心底层能力。其目标是重建场景的三维几何结构与光照外观，使用户可以在虚拟空间中自由移动并获得稳定、真实的视觉体验。视点合成技术从早期的基于显式深度渲染方法不断演变至基于学习的深度重建方法，并在近几年随着隐式神经表示及可微分渲染的发展迎来新一轮突破。总体而言，其技术发展经历了基于深度图的渲染、基于体积与距离场的空间融合、基于深度学习的多视图几何估计、基于隐式神经场的体渲染方法以及以三维高斯泼溅为代表的高效可微点云表示等阶段，各阶段在精度、鲁棒性、实时性与可扩展性方面均呈现出显著差异。

2.2.1　深度图像渲染

早期的视点合成主要依赖显式几何建模，通过对输入相机的深度图或立体视觉估计结果进行投影变换生成新视图。其中最具代表性的技术路线是深度图像渲染（Depth Image-Based Rendering, DIBR）[25]。DIBR 将输入图像与深度图结合，通过像素反投影与再投影直接构造虚拟视点图像。该方法在理论上简单高效，不需要完整的三维网格模型，因此被广泛应用于早期的三维电视、视频会议与自由视点娱乐系统中。然而，由于深度图往往存在噪声与边缘误差，DIBR 在遮挡边界处容易产生裂缝、拉伸与空洞等伪影。Zitnick 等[26] 通过分层表示减少空洞和遮挡；Oh 等[27] 提出基于深度图的修复方法来解决自由视点电视中的空洞问题；Do 等[28] 通过深度图中值滤波和像素反向映射的方法来避免深度边界处的颜色混合和修复空洞；Scharstein[29] 通过增加虚拟相机与参

考相机之间的距离比例来提高采样率，减少因采样不足引起的轮廓伪影；Mueller 等[30]基于边缘检测进行深度图的可靠和不可靠区域分离，分别处理后再融合以减少深度不准确带来的轮廓伪影和失真。这些研究通过自适应视差优化、边缘增强、基于图像修补的遮挡填充等方式有所改善，但其根本受限于输入深度的准确性，无法应对复杂动态场景或稀疏视角条件。

除基于深度图的几何投影路线外，视点合成的另一条重要发展脉络是图像式渲染（Image-Based Rendering, IBR）。早期工作如 Plenoptic Modeling[31] 和基于图像先验的渲染方法[32] 表明，在已知或近似已知相机位姿的条件下，可以不显式恢复完整三维网格，而是通过对输入图像之间的采样、重投影和颜色约束直接生成新视图。这类方法在一定程度上缓解了显式几何重建的不稳定性，并为后续基于学习的视点合成提供了重要思想基础，即将新视点生成看作由多视图观测到目标图像的映射问题，而非严格依赖中间几何模型的显式恢复问题。

2.2.2 基于深度学习的视点合成

随着深度学习的发展，研究者开始尝试利用端到端网络直接学习由源视图到目标视图的映射关系。DeepStereo[33] 是较早将深度学习系统性应用于真实场景新视点生成的代表性工作，其通过体积化代价表示与学习式颜色融合实现高质量图像合成。随后，Stereo Magnification[34] 提出了多平面图像（Multiplane Images, MPI）表示，通过在若干深度平面上分层存储颜色与透明度信息，有效提升了窄基线视角外推能力。再之后，SynSin[35] 将潜在三维点云和可微渲染结合起来，使单幅图像条件下的视点合成也能够获得较好的几何一致性。这些方法共同推动视点合成从基于显式几何的投影渲染逐步过渡到基于深度学习的前馈式生成，为后续基于神经场和高斯表示的实时新视点合成奠定了基础。

为提升几何一致性与视点合成质量，多视图体积融合（volumetric fusion）逐渐成为主流方向，其中最典型的技术是截断符号距离函数（Truncated Signed Distance Function, TSDF）方法[36]。TSDF 将多视角深度图融合到同一体素网格中，以加权方式消除深度噪声，从而形成一致的体积几何表示。基于 TSDF 的光线投射能够生成平滑的表面并改善几何稳定性，尤其在多视图密集采集条件下表现显著优于单视图 DIBR。随后 KinectFusion[37] 将 TSDF 融合与 GPU 并行计算结合，实现了实时的三维表面重建，为后续实时三维会议和自由视点系统提供了一条可行路径。TSDF 方法在大规模场景中仍然存在内存占用高、缺乏表达力（例如对透明、细薄或非刚体结构）的不足，难以满足复杂沉浸式通信环境的需求。

随着深度学习的发展，研究者开始探索利用神经网络解决传统立体匹配与多视图

几何恢复中的关键难题。典型代表如 MVSNet[38-40]、Cascade-Stereo[41] 等，它们从多视图 RGB 图像构建代价体，并通过三维卷积网络推断高精度深度图。这类方法能够在稀疏视角条件下恢复结构一致的几何信息，为高质量视点合成提供可靠基础。同时，基于图像融合的深度学习方法如 Deep Blending[42] 在遮挡处理方面展现了优越性，通过学习融合权重与纹理补全策略显著降低几何误差带来的伪影。尽管这些方法在静态或半动态场景中取得了优异表现，但推理复杂度较高，通常无法在 30–60 FPS 的实时通信场景中达到端到端延迟要求。

值得注意的是，在面向双目或窄基线输入的视点生成任务中，基于光流估计与图像插值的技术路线近年来受到越来越多关注。尽管视频帧插值最初主要服务于时间域中间帧生成，但其核心思想——通过估计输入图像之间的稠密对应关系，并结合可微扭曲与遮挡推理恢复目标图像，与窄基线条件下的新视点生成具有较强共通性。Super SloMo[43] 通过双向光流与可见性掩膜联合建模实现多中间帧插值；Niklaus 等提出的 Softmax Splatting[44] 通过前向扭曲与冲突消解机制改善了运动边界区域的合成效果；RIFE[45] 则提出直接估计中间流场的 IFNet 结构，在运行速度与重建质量之间取得了较好平衡。虽然这些方法最初针对的是视频时间插值，但其在对应关系估计、遮挡建模、多尺度细化与高效推理等方面提供的思路，已经被逐步迁移到双目视图插值、窄基线视点生成以及轻量级自由视点渲染等任务中。这类方法尤其适合显示端只需在相邻少量视点间快速生成双目图像或局部虚拟视图的应用场景，例如裸眼三维显示中的左右眼图像生成、窄范围头动补偿以及稀疏多摄像机条件下的实时视图补全。然而，该类方法也存在对基线范围敏感、对大遮挡区域恢复能力有限、缺乏统一三维几何约束等不足，因此常需要结合几何校正、特征提取和多尺度融合机制提升合成质量。总体而言，基于光流/插值的视点合成方法构成了实时自由视点视频系统中一条兼顾效果与复杂度的重要技术路线。

2.2.3　隐式神经表示

视点合成领域真正的范式转变发生在隐式神经表示（Implicit Neural Representation, INR）兴起之后，其代表性成果是 NeRF（Neural Radiance Fields）[4]。NeRF 使用多层感知机隐式建模三维场景中每个空间点的体密度与颜色，并利用可微体渲染实现高质量自由视点生成。得益于其对光照、遮挡、细节纹理的建模能力，NeRF 很快成为计算机图形学与计算机视觉领域的研究热点。其主要优势在于不依赖显式几何结构、能表达复杂材质，并在有限视图下也能恢复全局一致的场景。然而，NeRF 的体渲染需要对每条射线进行数百次采样并反复调用 MLP，导致其训练与推理速度难以满足实时需求。例如，标准 NeRF 的训练通常需要数小时至数天，而每帧渲染耗时可达数百毫秒甚

至数秒。加速方法和衍生变体如 PlenOctrees[46]、FastNeRF[47]、IBRNet[48]、ENeRF[49]、Instant-NGP[50] 虽然在训练和渲染速度上取得显著进展，但仍未解决实时通信系统中端到端延迟严格受限的问题。

2.2.4 三维高斯重建

近年，三维高斯泼溅（3D Gaussian Splatting, 3DGS）[5] 的提出为视点合成带来了新的突破。3DGS 不再依赖体密度采样，而是将场景表示为一组带有空间位置、各向异性协方差、透明度与颜色属性的三维高斯分布。渲染时，每个高斯作为一个可微的点基元进行泼溅，并通过光栅化排序实现高效渲染。得益于图形管线的高度并行性，3DGS 能够在消费级 GPU 上以 100 FPS 以上的速度进行高质量渲染，真正达到了实时应用的标准。大量研究基于 3DGS 进行了扩展，例如 SplaTAM[51] 将 SLAM 与高斯表示结合用于实时场景重建，Dai 等[52] 提出的基于高斯面片的表示方法通过去除由 alpha 混合生成的深度图中的错误点来提升重建精度。然而，原始的 3DGS 需要在几分钟内进行场景特定的训练，pixelSplat[53]、Szymanowicz 等[54]、MVSGaussian[55]、GPS-Gaussian[56] 等研究尝试通过在二维图像平面上构建高斯映射来解决这种泛化性问题；VGGT[57]、AnySplat[58] 则采用 ViT[59] 的编码-解码架构从多视角图像推理场景几何与外观等三维属性，并进一步生成用于实时新视点渲染的 3DGS 表示。3DGS 在视觉质量、运行效率和可优化性之间取得了良好平衡，使其成为当前最有潜力用于沉浸式通信系统的新视点生成技术。

总体来看，视点合成技术正从依赖显式深度图与体积几何的传统方法演化到以神经网络为核心的隐式表示与可微分点云渲染方法。尽管当前的 NeRF 与 3DGS 在视觉质量上均取得突破，但要满足沉浸式通信系统对低时延、高分辨率、低计算资源消耗和动态一致性和可传输性的综合要求仍然存在诸多挑战。例如，在快速表情变化、头发等高频几何区域，现有方法仍可能出现抖动与不稳定；在高分辨率光场显示场景下，3DGS 仍需进一步提高像素级细节与抗噪能力；在传输链路上，高斯点云优化参数量大，如何实现实时压缩与编码仍是开放问题。

2.3 沉浸式视频通信系统

随着虚拟现实、增强现实、自由视点视频和体积视频等沉浸式媒体技术的快速发展，沉浸式视频通信逐渐被视为下一代信息通信系统的重要形态。相比传统二维视频会议，沉浸式通信更加关注三维几何的真实重现、自然的视线交互、空间化音频以及端到端低时延等综合指标，其目标是在不依赖头戴式显示设备的前提下，为远程用户提供接近面对面交流的共同在场感。围绕这一目标，学术界与工业界提出了多种具有

代表性的系统原型，涵盖从基于几何重建的 FVV 系统，到以体积视频与隐式表示为核心的实时流式系统，再到以神经场与三维高斯表示为代表的新一代通信架构。图 2.1 展示了近年来具有代表性的几款沉浸式通信系统，表 2.1 对现有典型沉浸式视频通信系统在采集、编码与传输、渲染、显示、时延与硬件资源等方面的关键参数进行了对比。

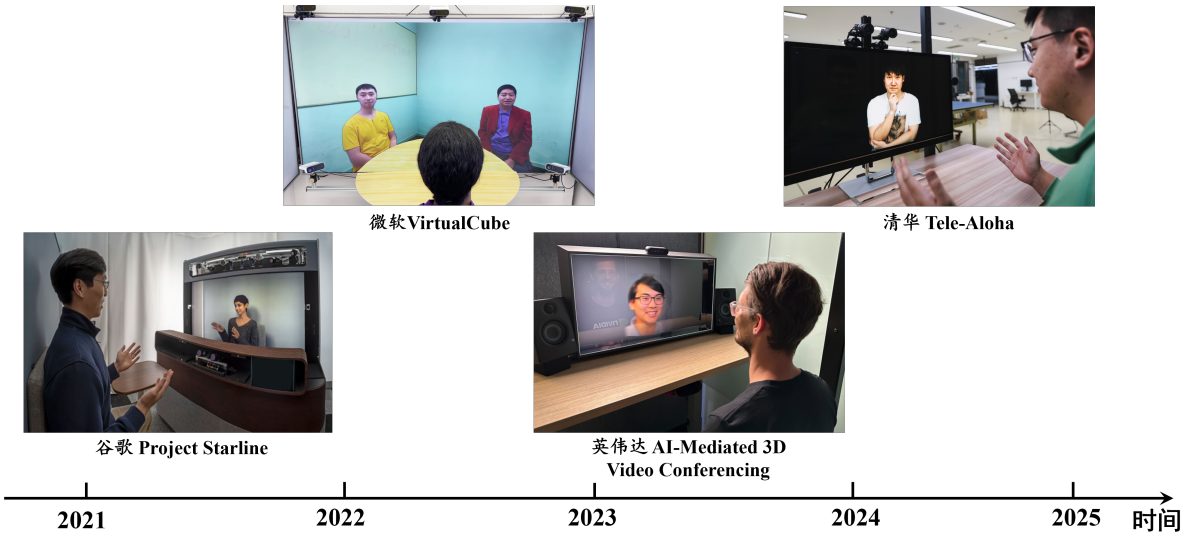


图 2.1　沉浸式通信系统发展

FVV Live[60] 面向事件直播和多人视频会议场景，构建了一套端到端 FVV 传输方案。系统通过 9 路 ZED RGBD 摄像机获取多视点纹理与深度信息，在编码侧采用颜色与深度分离的策略：彩色视频使用 H.264/AVC 编码以节省带宽，深度数据则采用无损压缩以保证视点合成质量，并通过 RTP/UDP 实现低时延传输。在渲染端，系统基于深度图像渲染（DIBR）完成新视点合成，并结合裂缝补偿与遮挡修复等后处理，提高虚拟视点的视觉一致性。FVV Live 的优势在于采用商用深度相机与成熟的视频编码工具，系统架构清晰、部署门槛较低，能够在 1080p、30 FPS 条件下实现约 250–360 ms 的端到端时延，适用于中等规模的沉浸式会议。然而，该系统对深度质量高度敏感，在复杂遮挡、细粒度几何和非刚体场景下仍存在伪影，且多相机同步与校准的工程代价较高。

VirtualCube 系统[2] 则面向小规模多人会议场景，通过虚拟立方体会议舱的形态实现三维沉浸式沟通。其采集端由 6 路 Kinect RGBD 摄像机构成，利用多视图立体（MVS）和深度估计得到用户的三维表示，并在渲染端结合深度引导的纹理融合与光照建模生成多视点视图。系统通过三块 65 英寸 4K 显示屏呈现近似等身比例的远程人物，并通过环绕布局恢复自然的视线关系。VirtualCube 在显示与空间交互上显著优于传统 2D 视频会议，能够支持 23–30 FPS 的实时通信，但其整体延迟约为 300 ms，且需要多块高端 GPU 支撑复杂的 MVS 与渲染流程，硬件与部署成本较高。

Project Starline[1] 是目前工业界最具代表性的高保真沉浸式远程临场系统之一，完

表 2.1　典型沉浸式通信系统对比

沉浸式通信系统	年份	公　司/机构	采集	编码/传输	渲染	显示	帧　率/时延	硬件
FVV Live[60]	2020	-	9×RGBD (ZED)	AVC RTP/UDP	DIBR	1080p 显示器	30 FPS 361ms	3×(GTX1080+P4000) 1×GTX1080
Project Starline[1]	2021	Google	3×RGBD	HEVC WebRTC	光线投射算法 + 加权彩色融合	65 inch 8K 追踪式自由立体显示器	60 FPS 105ms	2×RTX 6000 2×Titan RTX
VirtualCube[2]	2021	Microsoft	6×RGBD (Kinect)	JPEG TCP/IP	MVS 深度估计 + 纹理融合	3× 65 inch 4K (1280×960)	23-30 FPS 300ms	2×RTX 3090 1×RTX 2080
Live4D[61]	2023	ByteDance	10×　双目 RGB	NVIDIA Video SDK + Draco	隐式表面重建	/	24 FPS 95ms	4×RTX 3090 1×RTX 4090
AI-mediated 3DVC[62]	2023	Nvidia	1×RGB	-	基于 ViT 三平面表示 + NeRF / 眼神交互	眼动追踪 3D 显示器	100ms	1×A6000 Ada 1×RTX 4090
Tele-Aloha[3]	2024	THU	4×RGB	HEVC WebRTC	Feed-forward 3DGS	眼动追踪 3D 显示器 (2048×2048)	30 FPS 150ms	1×RTX 4090
Google Beam	2025	Google	6×RGB	-	-	光场显示器	60 FPS	-

成了对采集、重建、编码、渲染、显示全链路的系统级协同优化。系统使用多组 RGB 与 NIR 摄像头构成的采集阵列，在 60–180 Hz 采样频率下生成多视图彩色和深度信息，并通过基于 ESPReSSo[63] 的时空立体算法和 TSDF 融合[36] 完成高精度几何重建。编码侧采用 HEVC/H.265 对多路彩色和深度流进行压缩，并通过 WebRTC 进行低时延传输；显示侧则使用 65 英寸 8K 头跟踪式裸眼立体显示器，结合精确的眼位追踪和空间音频重建，实现高度自然的三维临场感。该系统在 60 FPS 的帧率下能够将端到端时延控制在约 105 ms 量级，在用户主观体验和行为指标方面显著优于二维视频会议。其主要不足在于硬件成本和系统复杂度极高，短期内难以大规模部署到普通办公或家庭环境。

与上述以显式或 TSDF 几何为中心的系统不同，Live4D[61] 将体积视频重建与隐式神经表示相结合，面向 VR/AR 场景提供低成本、低带宽的实时体积视频流。系统通过稀疏布置的双目相机阵列获取多视图图像，在重建端借助局部 TSDF 与神经隐式场（Neural SDF）联合建模，生成水密网格，并使用 Draco 等几何压缩工具大幅降低传输数据量。在优化的 CUDA/TensorRT 实现下，Live4D 在中高端 GPU 上可实现 24 FPS 重建与约 95 ms 的端到端时延，相比传统体积视频方案在时延与带宽上均有数量级的改进。其优势在于重建结果更适合与现有 VR/AR 渲染管线结合，但系统仍需依赖较强的 GPU 算力，且隐式场训练与更新对动态场景的适应性仍有限。

NVIDIA 提出的 AI-mediated 3D Video Conferencing 系统[62] 则通过单目采集和三平面表示的轻量化路线实现三维通信。该系统在采集端仅使用一台普通 RGB 摄像头获

取二维视频流，通过基于 ViT 的编码器将图像映射为紧致的三平面隐式表示，再在接收端通过神经渲染生成等身大小的三维人像。为了提升社交临场感，系统还利用专门的视线重定向算法[64] 实现自然的眼神交互，并支持裸眼立体和光场显示设备。在编码与传输层面，系统保持与传统视频会议近似的带宽占用，使三维通信在消费级终端上成为可能。其主要优势在于极大降低了多摄像机布设与几何标定的工程门槛，但在复杂动作、快速运动和大范围场景变化下，基于学习的升维模型仍可能出现漂移与伪影。

Tele-Aloha[3] 是近年来具有代表性的低成本沉浸式通信系统之一，其突出特点是在稀疏多视图采集与 3DGS[5] 渲染之间构建了一条高效的技术路径。系统仅依赖四台消费级 RGB 摄像机采集用户上半身影像，在编码端利用 HEVC 和 WebRTC 完成低时延压缩与传输，并在接收端通过级联视差估计与基于 3DGS 的前馈新视点合成网络实现 2048×2048 分辨率、30 FPS 的实时渲染。得益于 3DGS 在渲染效率和表达能力上的优势，Tele-Aloha 在单张 RTX 4090 GPU 上即可完成完整的三维重建与渲染流程，将端到端时延控制在约 150 ms 内。Tele-Aloha 在硬件和部署成本上大幅下降，更符合未来家庭或小型办公场景的应用需求。

综合上述代表性系统可以看出，当前沉浸式通信技术在采集、编码传输、渲染与显示等环节各有取舍。基于 RGBD 相机采集与 DIBR/MVS 的方案在工程上相对成熟，适合中等成本的多视点会议场景，但在视点灵活性和细节表现方面存在天花板；基于 TSDF 与隐式神经场的体积视频方案在几何与外观质量上更加出色，但训练与推理开销较大；三平面与三维高斯驱动的方案则在硬件门槛和实时性之间取得了更佳平衡，代表了面向消费级应用的重要方向。总体来看，当前沉浸式通信系统尚未形成统一的最佳实践路线，技术选择高度依赖目标场景的带宽条件、算力预算、用户规模和体验需求等因素，这也为后续研究在系统架构设计与跨层协同优化上留下了广阔空间。

第三章　基于光流预测的快速视点合成方法

3.1　引言

在沉浸式自由视点通信中，接收端往往需要根据观察者位置动态生成与当前视点相匹配的图像内容，以满足用户视角连续变化下的裸眼 3D 稳定显示的要求。

现有视点合成方法大多建立在显式几何建模基础之上，例如基于深度图像的渲染（DIBR）框架通常需要依次完成深度估计或深度融合、纹理映射、空洞修复与目标视点渲染等多个处理环节。该类方法在多数场景下能够保持较稳定的空间结构，但各处理阶段之间存在较强依赖，对深度估计与相机标定误差较为敏感。当场景中存在遮挡、边缘细节、高频纹理以及快速运动目标时，深度误差往往会被逐级放大，进而导致目标视图出现错位、撕裂、空洞和边界伪影等问题。此外，显式几何重建通常需要较大的计算与传输开销，在实时通信场景下容易受到时延与硬件资源限制。

近年来，基于神经网络的视图插值与视频帧插值方法快速发展，为快速视点合成提供了新的实现方式。其中，光流驱动的图像变形方法能够直接在图像域建立不同视图之间的像素对应关系，通过估计目标视图到参考视图之间的运动场，实现目标视图的高效重建。与传统依赖显式深度建模的方案相比，这类方法避免了复杂的三维几何恢复过程，能够以更低的计算开销完成视点合成任务。特别是 RIFE[45] 实时光流插值模型，采用多尺度光流估计与逐级细化结构，在保持较低计算开销的同时能够实现高效推理，适用于实时视点合成任务。

然而，面向实际沉浸式通信系统部署时，直接采用通用光流插值模型仍然面临挑战，沉浸式显示通常要求系统在较高分辨率和较高帧率下稳定生成双目视图，这对模型的推理速度和显存占用提出了严格要求。因此，如何在保证合成质量的前提下，降低视点合成算法的计算复杂度，实现适配沉浸式通信场景的实时部署，是沉浸式自由视点视频通信中的关键问题。

针对上述问题，本章围绕基于光流预测的快速视点合成方法展开研究，旨在设计一种适用于沉浸式通信接收端的实时视图生成方法。该方法以相邻参考视图及目标视点插值参数为输入，利用多尺度光流预测网络直接建立目标视图与参考视图之间的像素映射关系，通过图像扭曲与融合实现新视点图像重建，从而避免传统深度估计、几何重建与显式渲染所带来的复杂处理流程。在此基础上，本章进一步结合模型轻量化与工程部署需求，从输入尺度压缩、半精度量化以及高性能推理框架实现等方面对网络推理过程进行加速优化，以满足高分辨率、低时延沉浸式显示场景下的实时运行需求。

本章的研究目标主要体现在以下三个方面：其一，构建适用于双参考视图输入的目标视点合成模型，建立视点插值参数与像素变换之间的映射关系；其二，分析并设计面向实时应用的多尺度光流预测与图像融合机制，提高模型在遮挡区域、边缘区域和运动区域中的合成鲁棒性；其三，结合沉浸式通信系统的资源约束，对模型推理流程进行低复杂度优化，在保证合成质量基本稳定的前提下显著提升运行速度。通过上述研究，本章为后续沉浸式通信系统中的双目视图实时生成提供算法基础，也为后续系统集成与性能验证奠定方法支撑。

本章后续内容首先对视点合成任务进行形式化建模，随后介绍基于多尺度光流预测的快速视点合成方法，并阐述采用的网络训练策略与推理加速方案，最后通过实验对所提出方法的生成性能和运行效率进行分析与验证。

3.2　问题描述与视点合成框架

为实现沉浸式通信场景下的实时新视点生成，本节对视点合成任务进行形式化描述。本章关注的场景为：发送端通过线性排列的多摄像机阵列采集人物图像，接收端根据观察者当前位置，从与目标视点最邻近的两个参考摄像机视图出发，实时生成与当前观察位置相匹配的目标视图。

首先给出视点合成问题描述与基本假设。设经过标定、去畸变和极线校正后的两幅相邻参考视图分别记为 $I_L, I_R \in \mathbb{R}^{H\times W\times 3}$。其中，$H$ 和 W 分别表示图像的高度与宽度，I_L 和 I_R 对应于目标视点左右两侧最邻近的参考相机所采集的彩色图像。目标是根据参考视图以及目标视点在两参考视图之间的位置关系，生成目标视图 $I_V \in \mathbb{R}^{H\times W\times 3}$。

为便于问题建模，本章作如下基本假设：

1. 多摄像机经过离线标定与校正后，其图像被映射到统一参考成像平面，不同视图之间的主要差异体现在由视差引起的像素位移；
2. 摄像机沿近似水平方向线性排列，观察者在显示区域内的主要运动也以水平方向变化为主，因此目标视图相对于参考视图的几何变化主要表现为水平视差及其局部非刚性扰动；
3. 为降低建模难度并提高生成稳定性，目标视图始终由其左右两侧最邻近的参考视图共同合成，即本章主要考虑插值型视点合成问题，而不考虑远距离外推型视图生成。

在此基础上，视点合成问题可形式化表示为如下映射关系：

$$I_V = \mathcal{G}(I_L, I_R, v), \tag{3.1}$$

其中，v 表示目标视图在两参考视图之间的插值参数，$\mathcal{G}(\cdot)$ 表示由参考视图到目标视图

的视点合成映射函数。

接下来给出目标视点的参数化表示。设左、右参考摄像机的光心在统一基线坐标系中的位置分别为 p_L 和 p_R，目标视点位置为 p_V，且满足 $p_L < p_V < p_R$。则目标视点相对于左右参考视图的位置可由归一化插值参数 v 描述为

$$v = \frac{p_V - p_L}{p_R - p_L}, \qquad v \in (0, 1). \tag{3.2}$$

当 $v = 0$ 时，目标视点与左参考视图重合；当 $v = 1$ 时，目标视点与右参考视图重合；当 $v \in (0, 1)$ 时，目标视图位于两参考视图之间。

对于沉浸式显示场景，观察者左眼和右眼通常对应两个彼此邻近但不完全相同的目标视点。设左眼和右眼对应的目标视点分别为 p_{V_l} 和 p_{V_r}，则可得到两个插值参数

$$v_l = \frac{p_{V_l} - p_L}{p_R - p_L}, \qquad v_r = \frac{p_{V_r} - p_L}{p_R - p_L}. \tag{3.3}$$

传统基于显式深度建模的方法通常先恢复场景几何，再依据投影关系完成新视点渲染。与之不同，本章采用图像域直接变换思路，将目标视图生成建模为从目标视点到参考视图的反向像素映射问题。

设目标视图中像素位置为 $\mathbf{x} = (x, y)^\top$，定义从目标视点到左、右参考视图的反向光流场分别为 $F_{V\to L}$, $F_{V\to R} \in \mathbb{R}^{H\times W\times 2}$。其中，$F_{V\to L}(\mathbf{x})$ 表示目标视图像素 $\mathbf{x}$ 在左参考图像中的对应采样偏移，$F_{V\to R}(\mathbf{x})$ 具有相同含义。于是，左右参考图像经反向扭曲后的结果分别可表示为

$$\begin{aligned} \tilde{I}_L(\mathbf{x}) &= \mathscr{W}(I_L, F_{V\to L})(\mathbf{x}) = I_L\big(\mathbf{x} + F_{V\to L}(\mathbf{x})\big), \\ \tilde{I}_R(\mathbf{x}) &= \mathscr{W}(I_R, F_{V\to R})(\mathbf{x}) = I_R\big(\mathbf{x} + F_{V\to R}(\mathbf{x})\big), \end{aligned} \tag{3.4}$$

其中，$\mathscr{W}(\cdot)$ 表示可微扭曲算子，实际实现中通常采用双线性插值完成非整数坐标的像素采样。

由于左右参考视图对目标像素的可见性并不完全一致，在遮挡区域、边缘区域或反射区域中，单纯依赖某一侧参考图像往往难以获得稳定重建结果。为此，引入融合掩模 $M \in [0, 1]^{H\times W}$，其中，$M(\mathbf{x})$ 表示目标像素位置 $\mathbf{x}$ 处左参考视图对最终重建结果的贡献权重，则目标视图可以表示为

$$\hat{I}_V(\mathbf{x}) = M(\mathbf{x}) \odot \tilde{I}_L(\mathbf{x}) + \big(1 - M(\mathbf{x})\big) \odot \tilde{I}_R(\mathbf{x}), \tag{3.5}$$

其中，$\odot$ 表示逐元素乘法运算，$\hat{I}_V$ 为合成得到的目标视图。

据此，式 (3.1) 所定义的视点合成映射可以分解为两个子过程：

$$\begin{aligned}\left(F_{V\to L}, F_{V\to R}, M\right) &= \Phi_{\Theta}(I_L, I_R, \upsilon),\\ \hat{I}_V &= \Psi\left(I_L, I_R, F_{V\to L}, F_{V\to R}, M\right),\end{aligned} \tag{3.6}$$

其中，$\Phi_{\Theta}(\cdot)$ 表示由参数 Θ 决定的光流与融合掩模预测模型，$\Psi(\cdot)$ 表示基于反向扭曲与加权融合的图像重建过程。这样，原始的视点合成问题被转化为一个以参考图像和目标视点参数为输入、以反向光流与融合权重为中间表示的学习问题。

3.3　基于多尺度光流预测的快速视点合成

本节给出面向沉浸式自由视点视频通信场景的快速视点合成方法。所提方法以目标视点两侧最邻近的参考视图作为输入，通过多尺度光流预测网络直接估计目标视点到参考视图的反向光流场，并结合融合掩模完成虚拟视图重建。与传统依赖深度估计、三维几何重建和显式渲染的处理流程不同，所提方法直接在图像域建立跨视点像素对应关系，避免了复杂的几何建模与中间表示传输过程，从而更适合于高分辨率、低时延的实时沉浸式显示场景。

3.3.1　参考视图预处理与几何校正

在多视点采集场景中，相邻参考视图之间的几何一致性会直接影响后续光流预测与虚拟视图重建效果。若输入参考视图之间存在较大的成像差异，则对应像素在不同视图中的位移幅度会显著增加，从而提高像素匹配难度，降低光流估计的稳定性。尤其在人像边缘、手部动作以及局部遮挡区域中，较大的跨视点形变更容易引入匹配误差，进而影响最终合成视图的连续性与真实感。因此，在进入视点合成网络之前，首先需要对原始采集图像进行预处理，将不同相机采集到的视图尽可能映射到统一、规整的标准成像平面。本节将该预处理过程划分为去畸变与几何校正两个步骤。

由于工业相机和镜头在实际成像过程中不可避免地存在径向畸变与切向畸变，原始采集图像中直线结构往往会发生弯曲，且图像不同区域的局部放缩关系也不完全一致。如果直接将这类图像输入光流预测网络，则网络除了需要学习由场景视差引起的像素位移外，还必须额外补偿镜头畸变所带来的几何偏差，这会增加网络建模难度，不利于后续视点合成。

设第 i 个摄像机采集到的原始图像记为 $\boldsymbol{I}_i^{raw}$，其内参矩阵和畸变参数分别记为 $\boldsymbol{K}_i$ 和 $\boldsymbol{D}_i$。首先通过相机标定得到各摄像机的成像模型，并对原始图像进行去畸变处理，记去畸变后的图像为

$$\boldsymbol{I}_i^{und} = \mathscr{U}(\boldsymbol{I}_i^{raw}; \boldsymbol{K}_i, \boldsymbol{D}_i), \tag{3.7}$$

其中，$\mathcal{U}(\cdot)$ 表示去畸变映射函数。经过该步骤后，径向畸变和切向畸变带来的局部非线性形变得到有效校正，从而为后续的视图对齐奠定基础。

完成去畸变后，不同相机图像之间仍然可能由于安装姿态误差、视轴偏移以及采集空间限制等因素而存在成像平面不一致的问题。为了减小相邻参考视图之间的运动幅度，本节对去畸变后的图像进行几何校正，将其重投影到预先定义的标准成像平面上。

设第 i 个相机去畸变后的原始视图为 $\boldsymbol{I}_i^{und}$，其实际成像姿态对应的旋转矩阵记为 $\boldsymbol{R}_i^*$。在系统设计阶段，根据预设的相机阵列几何布局，可以为该视图指定一个目标标准成像平面，其对应的理想旋转矩阵记为 $\boldsymbol{R}_i$。若标准视图对应的内参矩阵记为 $\boldsymbol{K}_i^r$，则原始视图到标准视图之间的单应矩阵可写为

$$\boldsymbol{H}_i = \boldsymbol{K}_i^r \boldsymbol{R}_i (\boldsymbol{R}_i^*)^{-1} \boldsymbol{K}_i^{-1}. \tag{3.8}$$

式 (3.8) 表明，只要通过离线标定获得相机内参与实际拍摄姿态，并根据标准采集几何关系确定目标视图的理想姿态，就可以直接计算出原始视图到标准视图的单应矩阵。对于原始视图中的任意像素点 $\boldsymbol{x}^* = [x^*, y^*, 1]^\top$，其在标准视图中的对应位置 $\boldsymbol{x} = [x, y, 1]^\top$ 满足

$$\boldsymbol{x} \sim \boldsymbol{H}_i \boldsymbol{x}^*, \tag{3.9}$$

其中符号 “$\sim$” 表示齐次坐标意义下的相等关系。根据上述映射关系，可以将去畸变后的原始图像重新采样到目标标准平面上，得到几何校正后的参考视图

$$\boldsymbol{I}_i = \mathrm{Remap}(\boldsymbol{I}_i^{und}; \boldsymbol{H}_i), \tag{3.10}$$

其中，$\mathrm{Remap}(\cdot)$ 表示基于单应矩阵的图像重投影操作，实际实现中采用双线性插值完成非整数坐标处的像素采样。

需要指出的是，该过程并不改变相机在真实空间中的位置，而是利用已标定的相机参数，将实际采集到的原始视图映射到预先定义的标准成像平面。这样做的目的在于尽可能减小相邻参考视图之间的姿态差异，使其几何变化主要反映真实场景视差，而非额外的成像偏差。经过几何校正后，相邻视图之间的像素运动幅度得到有效控制，从而有利于后续光流预测网络学习更稳定的跨视点对应关系。

图 3.1 给出了本章视点合成流程的示意图。首先，将原始参考视图 $\boldsymbol{I}_L^*$ 和 $\boldsymbol{I}_R^*$ 经过去畸变与单应重投影，校正为标准视图 $\boldsymbol{I}_L$ 和 $\boldsymbol{I}_R$；随后，以校正后的标准视图作为输入，预测参考视图到目标虚拟视图的光流场，并完成虚拟视图 $\hat{\boldsymbol{I}}_V$ 的重建。可以看出，去畸变与几何校正是整个视点合成流程的前置步骤，其作用并非直接生成新视图，而是通过统一参考视图的成像几何关系来降低后续光流估计的难度。

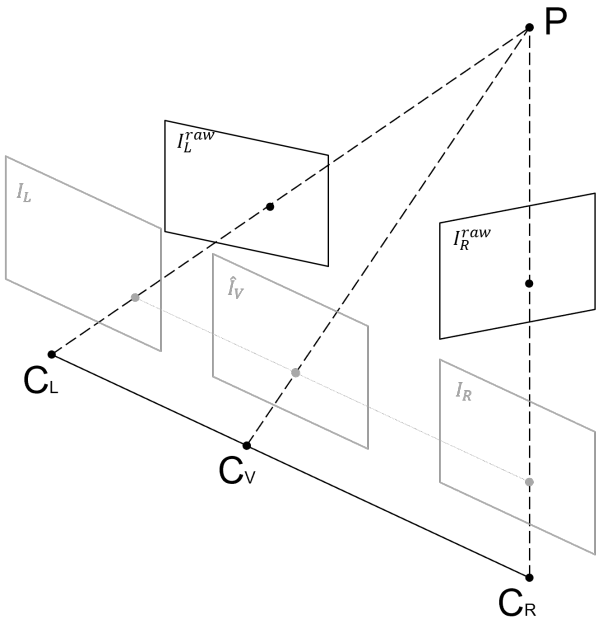


图 3.1　视点合成流程示意图

经过上述处理后，不同视图被尽可能对齐到统一的参考平面，跨视点差异主要表现为由场景深度变化所引起的视差位移，而由镜头畸变和相机姿态误差引入的额外干扰被显著抑制。在此基础上，再从几何位置上选取目标视点两侧最邻近的参考视图作为网络输入。设当前目标视点对应的左右参考图像分别为 $\boldsymbol{I}_L$ 和 $\boldsymbol{I}_R$，其选择依据由接收端眼动跟踪模块提供的观察者双眼位置确定。由于目标视点始终位于两参考视图之间，因此本章仅处理插值型视点合成问题，从而保证合成结果具有更好的几何可解释性。

3.3.2　多尺度光流预测网络结构

在完成参考视图预处理与几何校正后，采用基于多尺度级联细化的光流预测网络对目标视点进行建模。网络结构如图 3.2 所示。该网络以左右参考图像和目标视点插值参数为输入，直接预测目标视点到左右参考视图的反向光流场以及融合掩模，并在逐层细化的过程中同时维护中间特征表示，以提升不同尺度下的像素对应建模能力。

设经过预处理后的左右参考图像分别为 $\boldsymbol{I}_L$ 和 $\boldsymbol{I}_R$，目标视点插值参数为 $v \in (0,1)$。为将视点位置信息显式注入网络，首先构造与输入图像同尺寸的单通道参数图

$$V(x,y) = v, \qquad \forall (x,y) \in \Omega, \tag{3.11}$$

其中，Ω 表示图像定义域。该参数图在空间维度上取值恒定，但在通道拼接后能够为网络提供目标视点相对于左右参考视图的位置信息。

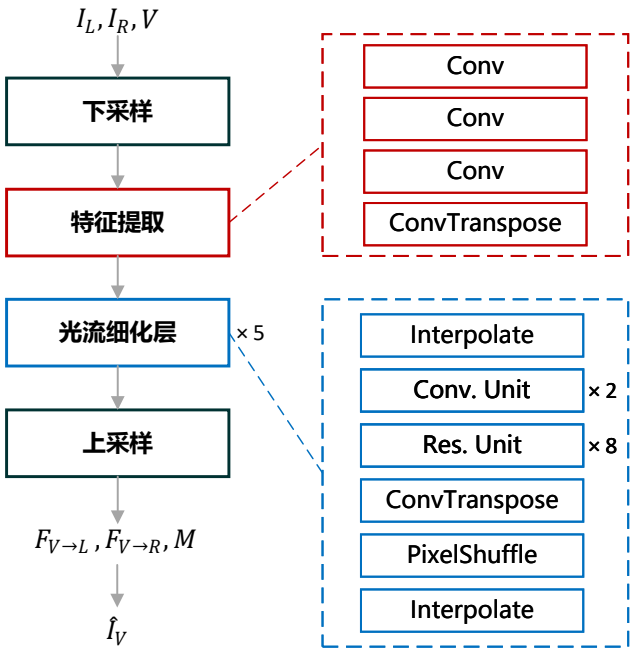


图 3.2　多尺度级联细化的光流预测网络结构

3.3.2.1　浅层图像特征提取

首先利用轻量级图像特征提取器对左右参考图像进行编码。设该特征提取器记为 $\mathscr{E}(\cdot)$，则左右参考图像对应的浅层特征分别表示为

$$f_L = \mathscr{E}(I_L), \quad f_R = \mathscr{E}(I_R). \tag{3.12}$$

图像特征提取器由四层卷积操作构成。前三层采用卷积核大小为 3×3 的卷积运算，其中第一层步长为 2，用于在扩大感受野的同时压缩空间分辨率；第二层和第三层步长均为 1，用于进一步提取局部纹理特征。最后一层采用转置卷积对特征进行上采样，并将输出映射为较低通道数的特征表示。各层卷积之后均接 LeakyReLU 激活函数。由此，网络能够在保持较低计算复杂度的前提下，从左右参考图像中提取适用于跨视点匹配的浅层视觉特征。在后续各层光流细化过程中，这些图像特征不仅作为首层光流估计的输入，而且还会依据当前光流场进行同步扭曲，从而为后续细化层提供与当前目标视点更一致的中间表征。

3.3.2.2　多尺度级联光流预测框架

本节所采用的多尺度光流预测网络由 $K = 5$ 个串联的光流细化层组成，每个层分别工作在不同尺度下，其尺度系数设定为

$$\{s_0, s_1, s_2, s_3, s_4\} = \left\{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\right\}. \tag{3.13}$$

随着层数加深，网络逐步从粗尺度过渡到原始分辨率尺度，以实现由粗到细的反向光流估计与图像融合。

在第 k 层，网络输出目标视点到左右参考视图的四通道反向光流 $F^{(k)} = \left[F_{V\to L}^{(k)},\ F_{V\to R}^{(k)}\right]$，其中 $F_{V\to L}^{(k)} \in \mathbb{R}^{H\times W\times 2}$ 和 $F_{V\to R}^{(k)} \in \mathbb{R}^{H\times W\times 2}$ 分别表示从目标视图到左、右参考图像的二维反向位移场。同时，第 k 层还输出融合掩模 $M^{(k)}$ 以及中间特征 $Z^{(k)}$，用于后续层的进一步细化。因此，第 k 层可统一表示为

$$\left(F^{(k)}, M^{(k)}, Z^{(k)}\right) = \Phi^{(k)}(\cdot). \tag{3.14}$$

在第一层中，网络直接利用左右参考图像、对应图像特征以及视点参数图生成初始光流、掩模和中间特征。其输入可表示为

$$X^{(0)} = \mathrm{Concat}(I_L, I_R, f_L, f_R, V), \tag{3.15}$$

其中 Concat(·) 表示沿通道维拼接操作。于是，首层预测过程可写为

$$\left(F^{(0)}, M^{(0)}, Z^{(0)}\right) = \Phi^{(0)}\left(X^{(0)}; s_0\right). \tag{3.16}$$

由于首层在最粗尺度下工作，其主要任务是根据左右参考图像之间的整体结构差异和目标视点参数，建立初始的跨视点像素对应关系。该层所估计的光流场主要负责捕获主体轮廓、躯干位置和大范围位移变化，为后续更高分辨率下的细化过程提供初始条件。

从第二层开始，网络采用递进式残差细化策略，对已有光流进行逐层修正。设第 $k-1$ 层得到的反向光流为 $F^{(k-1)}$，则首先利用该光流对左右参考图像及其图像特征进行扭曲，得到

$$\begin{aligned} \hat{I}_L^{(k-1)} &= \mathscr{W}\left(I_L, F_{V\to L}^{(k-1)}\right), \qquad \hat{I}_R^{(k-1)} = \mathscr{W}\left(I_R, F_{V\to R}^{(k-1)}\right), \\ \hat{f}_L^{(k-1)} &= \mathscr{W}\left(f_L, F_{V\to L}^{(k-1)}\right), \qquad \hat{f}_R^{(k-1)} = \mathscr{W}\left(f_R, F_{V\to R}^{(k-1)}\right). \end{aligned} \tag{3.17}$$

随后，将扭曲后的图像、扭曲后的图像特征、视点参数图、上一层的融合掩模以及上一层的中间特征进行拼接，构成当前层的主输入：

$$X^{(k)} = \mathrm{Concat}\left(\hat{I}_L^{(k-1)}, \hat{I}_R^{(k-1)}, \hat{f}_L^{(k-1)}, \hat{f}_R^{(k-1)}, V, M^{(k-1)}, Z^{(k-1)}\right), \qquad k = 1,2,3,4. \tag{3.18}$$

与此同时，上一层的光流 $F^{(k-1)}$ 也会作为额外输入参与当前层计算。于是，第 k 层的预测过程可表示为

$$\left(\Delta F^{(k)}, M^{(k)}, Z^{(k)}\right) = \Phi^{(k)}\left(X^{(k)}, F^{(k-1)}; s_k\right), \qquad k = 1,2,3,4. \tag{3.19}$$

其中，$\Delta F^{(k)}$ 表示当前层对光流的修正量。由此得到第 k 层更新后的光流为

$$F^{(k)} = F^{(k-1)} + \Delta F^{(k)}. \tag{3.20}$$

3.3.2.3 光流细化层内部结构

每一个光流细化层都包含下采样、特征变换和上采样三个阶段。设输入特征为 X，尺度系数为 s_k，则首先采用双线性插值将输入缩放到当前层对应的工作分辨率：

$$X_s = \mathrm{Interpolate}(X; s_k), \tag{3.21}$$

若当前层不是首层，则还需将输入光流同步缩放并在通道维进行拼接：

$$\begin{aligned} F_s &= s_k \cdot \mathrm{Interpolate}(F^{(k-1)}; s_k), \\ \tilde{X}_s &= \mathrm{Concat}(X_s, F_s). \end{aligned} \tag{3.22}$$

在下采样阶段，依次使用两层步长为 2 的卷积层提取特征，并逐步将通道数映射到当前层所设定的特征维度。随后，在特征变换阶段，网络串联多个残差卷积单元，对特征进行非线性建模。其中每个残差卷积单元均由 3×3 卷积和 LeakyReLU 激活构成，并采用残差连接保留输入信息，以提高深层特征传播的稳定性。

在上采样阶段，网络利用转置卷积和 PixelShuffle 操作恢复空间分辨率，再通过双线性插值映射回原始尺度。最终输出被划分为三部分：前 4 个通道对应左右反向光流，接下来的 1 个通道对应融合掩模，其余通道作为中间特征传递到下一层。

在多层级联细化全部完成后，网络得到最终的累积反向光流 $F^{(4)}$ 和融合掩模 $M^{(4)}$。此时，再次利用最终光流对左右参考图像进行反向扭曲，得到最终对齐结果：

$$\hat{I}_L = \mathscr{W}\left(I_L, F^{(4)}_{V\to L}\right), \qquad \hat{I}_R = \mathscr{W}\left(I_R, F^{(4)}_{V\to R}\right). \tag{3.23}$$

为了将融合掩模约束到 $[0, 1]$ 区间，对其施加 Sigmoid 激活：

$$\bar{M} = \sigma\left(M^{(4)}\right), \tag{3.24}$$

其中，$\sigma(\cdot)$ 表示 Sigmoid 函数。最终虚拟视图通过式 3.5 将左右扭曲结果的逐像素加权融合得到。对于目标视图中的不同空间位置，融合掩模 $\bar{M}$ 会自适应地决定左右参考图像的贡献比例：当左参考图像在当前像素处的对齐结果更可靠时，$\bar{M}$ 的取值更接近 1；反之，当右参考图像更可信时，$\bar{M}$ 更接近 0。借助这种自适应融合机制，网络能够在遮挡切换、边界错位和局部纹理差异较大的区域获得更加平滑和自然的重建结果。

3.4　网络训练策略与推理加速方法

在实际训练与部署过程中，仅依赖网络结构本身难以同时兼顾重建精度与实时性能。训练数据通常只包含目标视点的真实图像，而不提供中间光流场和融合掩模的显式监督，因此网络需要在弱监督条件下学习目标视点到参考视图的像素对应关系。沉浸式通信系统要求接收端在高分辨率和高帧率条件下稳定生成双目视图，这对模型推理效率提出了严格要求。基于此，本节分别从训练策略和推理加速两个方面展开讨论：针对缺乏显式光流监督的问题，采用特权蒸馏训练策略提升学生模型的优化效果；结合实际部署需求，对模型推理流程进行轻量化与工程化加速，以满足实时系统的运行约束。

(1) 特权蒸馏训练策略

本章所研究的视点合成任务属于典型的弱监督图像生成问题。对于训练样本，通常仅能够获得目标视点对应的真实图像 I_V^{GT}，而无法直接获得目标视点到左右参考视图的真实光流场以及融合掩模。因此，若仅依赖最终重建误差来优化网络，模型需要同时完成像素对应关系学习、遮挡区域建模和图像融合权重估计，训练难度较大，尤其在边界、弱纹理和局部遮挡区域中容易出现优化不稳定的问题。

为缓解上述困难，本节引入特权蒸馏策略。在训练阶段，除原始学生模型外，额外构造一个仅在训练阶段使用的教师模型。教师模型与学生网络的基本结构保持一致，但在输入端额外引入目标视点的真实图像 I_V^{GT} 作为特权信息，从而利用更充分的先验辅助光流场和掩模的学习。由于教师模型能够访问学生模型在测试阶段不可获得的真实中间视图，因此其预测结果通常比学生模型更精确，可用于为学生模型提供额外监督。

设学生模型经过最后一层细化后得到的累积反向光流和掩模分别为 $F^{(4)}$ 和 $M^{(4)}$，对应的学生模型合成结果为 $\hat{I}_V$。在此基础上，教师模型被附加在学生模型最后一个细化模块之后，其输出可表示为

$$(F^{(\mathrm{Tea})}, M^{(\mathrm{Tea})}) = (F^{(4)}, M^{(4)}) + \Phi^{\mathrm{Tea}}\left(F^{(4)}, M^{(4)}, \hat{I}_V, V, I_V^{GT}\right), \tag{3.25}$$

其中，Φ^{Tea} 表示教师模型对应的附加细化模块，V 为目标视点参数图。与学生模型不同，教师模型在细化过程中显式引入真实目标视图 I_V^{GT}，从而能够生成更准确的光流场 $F^{(\mathrm{Tea})}$ 和融合掩模 $M^{(\mathrm{Tea})}$。

通过这种方式，教师模型实际上承担了训练期辅助优化器的角色。其作用并不是替代学生模型完成最终推理，而是在训练阶段为学生模型提供更可靠的目标光流方向，帮助学生模型在复杂区域中学习更合理的像素对应关系。教师模型仅在训练阶段存在，在测试和实际部署阶段将被完全移除，因此不会引入额外的推理开销。

(2) 损失函数设计

在特权蒸馏框架下，总损失由学生模型重构损失、教师模型重构损失以及蒸馏损失三部分组成。

蒸馏损失用于约束学生模型与教师模型在最终光流预测上的一致性。设学生模型最终输出的左右反向光流分别为 $F_{V\to L}$ 和 $F_{V\to R}$，教师模型输出的对应光流分别为 $F_{V\to L}^{(\mathrm{Tea})}$ 和 $F_{V\to R}^{(\mathrm{Tea})}$，则蒸馏损失表示为

$$\mathscr{L}_{dis} = \left\| F_{V\to L} - F_{V\to L}^{(\mathrm{Tea})} \right\|_2 + \left\| F_{V\to R} - F_{V\to R}^{(\mathrm{Tea})} \right\|_2 . \tag{3.26}$$

该损失鼓励学生模型的光流预测尽可能逼近教师模型输出，从而在缺少显式光流标注的条件下提升网络的收敛稳定性。训练过程中，蒸馏损失的梯度仅回传至学生模型，教师模型本身不通过该项损失进行反向更新。

学生模型和教师模型使用目标视图进行图像重建监督。学生模型和教师模型的重构损失分别为

$$\begin{aligned} \mathscr{L}_{rec} &= d(\hat{I}_V, I_V^{GT}), \\ \mathscr{L}_{rec}^{(\mathrm{Tea})} &= d(\hat{I}_V^{(\mathrm{Tea})}, I_V^{GT}), \end{aligned} \tag{3.27}$$

其中，其中，$\hat{I}_V$ 表示学生模型合成的目标视图，$\hat{I}_V^{(\mathrm{Tea})}$ 表示教师模型基于 $F^{(\mathrm{Tea})}$ 和 $M^{(\mathrm{Tea})}$ 得到的目标视图。

综合上述三项约束，得到总损失函数：

$$\mathscr{L}_{total} = \mathscr{L}_{rec} + \lambda_{\mathrm{Tea}} \mathscr{L}_{rec}^{(\mathrm{Tea})} + \lambda_{dis} \mathscr{L}_{dis}, \tag{3.28}$$

对于式 (3.27) 中的距离度量函数 $d(\cdot,\cdot)$，本节采用基于拉普拉斯金字塔的多尺度 L_1 损失来度量合成图像与真实图像之间的差异。通过对不同尺度下的重建误差进行加权求和，网络不仅能够学习低频结构信息，还能够更好地恢复头发、手指和轮廓边缘等高频细节区域。其定义为

$$d(I, I^{GT}) = \sum_{l=1}^{5} 2^{l-1} \left\| L^l(I) - L^l(I^{GT}) \right\|_1 , \tag{3.29}$$

其中，$L^l(\cdot)$ 表示拉普拉斯金字塔的第 l 层。

(3) 推理加速方法

本章采用的预测网络在原始实现条件下，难以直接支撑沉浸式通信场景中双目 1080p 视图的高帧率生成。为此，从输入分辨率、数值精度以及运行框架三个方面对推理过程进行加速优化。推理阶段采用的加速方法包括：

1. 将输入参考视图下采样至 1/4 分辨率，在低分辨率空间中完成光流和掩模估计，

再将结果上采样至目标分辨率；

2. 将网络参数由 32 位浮点格式量化为 16 位浮点格式，以减少显存占用并提升计算吞吐效率；
3. 使用 C++ 与 TensorRT 对网络进行部署实现，替代原始的 Python/PyTorch 推理流程。

采用基于降采样输入的低分辨率推理策略。具体而言，在推理阶段将输入参考图像 I_L 和 I_R 下采样至原始分辨率的 1/4，并在该分辨率下完成光流场与融合掩模估计。网络推理输入为

$$I_L^{\downarrow} = \text{Interpolate}(I_L; 1/4),\ I_R^{\downarrow} = \text{Interpolate}(I_R; 1/4). \tag{3.30}$$

学生网络基于低分辨率输入得到降采样光流 $F_{V\to L}^{\downarrow}$、$F_{V\to R}^{\downarrow}$ 和融合掩模 $M^{\downarrow}$。随后，再将其上采样回目标分辨率，得到

$$\begin{gathered} F_{V\to L} = \text{Interpolate}(F_{V\to L}^{\downarrow}; 4),\ F_{V\to R} = \text{Interpolate}(F_{V\to R}^{\downarrow}; 4), \\ M = \text{Interpolate}(M^{\downarrow}; 4), \end{gathered} \tag{3.31}$$

通过这一方式，网络的大部分计算都在低分辨率空间中完成，从而显著降低卷积运算开销。

采用半精度量化策略，将网络参数和部分中间计算从 32 位浮点格式转换为 16 位浮点格式。采用半精度量化后，模型能够在保持较高数值表示能力的同时，减少显存占用并提升 GPU 上的并行吞吐效率，从而降低推理延迟。

采用面向部署的高性能推理实现方案。与原始 Python/PyTorch 推理方式不同，本节使用 C++ 对网络进行工程化实现，并基于 Nvidia TensorRT 对计算图进行优化与加速。TensorRT 能够自动完成层融合、算子调度优化以及显存访问优化，从而提高卷积运算和张量变换的执行效率。结合前述低分辨率推理与 FP16 量化策略，最终形成面向实时沉浸式通信的高效部署方案。

3.5　实验设置与结果分析

本节围绕所提出视点合成方法的有效性与实时部署能力展开实验验证。首先介绍实验所采用的数据集、评价指标以及训练与测试配置，随后从实时采集场景下的主客观合成质量、不同部署方案下的推理效率与实时性，以及消融实验三个方面，对所提方法进行系统评估。

3.5.1 数据集与评价指标

为保证网络训练的稳定性并提高收敛效率，在训练设置上参考了 RIFE[45] 的实现细节。训练阶段以 Vimeo90K[65] Triplet 数据作为基础训练集，每个样本由 3 帧连续视频图像构成，原始分辨率为 448×256。训练过程中，采用随机水平翻转、随机垂直翻转、时间顺序反转以及 90° 旋转等数据增强操作，以提升模型对不同运动方向和视角变化的适应能力。

优化器采用 AdamW[66]，权重衰减系数设置为 10^{-4}。训练时从输入图像中随机裁剪 224×224 的图像块作为网络输入，批大小设置为 64。学习率在整个训练过程中采用余弦退火策略，由 10^{-4} 逐步衰减到 10^{-5}。模型在 2 块 NVIDIA GeForce RTX 4090 GPU 上训练 300 个 epoch。

本节使用的测试数据集由第五章中所述系统实时采集获得。为了全面评估方法在不同动作模式下的表现，设计了三类测试场景，每类场景的视频持续时间均为 50 s：

1. 场景 I：轻微运动场景。被试基本保持静止，仅存在口部运动、表情变化及轻微手势调整；
2. 场景 II：上肢显著运动场景。被试保持坐姿，但伴随挥手、摆臂、挠头等明显上肢动作；
3. 场景 III：全身移动场景。被试在采集范围内缓慢走动，并伴随表情变化和肢体动作。

在自建测试数据集上，分别从主观质量和客观质量两个方面对算法性能进行评价。在客观评价中，统计 8 路视图在编码后的平均峰值信噪比（PSNR），并选取视差最大的相邻参考视图对 C_3 和 C_4 作为输入，在视点插值因子 $v = 0.5$ 的条件下合成中间目标视图，并以测试相机 C_t 采集的视频帧作为真实参考，计算合成结果的 PSNR、沉浸式视频峰值信噪比（Immersive Video PSNR, IV-PSNR）[67] 以及结构相似性指标（SSIM）[68]。为提高统计结果的稳定性，客观评价采用每秒采样一帧的方式进行计算，并对各场景下的结果取平均值。

在主观评价中，邀请了 5 名测试者在接收端对显示效果进行平均主观意见分（Mean Opinion Score, MOS）打分，其中 MOS 为 5 表示观看体验接近于真实人物出现在显示设备对应空间位置。实验中，多视点视频编码采用恒定码率控制（Constant Bit Rate, CBR）方式，码率设置为 15 Mbps，以保证在传输带宽受限条件下仍具备较好的视频质量。

除实时采集的数据外，本节还引入 3 个公开数据集用于消融研究，以便考察推理加速策略对视点合成性能的影响。其中，公开数据集包括：*Head Pose Image Database*[69]，包含不同头部姿态和观察角度下的人脸图像，可用于测试算法在头部转动、面部视角变化以及局部遮挡条件下的视点合成效果；*Free Viewpoint RGB-D Video Dataset*[70]，包

含多场景下的人体 RGB-D 视频序列以及相机参数信息，适合用于评估自由视点视频合成算法在动态人体场景中的表现；*nerf_synthetic* 数据集[4]，包含多个由计算机图形学方式渲染得到的合成场景，能够提供高质量 RGB 图像以及精确的相机位姿信息。

3.5.2　视点合成质量分析

表 3.1 给出了系统实时采集场景下的主客观评价结果。从主观评价来看，场景 I 和场景 III 的 MOS 分别达到 4.2 和 4.0，说明在轻微运动和全身移动条件下，系统能够较稳定地提供自然、连续的双目立体观看体验。相比之下，场景 II 的 MOS 为 3.4，低于另外两个场景。这表明当被试存在较大幅度的挥手、摆臂等上肢运动时，视点合成在快速运动区域仍会受到一定影响，尤其在手指边缘和局部遮挡交界区域更容易出现失真。

表 3.1　实时采集场景下的视点合成质量主客观评价结果

场景	MOS↑	压缩 PSNR/dB↑	合成 PSNR/dB↑	IV-PSNR/dB↑	SSIM↑
I	4.2	38.57	28.74	33.28	0.751
II	3.4	34.32	29.05	34.06	0.741
III	4.0	31.39	31.09	36.82	0.795
平均	–	–	29.63	34.72	0.762

从客观评价结果来看，多视点压缩视频的 PSNR 随着场景运动复杂度增加而逐渐下降，由场景 I 的 38.57 dB 下降到场景 III 的 31.39 dB。这是因为被试动作幅度增大后，视频编码中的时空冗余降低，从而导致压缩失真加重。相比之下，所提出方法合成视图的 PSNR 在三个场景中保持在相对稳定的范围内，分别为 28.74 dB、29.05 dB 和 31.09 dB，说明所提方法在不同动作模式下具有较好的稳定性。

3.5.3　推理效率与实时性分析

为分析 3.4 节所提出的推理加速方法的有效性，表 3.2 给出了 5 种不同实现方案下的运行时间和合成质量。实验在配置了 Intel(R) Core(TM) i9-13900KF CPU 和 NVIDIA GeForce RTX 4090 GPU（24 GB 显存）的工作站上进行。方案 A 为原始实现，即在 1920×1080 分辨率下采用 Float32 精度并基于 PyTorch 运行；方案 B 在保持 Float32 精度不变的条件下，将输入分辨率降为 960×540；方案 C 在方案 B 的基础上进一步采用 Float16 半精度推理；方案 D 在方案 A 的基础上改用 TensorRT 进行部署加速；方案 E 则在方案 A 的基础上采用了降分辨率、Float16 半精度推理和 TensorRT 部署三种加速方法。

表 3.2　视点合成网络不同部署方案的配置与效率对比

方案	输入分辨率	数值精度	推理框架	平均 PSNR/dB	推理时间/ms
A	1920×1080	Float32	PyTorch	29.41	34.9
B	960×540	Float32	PyTorch	29.61	11.5
C	960×540	Float16	PyTorch	29.61	10.6
D	1920×1080	Float32	TensorRT	29.40	23.8
E	960×540	Float16	TensorRT	29.63	7.7

由表 3.2 可以看出，随着输入降采样、半精度量化和 TensorRT 加速的逐步引入，模型推理时间显著下降，而合成质量基本保持不变。总体上，方案 A 的平均推理时间为 34.9 ms，已经超过实时（30 FPS）双目视图生成所允许的单帧 33 ms 时间预算；而方案 E 的平均推理时间降低至 7.7 ms，相比原始实现提升约 4.5 倍。从各项加速策略的贡献来看，由方案 A 到方案 B，网络输入分辨率减半后，推理时间从 34.9 ms 大幅下降到 11.5 ms，而平均 PSNR 反而略有提升，这说明低分辨率光流估计结合后续上采样与融合策略，可以在较大程度上保留合成质量。由方案 B 到方案 C，采用 Float16 后推理时间由 11.5 ms 降至 10.6 ms，表明半精度量化在提升吞吐率方面具有积极作用。而方案 A 到方案 D 到，采用 TensorRT 部署后推理时间由 34.9 ms 下降至 23.8 ms，表明基于 TensorRT 的工程化推理框架有效提高了神经网络的执行效率。最后，方案 E 同时采用输入降采样、半精度量化和 TensorRT 加速三种加速方法，将推理时间压缩到 7.7 ms。实验结果表明，3.4 节提出的推理加速方案能够有效降低运行开销，满足沉浸式通信中双目 1080p 视图实时生成的应用需求。

3.5.4　消融实验

本节在公开测试数据集上展开消融实验，以验证关键设计在不同数据场景下的有效性。本节重点考察两类影响视点合成效果的关键因素：其一是模型量化对合成质量的影响，其二是低分辨率光流推理对合成质量的影响。

3.5.4.1　模型量化对视点合成质量的影响

本节分析 Float16 半精度量化对视点合成质量的影响。实验分别在 *nerf_synthetic*、*Head Pose Image Database* 和 *Free Viewpoint RGB-D Video Dataset* 中各选择一个代表场景，并比较采用 Float16 和 Float32 两种精度推理时的 PSNR 与 SSIM。结果如表 3.3 所示。

表 3.3 中，场景 1、2、3 分别对应 *nerf_synthetic*、*Head Pose Image Database* 和 *Free Viewpoint RGB-D Video Dataset* 中的代表样本。可以看出，采用半精度量化后，各场景

表 3.3　模型量化对视点合成质量的影响

精度	场景 1		场景 2		场景 3	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
FP16	36.89	0.88	34.08	0.91	31.19	0.83
FP32	36.89	0.88	34.08	0.92	31.18	0.83

下的 PSNR 和 SSIM 变化都非常有限，几乎没有明显差异。例如，场景 1 中两种精度的 PSNR 完全一致，均为 36.89 dB；场景 2 中两者的 PSNR 也完全一致，仅 SSIM 由 0.92 轻微变化到 0.91；场景 3 中两种精度的 PSNR 仅相差 0.01 dB。该结果表明，对于本章所研究的视点合成任务而言，Float16 量化带来的数值精度损失并不会明显破坏合成图像的几何一致性和整体外观，因此半精度推理是一种有效且代价较低的模型加速手段。

3.5.4.2　低分辨率光流估计对视点合成质量的影响

本节分析低分辨率光流推理对视点合成质量的影响。实验从 *Head Pose Image Database* 中选取 3 个不同的人像场景，并从 *nerf_synthetic* 中选取 *lego* 场景作为补充测试，共构成 4 个测试场景。所有参考图像统一缩放到 1080p 分辨率，然后分别在 1.0x、0.5x 和 0.25x 三种光流推理分辨率下进行视点合成，并统计其 PSNR 和 SSIM。结果如表 3.4 所示，部分定性结果如图 3.3 所示。

表 3.4　低分辨率光流估计对视点合成质量的影响

分辨率	场景 1		场景 2		场景 3		场景 4	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
1.0x 光流	33.18	0.89	34.08	0.91	36.03	0.94	36.89	0.88
0.5x 光流	33.36	0.89	34.13	0.91	36.20	0.95	36.50	0.87
0.25x 光流	33.36	0.89	34.12	0.91	36.15	0.94	35.84	0.86

表 3.4 中，场景 1、2、3 来自 *Head Pose Image Database*，场景 4 来自 *nerf_synthetic* 的 *lego* 场景。可以看出，在前 3 个以人像为主的场景中，采用 0.5x 甚至 0.25x 分辨率进行光流推理后，PSNR 与 SSIM 相比 1.0x 分辨率几乎没有下降，部分场景甚至略有提升。例如，场景 1 中 1.0x 光流的 PSNR 为 33.18 dB，而 0.5x 和 0.25x 光流均提高到 33.36 dB；场景 3 中 0.5x 光流取得了最高的 PSNR 和 SSIM，分别为 36.20 dB 和 0.95。该现象说明，在参考视图差异较大或者局部伪影较明显的情况下，适当降低光流估计分辨率反而有助于提升匹配稳定性，从而减轻局部重影和边界伪影。

然而，在 *lego* 这类具有更复杂纹理和细节结构的场景中，低分辨率光流推理会带来一定的质量下降。场景 4 中，1.0x 光流的 PSNR 为 36.89 dB，而 0.5x 和 0.25x 光流分



(a) 0.25x 光流　(b) 0.5x 光流　(c) 1.0x 光流　(d) 真实值

图 3.3　视点合成视觉质量定性对比

别下降到 36.50 dB 和 35.84 dB，SSIM 也由 0.88 降至 0.87 和 0.86。这表明，当场景中包含大量高频纹理、细小边缘和复杂局部结构时，较高分辨率的光流估计仍然具有优势，因为它能够提供更精细的像素对应关系，有助于更准确地恢复场景细节。

综合以上两组消融实验可以看出，3.4 节中引入的两类关键加速设计具有不同特性：Float16 半精度量化在三个公开数据集上都几乎不会引起可感知的质量退化，因此是一种稳定可靠的部署优化手段；低分辨率光流推理则在多数人像场景中能够保持质量稳定，甚至对伪影抑制有所帮助，但在纹理复杂场景中可能造成一定细节损失。因此，在实际应用中，应根据场景内容和实时性要求综合选择光流推理分辨率，以在计算效率和视点合成质量之间取得更合适的平衡。

3.6　小结

本章围绕低复杂度实时视点合成问题，提出了一种基于光流预测的快速视点合成方法。首先，针对沉浸式通信中实时双目视图生成的需求，建立了视点合成问题模型，并设计了包含去畸变与几何校正的参考视图预处理流程，以提升输入视图的一致性与后续光流估计的稳定性；随后，构建了基于多尺度光流预测的视点合成网络，通过逐层细化的反向光流估计与掩模融合，实现了目标视图的高效重建；在此基础上，结合特权蒸馏训练策略以及输入降采样、半精度量化和 TensorRT 部署等推理加速方法，兼

顾了重建质量与实时性能。实验结果表明，所提方法在不同场景下能够保持较稳定的视点合成质量，并满足双目 1080p 视图实时生成的应用需求，为后续沉浸式通信系统的实现提供了有效的算法支撑。

第四章　基于三维高斯重建的多视点视频编解码

4.1　引言

多视点视频能够从多个空间位置对同一场景进行同步采集与呈现，是自由视点视频、裸眼三维显示和沉浸式通信系统的重要基础。相较于传统二维视频，多视点视频不仅能够提供双目视差与运动视差信息，还能够支持用户在一定范围内的视点切换，从而提升远程交互的沉浸体验。

现有方法通常基于多视点视频编码标准[7-8]，通过利用相邻视点之间的跨视点相关性减少冗余信息。传统多视点视频编码虽然能够在一定程度上利用视点间冗余，但通常仍需传输较多视点，带宽压力较大。Hu 等[70] 在传统编码基础上，采用视点合成思路，仅传输部分参考视点，再在接收端通过深度估计或光流插值恢复中间视图。该方法虽然减少了传输视点数量，但仅在二维图像域建立对应关系，对场景三维空间结构利用不足，在视角跨度增大或遮挡较为复杂时，容易导致合成质量下降。

近年来，三维高斯溅射[5] 作为一种新兴的显式三维表示方法，在新视点合成与三维重建任务中展现出优异的重建质量与渲染效率。该方法通过一组具有位置、颜色、尺度、旋转和不透明度属性的三维高斯基元对场景进行建模，并通过可微渲染完成视图生成。与基于体渲染的隐式表示方法[4] 相比，三维高斯表示在渲染速度方面具有明显优势，使其具备应用于实时多视点视频解码的潜力。传统三维高斯重建方法通常依赖多视点密集输入和逐场景迭代优化，生成的三维高斯点云传输数据规模大，难以直接适配低时延、多帧连续的实时视频通信场景。

随着视点数量的增加，多视点视频通信面临着传输数据规模大、压缩复杂度高的挑战。当系统需要面向裸眼 3D 显示或自由视点显示传输较为密集的视图集合时，现有多视点视频传输方法仍面临编解码与网络传输开销大的问题。因此，如何在有限带宽和计算资源条件下，充分利用多视点视频中的空间结构信息，构建兼顾传输效率、重建质量与实时性的多视点视频编解码方法，是沉浸式自由视点视频通信中的关键问题。

针对上述问题，本章研究基于三维高斯重建的多视点视频编解码方法，基本思想是在编码端仅选择并传输少量关键视点，在解码端利用接收到的稀疏视图恢复场景几何与三维高斯表示，并进一步渲染得到完整多视点视频输出。不同于全视点传输和基于视图插值的多视点解码方法，本章从三维空间表示角度重新构建多视点视频编解码链路，联合利用稀疏视图中的纹理与几何信息，以更紧凑的表示方式完成多视点视频重建。该方法能够在保证重建质量的同时显著降低传输带宽，并为裸眼 3D 显示等应用提供可实时运行的多视点视频输入。

围绕这一目标，本章从解码端三维重建和编码端视点选择两个方面展开研究。在解码端，构建由多尺度特征提取、双目深度估计、像素级三维高斯参数回归以及高斯溅射渲染组成的前馈式多视点视频解码模型，以替代传统逐帧优化式三维重建过程，实现由稀疏输入视图到完整多视点输出的前馈式重建。在编码端，结合场景内容变化与目标深度估计，设计内容自适应视点选择策略，通过动态确定最优视点对，在控制码率开销的同时提高解码端三维重建质量。

本章后续内容首先介绍问题定义与总体框架，随后阐述基于三维高斯重建的多视点视频解码方法，包括基于三维高斯的多视点解码和内容自适应的视点选择方法，最后通过实验分析所提方法的带宽效率与重建性能。

4.2　问题描述与总体框架

本章研究的目标是在多视点视频通信场景中，构建一种兼顾带宽效率、重建质量与实时性能的编解码方法。与传统多视点视频系统直接传输全部视点不同，本章尝试利用场景在三维空间中的结构冗余，仅编码和传输少量关键视点，并在接收端借助深度估计与三维高斯重建恢复完整多视点视频输出。该问题涉及两个方面，一方面，需要在稀疏输入条件下实现准确的三维场景重建；另一方面，需要根据场景内容动态选择具有代表性的输入视点。

首先给出多视点视频编解码问题定义。设多视点视频采集系统共获得 N 路输入视点视频，记为 $\mathcal{V} = \{V_i\}_{i=1}^{N}$，其中 $V_i = \{I_i^t\}_{t=1}^{T}$ 表示第 i 个视点在时序长度为 T 的视频序列，I_i^t 表示第 t 帧图像。与各视点对应的相机参数记为 $\mathcal{P} = \{P_i\}_{i=1}^{N}$，其中 P_i 包含第 i 个相机的内参与外参信息。在本章考虑的多视点视频场景中，各相机的成像分辨率与焦距保持一致，且相机位姿在采集过程中固定不变。

传统方法通常对全部视点进行编码与传输，因此整体码率较高。为降低带宽开销，本章仅对部分关键视点进行编码。从 $\mathcal{V}$ 中选择一个视点子集作为待编码视图，记为 $\mathcal{V}_S = \{V_s\}_{s\in S}$，其中 $S \subseteq \{1, 2, \cdots, N\}$ 表示被选中视点的索引集合，且通常有 $|S| \ll N$。编码端仅对 $\mathcal{V}_S$ 进行压缩与传输；解码端在解码该稀疏视点集合后，进一步重建场景的三维高斯表示，并生成其余未传输视点的视频内容。解码完整输出视点集合记为 $\hat{\mathcal{V}} = \{\hat{V}_i\}_{i=1}^{N}$，其中对于已传输视点 $s \in S$，$\hat{V}_s$ 由传统解码结果直接获得；对于未传输视点 $j \notin S$，$\hat{V}_j$ 由解码端基于三维重建与视图合成生成。

因此，本章的多视点视频编解码问题可表述为：在给定多视点视频集合 $\mathcal{V}$ 及其对应相机参数 $\mathcal{P}$ 的条件下，设计一种视点选择、编码传输与三维重建联合优化方法，使得在仅传输少量关键视点的前提下，尽可能准确地恢复完整多视点输出 $\hat{\mathcal{V}}$，并在码率、

重建质量与实时性能之间取得良好平衡。其优化目标可表示为：

$$\min_{\mathscr{E},\mathscr{D},S} \mathscr{R}(\mathscr{V}_S) + \lambda \mathscr{L}\left(\hat{\mathscr{V}}, \mathscr{V}\right), \tag{4.1}$$

其中 $\mathscr{E}$ 和 $\mathscr{D}$ 分别表示编码与解码过程，$\mathscr{R}(\mathscr{V}_S)$ 表示所选视点对应的传输码率，$\mathscr{L}(\hat{\mathscr{V}}, \mathscr{V})$ 表示重建多视点视频与原始多视点视频之间的失真度量，λ 为平衡码率与失真的权重参数。

基于上述问题定义，本章构建了一个面向多视点视频通信、基于三维高斯重建的编解码框架。该框架包括编码端与解码端两部分，其总体流程如图 4.1 所示。

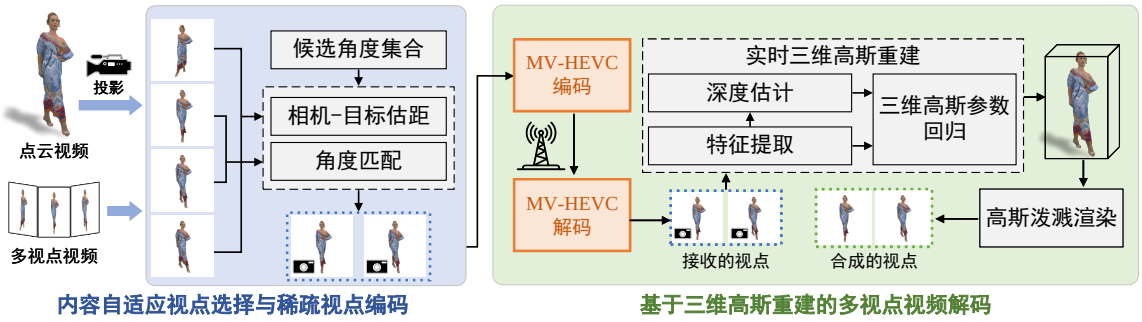


图 4.1　基于三维高斯重建的多视点视频编解码总体框架

编码端接收原始多视点视频序列作为输入，其核心任务是在控制码率的条件下生成适用于后续重建的稀疏视点表示。具体而言，编码端首先根据相机阵列分布对视点进行组织，并结合前景目标的空间位置、深度分布和视角覆盖关系进行关键视点选择。该过程会随着视频内容变化在时间维度上动态调整，以保证所选视点始终能够较好地描述当前场景的几何与纹理特征。

在完成视点选择后，编码端仅保留关键视点对应的视频序列，使用多视点视频编码器对其进行压缩，并通过网络将码流发送至接收端。与直接传输全部视点相比，这一策略显著减少了需要编码和发送的数据量。与此同时，由于所选视点通常以成对或成组形式分布在相机阵列中，它们不仅能够为解码端提供较好的纹理信息，还能为后续的双目深度估计提供必要的视差信号，从而为三维高斯重建创造条件。

解码端的任务是在仅获得稀疏参考视图的情况下恢复完整多视点视频输出。首先，解码端从网络中获得压缩码流，并通过传统视频解码器恢复出选定的稀疏参考视点图像序列。由于这些视点数量有限，解码端无法直接获得目标显示系统所需的完整视点集合，因此需要进一步执行三维重建与多视点渲染。具体而言，解码端首先从参考视图中提取多尺度图像特征，并对具有双目关系的参考视图执行视差估计与深度恢复，以建立二维图像平面与三维空间之间的几何联系。在此基础上，系统进一步结合图像特征与深度信息，对每个像素对应的三维高斯基元参数进行预测，包括空间位置、颜色、

旋转、尺度和不透明度等属性。由这些参数构成的高斯集合能够显式描述场景在三维空间中的几何与外观结构。

在得到三维高斯表示后，解码端针对目标输出视点执行高斯溅射渲染。与传统视点合成方式相比，这种方式直接利用场景的三维表示进行图像生成，因而在遮挡处理、视角跨度变化和纹理连续性方面具有更好的表达能力。最终，系统将已解码参考视图与新渲染视图共同组织为完整的多视点输出序列，用于后续显示设备播放或自由视点交互。

4.3　基于三维高斯重建的多视点视频解码

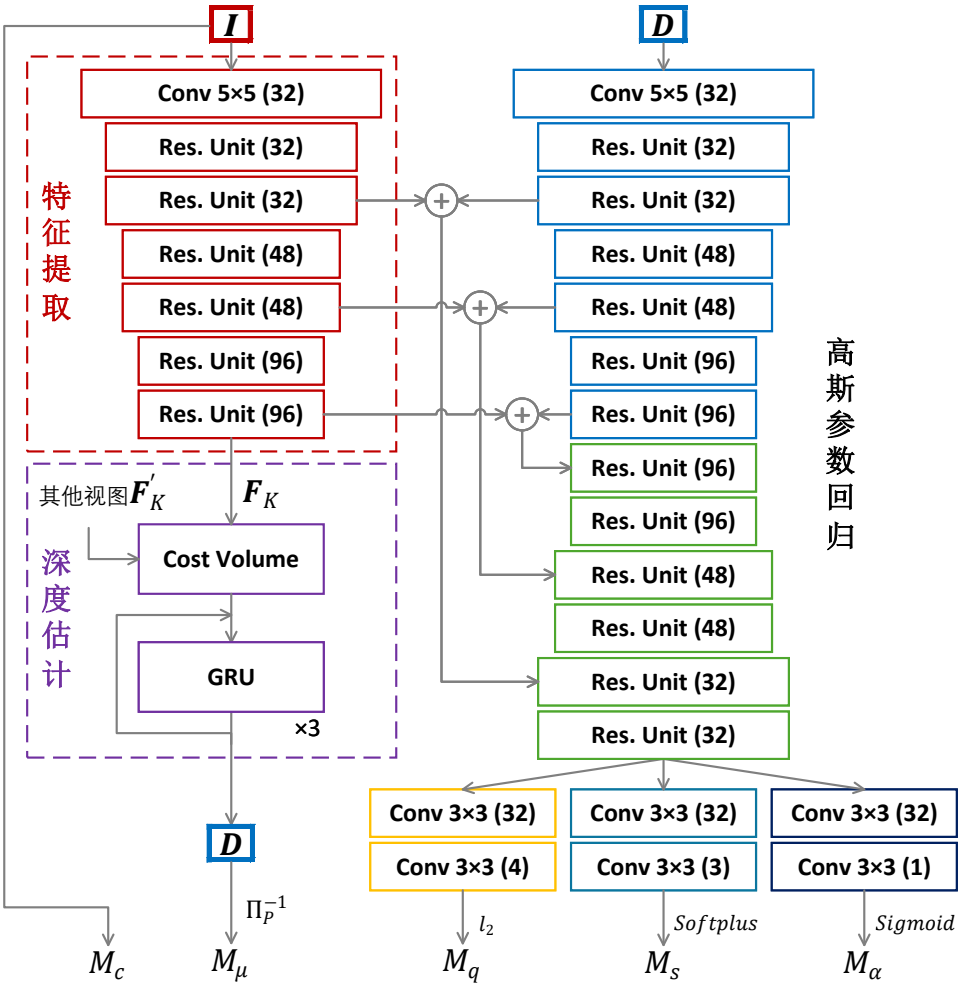


图 4.2　特征提取、深度估计与高斯参数回归网络结构图

在 4.2 节给出的总体框架中，编码端仅传输少量关键视点视频，完整多视点输出的恢复主要依赖解码端的三维重建能力。因此，如何在稀疏输入条件下高效恢复场景几

何结构，并进一步生成目标视点图像，是本章方法的核心问题。为此，本节构建了一种基于三维高斯重建的前馈式多视点视频解码模型，其基本思想是：首先从接收到的参考视图中提取多尺度图像特征，并通过双目深度估计建立图像平面与三维空间之间的几何联系；然后结合深度信息与纹理特征，对像素对应的三维高斯基元参数进行回归，构建场景的显式三维高斯表示；最后通过高斯溅射渲染恢复未传输视点，从而生成完整多视点视频输出。特征提取、深度估计与高斯参数回归模型的整体网络结构如图 4.2 所示。

与传统逐帧优化式三维重建方法不同，本节采用前馈式网络直接预测场景的三维高斯表示，避免了复杂的迭代优化过程，因而更适合实时视频通信场景。记解码端接收到的关键参考视图集合为 $\{I_s\}_{s\in S}$。其中，S 表示全部已选关键视点的索引集合，可将其分为 M 个双目视点组

$$S = G_0 \cup G_1 \cup \cdots \cup G_M, \tag{4.2}$$

其中 $G_m = \{l_m, r_m\}$ 表示第 m 组双目参考视图对应的索引集合，即每一组均由一对左右参考图像构成。由于各组双目参考视图在解码过程中采用相同的处理流程，下面不失一般性地以其中一组 $G_m = \{l, r\}$ 为例进行分析。针对任意一组双目参考视图，所提解码方法主要包括多尺度特征提取与双目深度估计、像素级三维高斯参数回归、基于高斯溅射的多视点重建与渲染三个部分。

4.3.1 多尺度特征提取与双目深度估计

双目深度估计的目标是从参考视图中恢复场景的几何结构，为后续三维高斯参数预测提供空间约束。考虑到稀疏输入场景下几何信息恢复的稳定性与实时性需求，本节采用基于立体匹配的深度估计方案，通过共享权重的多尺度特征提取网络和迭代式视差更新模块，对双目参考视图进行视差估计，并进一步将视差转换为深度。多尺度特征提取与双目深度估计的网络结构如图 4.2 左侧部分所示。

设输入左右参考图像分别为 I_l 和 I_r，首先利用共享参数的特征提取网络 $\mathscr{E}_{\text{img}}$ 提取其多尺度特征表示，即

$$\{F_{s,k}\}_{k=1}^{K} = \mathscr{E}_{\text{img}}(I_s), \quad s \in \{l, r\}, \tag{4.3}$$

其中，$F_{s,k} \in \mathbb{R}^{\frac{H}{2^k}\times\frac{W}{2^k}\times C_k}$ 表示第 k 个阶段输出的特征图，K 为特征提取网络的阶段数，H 和 W 分别表示输入图像的高和宽，C_k 表示第 k 个阶段的特征维度。

特征提取网络采用共享权重的编码器结构，由多个残差块和下采样层串联构成。对于输入图像 $I_s \in [-1, 1]^{H\times W\times 3}$，网络首先通过一个带归一化操作的残差块以及一次下采样，得到第一阶段高分辨率特征图 $F_{s,1}$；随后，将该特征继续输入后续残差块和下采

样层，逐级生成其余阶段的低分辨率特征图 $\{F_{s,k}\}_{k=2}^{K}$。因此，整个特征编码过程可表示为从输入图像到多阶段特征图的逐层映射。

在实现上，各阶段特征图的空间分辨率随阶段加深依次降为输入图像的 $1/2^k$，前几层输出较高分辨率特征，用于保留图像局部结构信息；后几层输出较低分辨率特征，用于扩大感受野并聚合上下文信息。后续深度估计模块中，采用最后一阶段特征 $F_{s,K}$ 构建双目相关体，以支持视差匹配与迭代更新。

在获得左右视图的多尺度特征后，利用末层特征构建双目相关体，以度量两个参考视图在不同视差假设下的匹配相似性。由于输入双目图像已完成极线校正，对应像素的位移主要沿水平方向变化，因此相关性计算仅需在具有相同纵坐标的像素之间进行，而无需像通用光流方法那样对所有像素对构建四维相关体。基于这一约束，采用三维相关体对双目匹配关系进行建模，从而在降低计算开销的同时保留视差估计所需的匹配信息。

具体而言，对于左视图特征图中位置 (i, j) 的特征向量，以及右视图特征图中同一行位置 (i, k) 的特征向量，通过通道维上的点积运算计算二者的匹配相似性，从而构建三维相关体：

$$\mathbf{C}_{ijk} = \sum_{h} \left(F_{l,K}\right)_{ijh} \cdot \left(F_{r,K}\right)_{ikh}, \tag{4.4}$$

其中，$\mathbf{C} \in \mathbb{R}^{\frac{H}{2^K} \times \frac{W}{2^K} \times \frac{W}{2^K}}$ 为三维相关体，前两个维度对应左视图特征图中的空间位置，第三个维度对应右视图同一行上的候选匹配位置。

由此得到的三维相关体刻画了左、右视图在不同视差假设下的特征匹配程度，其本质上是在同一极线上对所有候选匹配位置进行相似性聚合。相较于对全图所有像素对构建四维相关体的方式[71]，该表示更符合双目匹配问题的几何约束。

在构建三维相关体之后，采用基于多层卷积门控循环单元（GRU）的迭代更新机制对视差场进行逐步优化。该过程以零初始化视差为起点，通过反复查找相关体并更新隐藏状态，逐步预测左右视图的视差序列。设初始视差为

$$d_l^0 = 0, \quad d_r^0 = 0, \tag{4.5}$$

其中 d_l^0 和 d_r^0 分别表示左、右视图的初始视差估计。随后，在第 t 次迭代中，系统利用当前视差估计对三维相关体进行索引查找，提取与当前匹配位置对应的相关特征；同时，将当前视差场本身编码为视差特征，并与上下文特征共同输入更新网络。

具体而言，对于当前迭代的视差估计 d_s^t，首先根据其对应的匹配位置在三维相关体 $\mathbf{C}$ 中进行采样，得到相关特征：

$$f_{\text{corr},s}^t = \text{Lookup}(\mathbf{C}, d_s^t), \quad s \in \{l, r\}, \tag{4.6}$$

其中 Lookup(·) 表示基于当前视差估计的相关体索引操作。随后，将提取到的相关特征输入两层卷积网络进行编码；同时，当前视差估计 d_s^t 也经过两层卷积映射为视差特征。再结合由特征提取网络生成的上下文特征，共同构成更新网络的输入。于是，第 t 次迭代的更新过程可表示为

$$h_s^{t+1} = \mathrm{GRU}\Big(h_s^t, f_{\mathrm{corr},s}^t, f_{\mathrm{disp},s}^t, f_{\mathrm{ctx},s}^t\Big), \tag{4.7}$$

$$\Delta d_s^t = \Psi(h_s^{t+1}), \tag{4.8}$$

$$d_s^{t+1} = d_s^t + \Delta d_s^t, \tag{4.9}$$

其中，h_s^t 表示第 t 次迭代时 GRU 的隐藏状态，$f_{\mathrm{corr},s}^t$、$f_{\mathrm{disp},s}^t$ 和 $f_{\mathrm{ctx},s}^t$ 分别表示相关特征、视差特征和上下文特征，$\Psi(\cdot)$ 表示由隐藏状态预测视差增量的映射函数，Δd_s^t 表示当前迭代的视差更新量。

由此，经过 T 次迭代后，可得到左右视图在低分辨率特征空间上的最终视差估计 $\{d_l^T, d_r^T\}$。这些视差场的空间分辨率为输入图像的 $1/2^K$。为了恢复到原始图像分辨率，进一步采用与 RAFT[71] 相同的凸组合上采样方法，对低分辨率视差场进行上采样，得到最终的稠密视差图 $d_s(u,v)$。该上采样方式能够在保持边缘结构的同时提升视差图分辨率，从而为后续深度恢复提供更精细的几何信息。

在得到稠密视差图后，依据双目成像几何关系将其转换为深度图。对于经过极线校正的双目图像，像素视差与场景深度之间满足反比关系，因此像素 (u,v) 处的深度可表示为

$$D(u,v) = \frac{Bf}{d_s(u,v)}, \tag{4.10}$$

其中，$D(u,v)$ 表示像素 (u,v) 对应的深度值，$d_s(u,v)$ 表示该位置的视差估计，B 为双目相机之间的基线长度，f 为相机焦距。通过上述变换，系统将双目匹配得到的二维视差信息恢复为具有真实空间尺度意义的深度表示，进而完成从双目参考视图到场景深度图的估计。所得深度图不仅为后续像素级三维位置反投影提供了直接几何约束，也为三维高斯参数回归建立了必要的空间先验。

4.3.2 像素级三维高斯参数回归

在获得参考视图的深度信息后，需要进一步将二维图像中的像素映射到三维空间表示中。为此，本节采用像素级三维高斯参数回归方法，直接从图像特征与深度特征中预测场景的显式三维高斯表示。与传统 3DGS 依赖逐场景迭代优化不同，该方法通过前馈网络直接输出高斯参数，从而显著提高了解码速度，更适用于实时多视点视频重建任务。像素级三维高斯参数回归的网络结构如图 4.2 右侧部分所示。

设场景由若干各向异性三维高斯组成，其中一个高斯基元记为

$$\mathbf{g}_i = \{\boldsymbol{\mu}_i, \mathbf{c}_i, \mathbf{q}_i, \mathbf{s}_i, \alpha_i\}, \tag{4.11}$$

其中，$\boldsymbol{\mu}_i \in \mathbb{R}^3$ 表示高斯中心坐标，$\mathbf{c}_i \in \mathbb{R}^3$ 表示颜色向量，$\mathbf{q}_i \in \mathbb{R}^4$ 表示旋转四元数，$\mathbf{s}_i \in \mathbb{R}^3$ 表示尺度参数，$\alpha_i \in \mathbb{R}$ 表示不透明度。所有高斯基元共同构成用于后续新视点生成的场景表示。

对于参考图像 I 中的每个像素 $p = (u, v)$，为其关联一组高斯属性：

$$\mathscr{M}(p) = \left\{ M_\mu(p), M_c(p), M_q(p), M_s(p), M_\alpha(p) \right\}, \tag{4.12}$$

其中 M_μ、M_c、M_q、M_s 和 M_α 分别对应三维位置、颜色、旋转、尺度和不透明度映射。通过这些参数映射，可以将二维像素提升为三维空间中的高斯表示。

首先，三维位置由深度值与相机模型联合确定。设参考视图对应的相机参数为 P，世界到相机坐标系的投影函数为 Π_P，则该像素对应的三维点可由反投影得到：

$$M_\mu(p) = \Pi_P^{-1}\big(p, D(p)\big), \tag{4.13}$$

其中 $D(p)$ 为深度估计模块输出的深度值。

其次，在颜色建模方面，本章考虑的通信场景以人物为主，表面反射特性以漫反射为主，视角变化带来的颜色漂移相对有限，因此不再显式回归球谐光照系数，而是直接继承参考图像的颜色值：

$$M_c(p) = I(p). \tag{4.14}$$

与位置和颜色不同，旋转、尺度以及不透明度无法由相机几何直接给出，需要依赖数据驱动方式进行预测。为充分利用图像纹理和场景几何两方面信息，采用双分支编码、融合解码的网络结构。具体而言，一条分支对参考图像进行外观特征提取，另一条分支对深度图进行几何特征提取。设两者对应的编码器分别为 $\mathscr{E}_{\text{rgb}}$ 和 $\mathscr{E}_{\text{depth}}$，首先得到多尺度外观特征与几何特征，随后在各个尺度上将两类特征进行拼接，并输入解码器 $\mathscr{D}_{\text{param}}$，恢复得到全分辨率的高斯特征张量：

$$O = \mathscr{D}_{\text{param}}\big(\mathscr{E}_{\text{rgb}}(I) \oplus \mathscr{E}_{\text{depth}}(D)\big), \tag{4.15}$$

其中，$O \in \mathbb{R}^{H \times W \times C_Z}$，$C_Z$ 表示高斯特征通道数，$\oplus$ 表示在各个特征尺度上的拼接操作。通过这种方式，网络能够将图像外观信息与深度几何信息联合编码到统一的像素级高斯特征表示中。

在此基础上，通过三个独立的参数预测头对旋转、尺度和不透明度进行回归。设

h_q、h_s 和 h_o 分别表示旋转头、尺度头和不透明度头，则有

$$M_q(p) = \frac{h_q(O(p))}{\left\| h_q(O(p)) \right\|_2}, \tag{4.16}$$

$$M_s(p) = \mathrm{Softplus}\left(h_s(O(p))\right), \tag{4.17}$$

$$M_\alpha(p) = \mathrm{Sigmoid}\left(h_\alpha(O(p))\right), \tag{4.18}$$

其中，旋转项通过 ℓ_2 归一化约束为单位四元数；尺度项经 Softplus 激活后保证为正；不透明度项经 Sigmoid 映射后限制在 $(0, 1)$ 区间内。三个预测头均采用轻量级卷积结构实现，以控制在线推理开销。

对于左右两个参考视图，系统分别构建其对应的像素级高斯集合 $\mathcal{M}_l$ 和 $\mathcal{M}_r$，并合并为统一场景的三维高斯集合：

$$\mathcal{G} = \mathcal{M}_l \cup \mathcal{M}_r. \tag{4.19}$$

通过上述设计，解码端能够在不进行逐帧迭代优化的情况下，直接由参考视图回归场景的显式三维高斯表示，从而进行后续多视点渲染。

4.3.3 基于高斯溅射的多视点重建与渲染

在得到场景的三维高斯集合后，解码端需要根据目标相机参数生成各输出视点图像。为此，本节采用高斯溅射[5] 渲染方法，将三维高斯表示投影到目标图像平面，并通过可微的透明度混合规则合成最终图像。

对于第 i 个高斯基元，其空间分布可表示为

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{4.20}$$

其中，$\boldsymbol{\mu}$ 为高斯中心位置，$\boldsymbol{\Sigma}$ 为协方差矩阵。为便于参数化和优化，将协方差矩阵分解为旋转与尺度的形式：

$$\boldsymbol{\Sigma} = \mathbf{R}\mathbf{S}\mathbf{S}^\top\mathbf{R}^\top, \tag{4.21}$$

其中，$\mathbf{R}$ 由四元数 $\mathbf{r}$ 转换得到，$\mathbf{S} = \mathrm{diag}(\mathbf{s})$ 为由尺度向量构成的对角矩阵。该分解形式能够在保证协方差矩阵半正定的同时，提高参数预测的稳定性。

给定目标视点对应的相机位姿与投影矩阵，三维高斯需要从世界坐标系变换到目标相机坐标系，并进一步投影到二维图像平面。设视图变换矩阵为 $\mathbf{W}$，投影变换的局部雅可比矩阵为 $\mathbf{J}$，则高斯在目标图像平面上的二维协方差可表示为

$$\boldsymbol{\Sigma}' = \mathbf{J}\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^\top\mathbf{J}^\top. \tag{4.22}$$

该二维高斯描述了高斯基元投影到图像平面后的空间扩展范围，用于后续像素颜色计算。

在渲染过程中，系统首先对所有可见高斯基元按照深度进行排序，然后采用基于透明度的前向混合规则计算每个像素的颜色值。设某一像素被若干个高斯基元覆盖，则其最终颜色可表示为

$$\mathbf{C}_{\text{pixel}} = \sum_{i} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{4.23}$$

其中，$\mathbf{c}_i$ 表示第 i 个高斯基元在该像素处贡献的颜色，α_i 表示其对应的不透明度。该混合方式能够有效处理遮挡关系和前后景叠加问题，从而生成具有较高视觉一致性的目标视图。

对于所有未传输视点，系统均可按照其目标相机参数执行上述渲染过程，得到对应的重建视图。结合已解码得到的参考视图，最终即可组织形成完整的多视点视频输出序列。

4.3.4　损失函数与训练策略

为保证多尺度特征提取、深度估计和三维高斯参数回归模块能够稳定协同工作，本节采用两阶段训练策略对整个解码模型进行优化。第一阶段侧重于建立可靠的双目匹配与深度恢复能力，第二阶段则在此基础上联合优化深度估计与高斯参数回归，以提升最终渲染质量。

在第一阶段，使用具有真值视差标注的合成双目数据对特征提取网络和深度估计网络进行预训练。设网络在第 t 次迭代输出的视差预测为 d^t，真值视差为 d_{gt}，则深度估计损失定义为

$$\mathscr{L}_{\text{depth}} = \sum_{t=1}^{T} \gamma^{T-t} \left\| d^t - d_{\text{gt}} \right\|_1, \tag{4.24}$$

其中，γ 为迭代权重衰减系数。该损失形式对后期迭代结果赋予更高权重，有助于提升最终视差估计精度。

在完成深度模块预训练后，第二阶段引入多视点人物数据，对特征提取、深度估计和三维高斯参数回归模块进行联合微调。此阶段的优化目标不再局限于视差误差，而是直接约束重建视图与真实目标视图之间的图像差异。设渲染得到的目标图像为 $\hat{\boldsymbol{I}}$，其对应真值图像为 $\boldsymbol{I}$，则图像重建损失定义为

$$\mathscr{L}_{\text{img}} = \lambda_1 \mathscr{L}_1 + \lambda_2 \mathscr{L}_{\text{SSIM}}, \tag{4.25}$$

其中

$$\mathscr{L}_1 = \left\| \hat{I} - I \right\|_1, \tag{4.26}$$

$$\mathscr{L}_{\text{SSIM}} = 1 - \text{SSIM}(\hat{I}, I). \tag{4.27}$$

其中，$\mathscr{L}_1$ 用于约束像素级颜色误差，$\mathscr{L}_{\text{SSIM}}$ 用于增强图像结构一致性与感知质量[68]。

最终，将深度估计损失与图像重建损失联合构成总损失函数：

$$\mathscr{L} = \mathscr{L}_{\text{depth}} + \mathscr{L}_{\text{img}}. \tag{4.28}$$

通过这一训练策略，模型首先学习稳健的双目匹配与深度恢复能力，再逐步建立从稀疏参考视图到三维高斯表示及目标视图渲染的端到端映射关系。这样既能够保证几何估计的准确性，又有助于提升最终多视点重建结果的视觉质量。

4.4 内容自适应视点选择与稀疏视点编码

上一节从解码端角度介绍了基于三维高斯重建的多视点视频解码方法。然而，解码端的重建性能不仅取决于深度估计和高斯参数回归网络本身，还与编码端所选择的输入参考视图密切相关。对于基于双目深度估计和三维高斯重建的解码框架而言，若所选视点之间的空间分布不合理，则可能导致几何约束不足、纹理覆盖不完整或视差估计误差增大，进而降低最终多视点重建质量。因此，在保证码率受控的前提下，如何根据场景内容自适应地选择关键视点，是实现高效多视点视频编解码的重要环节。

本节设计了一种内容自适应视点选择与稀疏视点编码策略。该策略利用 4.3 节得到的深度估计能力，对场景内容和目标距离进行分析，并以视点夹角作为关键几何参数，在时间维度上动态调整最优视点对。

首先给出视点选择问题分析与经验观察。在基于三维高斯重建的多视点视频编解码框架中，解码端仅依赖少量参考视图恢复完整多视点输出，因此关键视点的选择会直接影响后续深度估计、三维高斯重建以及目标视图渲染的质量。

设相机阵列沿水平方向的圆弧分布，且每两个关于阵列中心轴对称的相机构成一组候选视点对。若所选视点对间距过小，则虽然图像内容相似度较高、双目匹配较为稳定，但参考视图对目标的观察存在冗余，难以充分刻画场景的整体三维结构；若所选视点对间距过大，则能够提升对目标的空间覆盖范围，但由于视差显著增大，易在人体边界、遮挡区域及复杂姿态位置引入更大的深度估计误差。因此，最优视点选择通常不会单调地偏向于较小或较大的相机间隔，而是与目标距离、场景内容以及相机阵列几何分布共同相关。

为分析视点选择与重建质量之间的关系，本节在公开多视点数据（详情参见 4.5

节）上开展了经验测量实验，结果如图 4.3 所示。实验中保持相机阵列位置固定，分别设置目标与相机阵列之间的距离为 1.5 m、2 m 和 3 m；在每种距离条件下，系统共采集 20 个视点，并组织为 10 组关于阵列中心轴对称的候选视点对。图中横轴表示候选视点对编号，编号越大表示该视点对中两台相机之间的间距越大；左纵轴和右纵轴分别表示在该视点对条件下解码得到的全部视图平均 PSNR 和平均 SSIM。

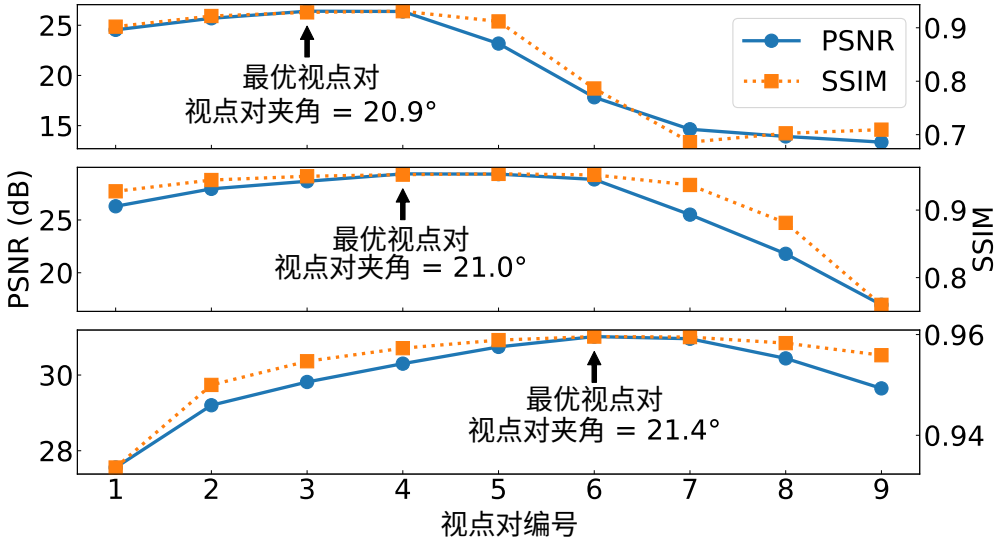


图 4.3　不同目标距离条件下候选视点对与重建质量之间的关系

从图 4.3 可以看出，不同候选视点对对应的重建质量曲线均呈现出先升后降的趋势，而非随视点间隔单调变化。这表明在给定场景内容和目标距离的条件下，确实存在能够同时兼顾空间覆盖能力和双目匹配稳定性的最优参考视图组合。进一步地，最优视点对会随目标距离变化而变化：当目标距离为 1.5 m 时，最优视点对为第 3 组；当目标距离为 2 m 时，最优视点对为第 4 组；当目标距离增大到 3 m 时，最优视点对则移动到第 6 组附近。这说明最优视点对并非固定不变，而具有显著的距离依赖性和内容依赖性。

进一步分析可知，尽管最优视点对编号会随目标距离变化，但图中标注的最优观察夹角却保持在较为接近的范围内。具体而言，三种距离条件下最优视点对对应的夹角分别为 20.9°、21.0° 和 21.4°。这一结果说明，相比于直接使用视点对编号或固定基线长度描述参考视图关系，以目标为顶点、连接左右相机光心所形成的观察夹角，能够更稳定地反映视点选择对解码质量的影响。基于这一观察，在后续内容自适应视点选择策略中采用观察夹角作为核心变量，通过动态更新最优角度并在候选视点对中进行匹配，实现场景内容变化条件下的自适应关键视点选择。

因此，将视点选择问题进一步转化为：在给定当前场景深度分布和相机阵列几何结构的条件下，寻找与最优角度参数最匹配的对称视点对，使其在受限码率条件下尽

可能提高解码端三维高斯重建与多视点渲染质量。

(1) 视点分组与前景深度估计

为了在多视点数量较多的情况下控制搜索复杂度并提升编码效率，首先对候选视点进行分组处理。考虑到实际系统中的相机阵列可能具有不同的视角范围和视点密度，若直接在全部视点中进行全局搜索，不仅计算代价较高，也不利于与分组编码和并行处理机制结合。为此，按照视点在阵列中的角度分布，将全部视点划分为若干组，每组覆盖固定角度范围 θ。设完整视点集合为

$$\mathscr{V} = \{V_i\}_{i=1}^{N}, \tag{4.29}$$

则可将其划分为

$$\mathscr{G} = \{\mathscr{G}_1, \mathscr{G}_2, \cdots, \mathscr{G}_M\}, \tag{4.30}$$

其中 $\mathscr{G}_m$ 表示第 m 个视点组，M 为分组总数。每组内部仅选择一对关于组中心对称分布的参考视点用于编码，从而降低需要传输的视点数量。

视点分组后，系统还需要根据场景内容确定与当前目标最匹配的视点间距。由于目标距离是影响最优视点对的重要因素，本节引入前景深度估计作为视点选择的几何依据。具体而言，在每个时间片或视频块开始时，系统首先利用中心参考视点对执行双目深度估计（4.3 节），得到当前帧的深度图 $\boldsymbol{D}$。随后，结合前景分割结果提取人物或目标区域，并计算其平均深度作为场景距离估计值。

设前景区域掩码为 $\boldsymbol{M}_f$，其中

$$\boldsymbol{M}_f(u,v) = \begin{cases} 1, & (u,v) \in \text{foreground}, \\ 0, & \text{otherwise}, \end{cases} \tag{4.31}$$

则前景平均深度可表示为

$$\bar{D}_f = \frac{1}{|\boldsymbol{M}_f|} \sum_{(u,v)\in \boldsymbol{M}_f} \boldsymbol{D}(u,v), \tag{4.32}$$

其中 $|\boldsymbol{M}_f|$ 表示前景像素数。该量刻画了当前目标与相机阵列之间的大致距离关系，可为后续角度参数更新提供重要依据。

(2) 最优角度更新机制

在获得视点分组结果和当前场景的前景深度后，下一步需要确定与当前内容相匹配的最优视点夹角。考虑到视频序列在时间上具有连续性，相邻时刻的目标距离、姿态和外观通常不会发生剧烈变化，因此最优角度参数也具有较强的时域相关性。基于这一特点，本节设计了一种面向视频块的最优角度更新机制，在保证自适应性的同时

降低搜索复杂度。

设视频序列被划分为若干编码块或 GOP（Group of Pictures），记为 $\{c_0, c_1, \cdots, c_T\}$。其中 c_t 表示第 t 个视频块。对于每个视频块，系统固定使用一个最优角度参数 o_t 来指导关键视点选择。若直接在全部候选视点对中采用暴力搜索寻找最优解，虽然能够获得较高精度，但需要为每个视频块遍历大量候选组合并评估其重建质量，计算代价过高，难以满足实时编码需求。

为此，本节利用相邻视频块之间的连续性，在上一时刻最优角度 o_{t-1} 附近构造一个局部候选角度集合：

$$\mathcal{O}_t = \left\{ o_{t-1} - k\Delta, \cdots, o_{t-1} - \Delta, o_{t-1}, o_{t-1} + \Delta, \cdots, o_{t-1} + k\Delta \right\}, \tag{4.33}$$

其中，Δ 为角度步长，k 为搜索半径。通过这种局部细化方式，系统只需要在有限个候选角度中进行比较，即可在较低复杂度下实现对最优角度的跟踪更新。

对于候选集合中的任一角度 $o_t^{(n)} \in \mathcal{O}_t$，系统可以基于当前深度估计结果和候选视点组合，对其可能得到的重建质量进行评估。由于本章方法的目标是提高解码端多视点恢复质量，因此将候选角度的评价准则定义为对应重建结果的视觉质量指标。设评价函数为 $Q(\cdot)$，则最优角度更新过程可写为

$$o_t = \arg \max_{o_t^{(n)} \in \mathcal{O}_t} Q\left(o_t^{(n)}\right). \tag{4.34}$$

在实际实现中，$Q(\cdot)$ 由 PSNR 或 SSIM 等指标进行标定。通过这种方式，系统能够根据视频内容变化动态调整最优角度参数，并在不同目标距离下保持较稳定的视点选择性能。与固定选择同一组参考视点相比，最优角度更新机制能够更好地适应场景动态变化。当目标靠近相机阵列时，系统倾向于选择夹角较小的参考视点对，以保证视差估计稳定；当目标距离增大时，则可以适当增大参考视点夹角，以提升空间覆盖能力和重建完整性。由此实现了视点选择策略从静态规则向内容自适应更新的转变。

(3) 角度匹配与关键视点确定

在得到当前视频块的最优角度参数后，编码端需要从候选视点集合中选出与该角度最匹配的一对关键视点作为最终编码输入。为保证左右视图在几何上具有较好的平衡性并便于双目深度估计，本节限定关键视点从各组中关于组中心对称分布的候选视点对中选取。

设目标中心的三维位置为 $\mathbf{x}_o$，第 i 个相机的光心位置为 $\mathbf{c}_i$，则对于任意一个候选对称视点对 (i, j)，可以定义其相对于目标的观察夹角为

$$\theta_{ij} = \arccos\left(\frac{(\mathbf{c}_i - \mathbf{x}_o)^\top (\mathbf{c}_j - \mathbf{x}_o)}{\|\mathbf{c}_i - \mathbf{x}_o\|_2 \cdot \|\mathbf{c}_j - \mathbf{x}_o\|_2} \right). \tag{4.35}$$

其中，$\mathbf{x}_o$ 可由前景区域的深度估计结果和相机参数近似恢复得到。该角度反映了左右参考视点围绕目标形成的观察张角，是衡量视点对空间覆盖能力与匹配稳定性的重要几何量。

给定最优角度参数 o_t 后，关键视点对的选择可表示为

$$(i^*, j^*) = \arg \min_{(i,j)\in\mathcal{P}_m} \left|\theta_{ij} - o_t\right|, \tag{4.36}$$

其中，$\mathcal{P}_m$ 表示第 m 个视点组中的候选对称视点对集合。即在当前组内选择观察夹角最接近最优角度参数的那一对视点，作为该组最终用于编码的关键视图。

这种角度匹配机制具有两个优点。首先，它避免了直接基于视点编号或固定基线长度进行选择所带来的不稳定性，因为相机编号差异并不能准确反映当前目标与相机的真实几何关系。其次，它将经验分析中观察到的最优视点对对应相近观察夹角的规律显式融入了视点选择过程，从而使关键视点的确定方式具有明确的几何解释。

通过上述匹配过程，系统能够在每个视点组中自动选出与当前场景内容最匹配的一对关键视点，并将其作为稀疏参考视图输入编码器。不同视频块之间，随着最优角度参数的更新，关键视点对也会相应变化，从而实现对内容变化的动态适应。

(4) 稀疏视点编码流程

在完成关键视点选择后，编码端仅对所选稀疏视点进行压缩和传输。结合前述内容，自适应稀疏视点编码流程可以概括为如下步骤。

首先，系统接收原始多视点视频序列作为输入，并根据相机阵列结构将全部视点划分为若干视点组。对于每个视频块，系统利用中心参考视点对估计前景深度，得到当前目标距离及其几何分布信息。随后，在上一视频块最优角度参数附近构造候选角度集合，并依据快速质量评价准则更新当前视频块的最优角度参数。

其次，对于每个视点组，系统计算组内各对称候选视点对相对于目标的观察夹角，并根据最优角度参数执行角度匹配，确定最终的关键视点对。由此，可得到当前视频块对应的稀疏参考视图集合：

$$\mathcal{V}_S^t = \{V_s^t\}_{s\in S_t}, \tag{4.37}$$

其中 S_t 表示第 t 个视频块中被选中的关键视点索引集合。

最后，系统利用多视点视频编码器对 $\mathcal{V}_S^t$ 进行压缩，并将码流发送至接收端。由于编码端仅保留少量关键视点，因此编码器可以将更多比特分配给这些视图本身，从而在给定总带宽预算下提升参考图像质量。解码端在接收到这些高质量稀疏视图后，再结合 4.3.1 节提出的双目深度估计、三维高斯参数回归与高斯溅射渲染过程，恢复其余未传输视点，从而形成完整多视点视频输出。

4.5　实验设置与结果分析

本节围绕所提出多视点视频编解码方法的带宽效率和重建质量方面展开实验验证。首先对数据集与训练配置、对比方法及评价指标进行统一说明，在此基础上从总体带宽、率失真性能和重建质量主客观表现等方面，对所提方法进行实验分析，并通过消融实验和关键参数分析，进一步验证所提出三维高斯解码框架及内容自适应视点选择策略的有效性。

4.5.1　实验设置

4.5.1.1　数据集与训练配置

本节构建了用于测试的多视点视频数据集，以验证所提方法在不同人物形态、动作模式和场景内容下的泛化性能。具体而言，实验选取 8i 数据集[72] 中的 Long Dress、Loot 和 Red and Black 三段动态点云，以及 Microsoft Upper Body 数据集[73] 中的 Andrew、David 和 Phil 三段上半身人物点云作为测试样本。上述数据覆盖了舞蹈、说话与手势等多种动态内容，具有较好的代表性。

测试数据生成时，对原始动态三维点云数据进行多视点渲染，得到面向实验的多视点视频序列。具体而言，在目标前方设置一组沿圆弧分布的虚拟相机，圆弧范围为 60°，相机到目标中心的半径为 2 m，共渲染 50 个视点，每个视点图像分辨率为 1024×1024。通过这种方式，可以获得与裸眼三维显示设备相匹配的高密度多视点视频输入。

在模型训练方面，采用两阶段训练策略。首先，在 SceneFlow 合成双目数据集[74] 上对特征提取与深度估计模块进行预训练，以增强网络对双目匹配关系和深度线索的建模能力。随后，在 THuman 2.1 数据集[75] 上对整个解码模型进行微调。THuman 2.1 数据集包含 2500 个高质量人体三维扫描样本，能够为三维高斯参数回归和新视点渲染提供较丰富的人体几何与纹理先验。

训练过程中，优化器采用 AdamW[66]，初始学习率设为 2×10^{-3}。损失函数权重设置与前文一致，其中深度迭代损失的时间衰减系数设为 $\gamma=0.9$，图像重建损失中的 L_1 项和结构相似性项权重分别设为 $\lambda_1=0.8$ 和 $\lambda_2=0.2$。在视点选择模块中，视点分组所采用的固定角度范围参数设为 $\theta=30°$。

4.5.1.2　对比方法与评价指标

为全面评估所提方法的性能，本节选取两类具有代表性的多视点视频传输方案作为对比方法。

第一类对比方法为基于多视点视频编码的方案，记为 x265Multiview。该方法对全部输入视点视频进行压缩与传输，采用 MV-HEVC[7] 作为核心编码机制，通过建模跨

视点相关性降低冗余。实验中使用 x265[76] 实现编码，接收端采用 FFmpeg 7.1[77] 完成对应解码，这些都是 MV-HEVC 的高效实现。

第二类对比方法为基于视图插值的方案，记为 ViewInterp。该方法仅传输部分参考视点（将传输视图数量减少到 8 个），在接收端利用实时中间光流估计模型[45] 对其余视点进行插值恢复。

在评价指标方面，本节从带宽效率与重建质量两个角度进行测量。对于带宽性能，采用总体码率（bitrate）衡量系统传输开销，并采用 Bjøntegaard Delta Rate（BD-Rate）评价在相同重建质量下相对于对比方法的平均码率节省。对于图像质量，采用 RGB 域峰值信噪比（PSNR）和 RGB 域平均结构相似性指标（SSIM）[68] 作为主要评价标准，以衡量完整多视点视频的重建保真度。

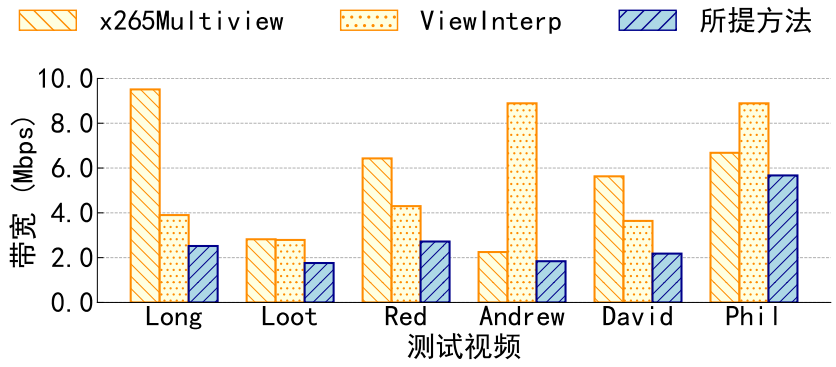
除上述离线重建指标外，第五章还统计了系统实时性能，包括编码帧率、解码帧率以及端到端时延等指标，以进一步验证所提方法在实时多视点视频通信场景中的可用性。

4.5.2 总体带宽与率失真性能分析

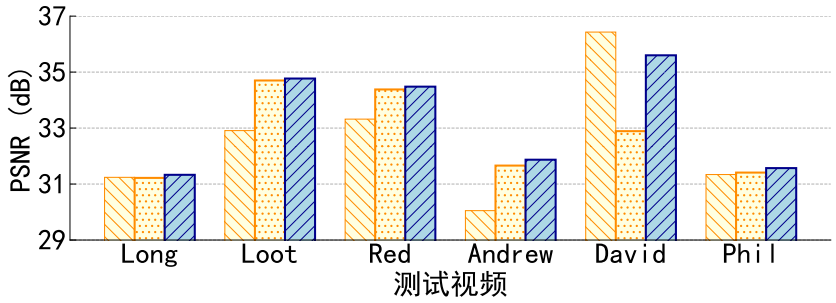
为了验证所提方法在带宽效率方面的优势，首先将所提方法与基于全视点编码的 x265Multiview 方法以及基于稀疏视点传输与视图插值的 ViewInterp 方法进行对比。总体实验结果如图 4.4 所示，其中图 4.4(a) 给出了六组测试视频在三种方法下的带宽开销，图 4.4(b) 和图 4.4(c) 分别给出了对应的 PSNR 和 SSIM 结果。

从带宽结果来看，所提方法在六组测试视频上均表现出明显的码率压缩优势。与 x265Multiview 相比，所提方法在 Long、Loot、Red、Andrew、David 和 Phil 六段视频上的带宽降幅分别为 73.48%、37.66%、57.65%、18.3%、45.9% 和 14.99%，平均降低 41.33%；与 ViewInterp 相比，相应降幅分别为 35.32%、37.09%、36.65%、79.28%、40.22% 和 36.14%，平均降低 44.12%。进一步地，所提方法在测试序列中的传输码率仅为 1.76 Mbps 至 5.67 Mbps，表明通过仅传输少量关键视点并在接收端进行三维高斯重建，可以有效压缩多视点视频中的跨视点冗余信息。

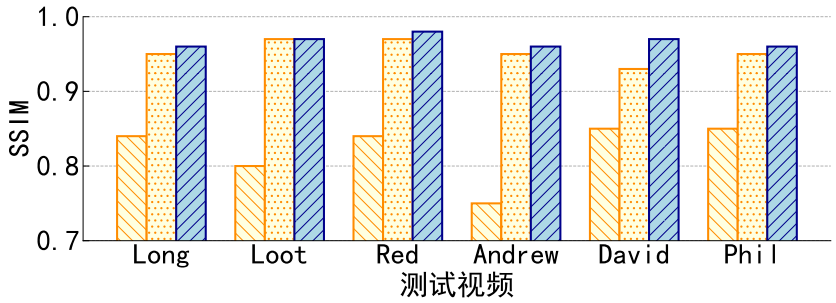
本节在三组测试视频上分析了率失真性能，结果如图 4.5 所示。图中给出了 Long、Loot 与 Red 三段视频在不同码率条件下的 PSNR-Rate 与 SSIM-Rate 曲线。总体来看，在中低码率区域，所提方法能够以更低的传输带宽获得更高的重建质量，其率失真性能整体优于两种对比方法；而在高码率区域，随着码率进一步提升，所提方法的解码质量增益逐渐趋于饱和，传统多视点编码方案则表现出更明显的率失真优势。例如，在三个测试序列中，随着码率持续增大，所提方法与 x265Multiview 分别在约 15 Mbps、5 Mbps 与 10 Mbps 附近出现性能接近区域。从适用场景来看，x265Multiview 更适用于



(a) 总体带宽



(b) PSNR



(c) SSIM

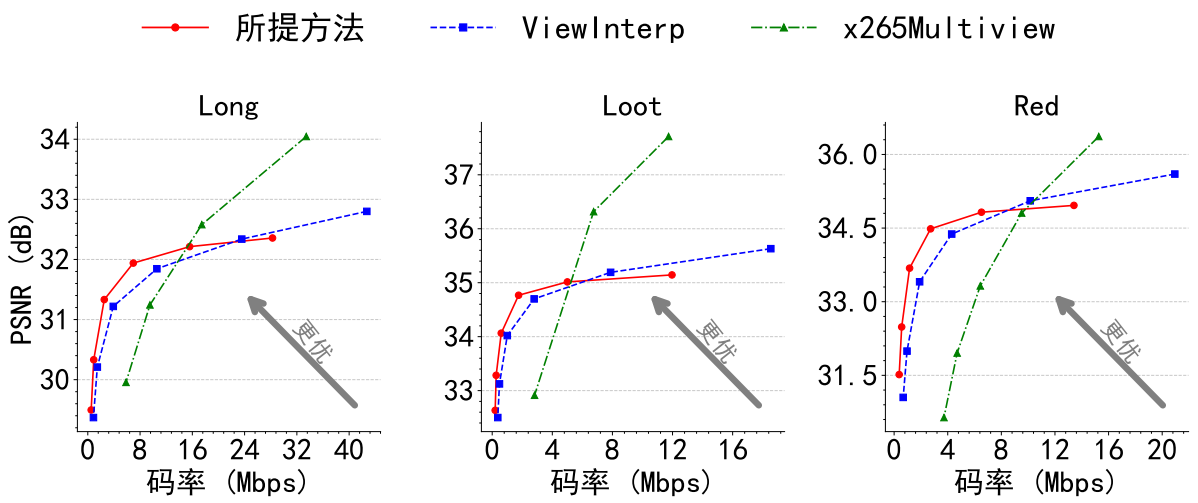
图 4.4　多视点视频编解码总体带宽与视觉质量评估

带宽资源充足、追求极致重建质量的高码率场景；相比之下，所提方法则更适用于带宽受限、强调码率利用效率与传输实时性的应用场景。

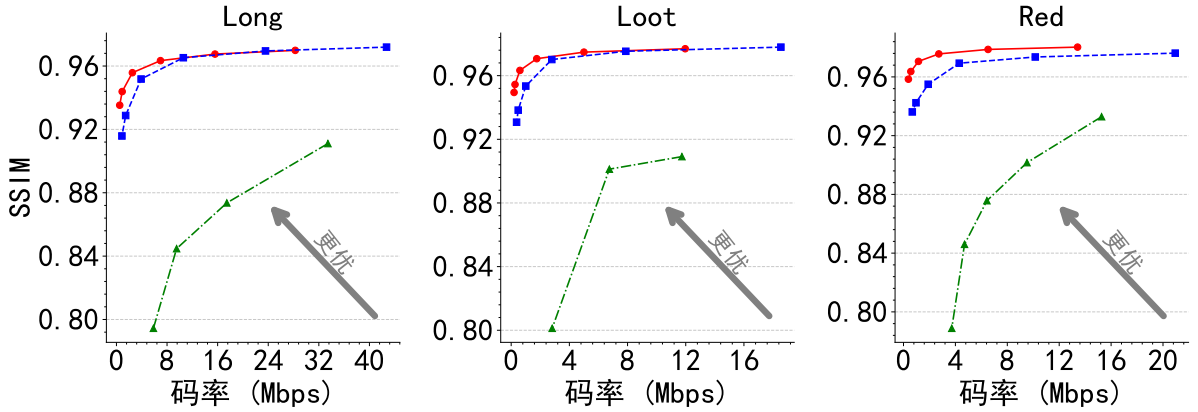
对 PSNR-Rate 曲线计算 BD-Rate 指标，评估同等质量下的码率节省。相比于 x265Multiview，所提方法在 Long、Loot 和 Red 三段视频上的 BD-Rate 分别为 -73.3%、-80.3% 和 -83.7%；相比于 ViewInterp，对应 BD-rate 分别为 -41.9%、-43.8% 和 -47.8%。总体而言，在相同重建质量下，所提方法相比于 x265Multiview 平均码率降低了 79.1%，相比于 ViewInterp 降低了 44.5%。

上述结果验证了所提方法在沉浸式自由视点视频通信场景中的带宽和率失真优势，相较于较传统全视点编码和视图插值方法，基于三维高斯重建的多视点视频编解码方

法能够更充分地挖掘空间结构信息，实现更优的压缩效率与重建质量平衡。



(a) PSNR 率失真曲线



(b) SSIM 率失真曲线

图 4.5　基于三维高斯重建的多视点编解码率失真曲线

4.5.3　重建质量主客观对比分析

在降低传输带宽的同时，重建质量是否能够保持稳定，是评价多视点视频编解码方法有效性的另一关键指标。本节从客观指标和主观视觉效果两个方面，对所提方法的重建质量进行了分析。

从客观评价结果来看，如图 4.4(b) 和图 4.4(c) 所示，所提方法在六组测试视频上的 PSNR 和 SSIM 均优于两类对比方法。相较于 x265Multiview，所提方法的 PSNR 提升幅度为 0.28%–9.46%，平均提升 4.27%；相较于 ViewInterp，提升幅度为 0.2%–10.78%，平均提升 2.13%。在 SSIM 方面，相较于 x265Multiview，所提方法提升了 12.75%–27.29%，平均提升 17.67%；相较于 ViewInterp，提升了 0.05%–4.09%，平均提升 1.16%。这一现

象表明，尽管所提方法仅编码并传输了少量关键视点，但得益于解码端三维高斯重建模型对场景几何与外观信息的有效建模，最终恢复出的多视点图像在整体结构一致性和纹理保真度上仍优于对比方法。

从主观视觉效果来看，图 4.8 和图 4.9展示了 Long、Loot、Red、Andrew、David 和 Phil 六组视频的典型重建结果，其中分别给出了 x265Multiview、ViewInterp、所提方法和真实图像之间的视觉对比。可以看出，x265Multiview 在低码率条件下容易出现压缩失真和细节退化；ViewInterp 虽然能够恢复中间视图，但在大视角变化、边界区域和人体复杂姿态处容易产生模糊、拖影或结构不连续现象。相比之下，所提方法恢复出的目标视图在人物轮廓、局部纹理以及遮挡区域的过渡上更接近真实图像，整体视觉质量更优。这说明三维高斯显式表示能够更好地描述人物场景的三维结构，从而在新视点生成过程中减轻仅基于视图插值所带来的几何失真。

值得注意的是,所提方法在显著降低带宽开销的同时仍能够获得优于 x265Multiview 的主客观质量结果，这并不矛盾。一方面，所提方法在每个视点组中仅编码少量参考视图，因此可以在有限带宽预算下为这些参考视图分配更高的目标码率，减轻压缩失真；另一方面，解码端引入的三维高斯重建与渲染机制也能够有效恢复未传输视图的空间细节。因此，所提方法在带宽效率和重建质量之间实现了较优平衡。

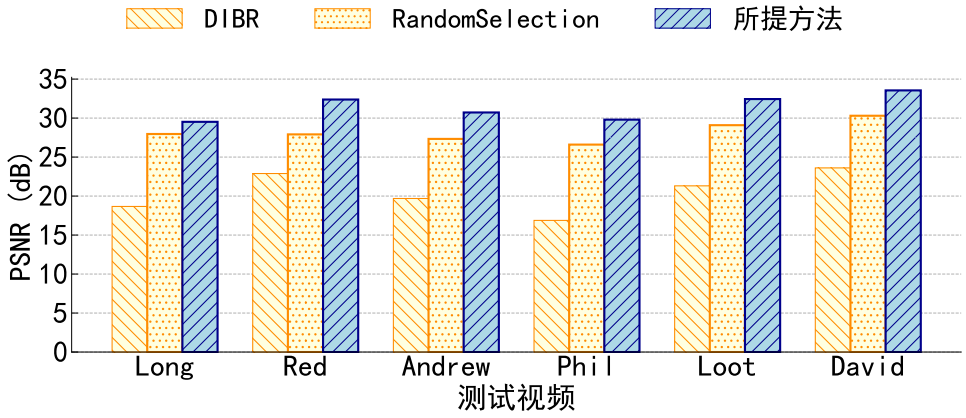
4.5.4　消融实验

为了进一步验证 4.3 节和 4.4 节中各关键设计的有效性，本节进行消融实验及分析。具体而言，设置两个变体方法与完整方法进行对比：其一为 DIBR 变体，即保留深度估计与视点选择过程，但将三维高斯重建与渲染替换为基于深度图的图像扭曲恢复；其二为 Random Selection 变体，即保留三维高斯解码框架，但不采用内容自适应视点选择，而是从候选视点中随机选取两路视图进行编码。由于这三种方案在实验中均仅编码两个视点，因此其带宽开销基本相近，本节主要比较其重建质量差异。

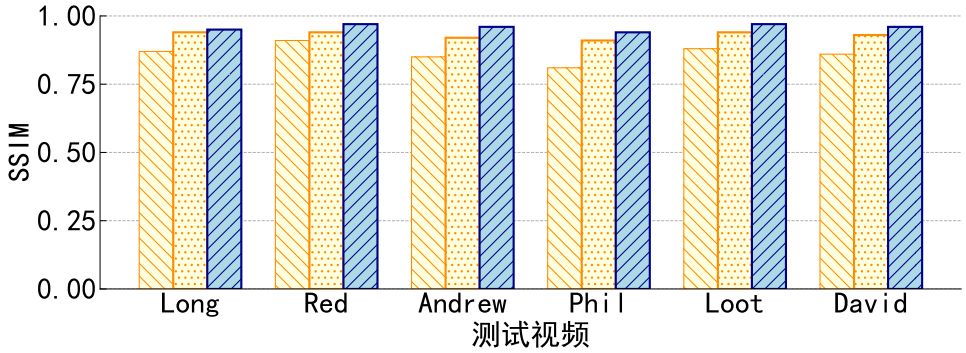
消融实验结果如图 4.6 所示，其中图 4.6(a) 和图 4.6(b) 分别给出了三种方法在六组测试视频上的 PSNR 和 SSIM 结果。可以看出，完整方法在所有测试序列上均取得了最优性能。与 DIBR 变体相比，所提方法的平均 PSNR 提升达到 54.46%，平均 SSIM 提升为 10.5%；与 Random Selection 变体相比，平均 PSNR 和 SSIM 分别进一步提升 11.41% 和 3.05%。

上述结果说明，所提方法的性能增益主要来自两个方面。首先，三维高斯重建与溅射渲染能够更好地处理目标视图生成中的遮挡关系、跨视角几何变形与纹理连续性问题，因此相较于仅基于深度图扭曲的 DIBR 方案具有明显优势。其次，内容自适应视点选择能够根据当前场景深度与几何关系动态调整输入视图组合，使解码端获得更

有利的立体匹配条件和更充分的空间覆盖，从而优于随机选点方案。由此可见，4.3 节提出的三维高斯解码模型与 4.4 节提出的自适应视点选择策略在整体框架中均发挥了关键作用。



(a) 所提方法和两个变体方法的 PSNR



(b) 所提方法和两个变体方法的 SSIM

图 4.6　基于三维高斯重建的多视点编解码及其变体的消融实验

4.5.5　关键参数影响分析

为了分析所提方法在不同视点密度条件下的扩展能力，本节进一步考察输入视点数量对系统性能的影响。实验中将总视点数从 10 增加到 50，记录不同设置下的带宽开销和重建质量，结果如图 4.7 所示。

从图 4.7 可以看出，随着总视点数从 10 增加到 50，系统的带宽消耗整体变化较小，基本保持稳定。这是因为所提方法无论在稀疏视点还是稠密视点设置下，编码端最终仅选取每组中的两路关键视点进行传输，因此总传输视点数并不会随着原始视点密度增加而显著增加。这一结果表明，所提方法在更大规模视点系统中具有较好的可扩展性。

与此同时，随着视点总数增加，系统的重建质量呈现持续提升趋势，PSNR 整体上升。其原因在于，当视点分布更加稠密时，编码端更容易在候选集合中找到与最优角

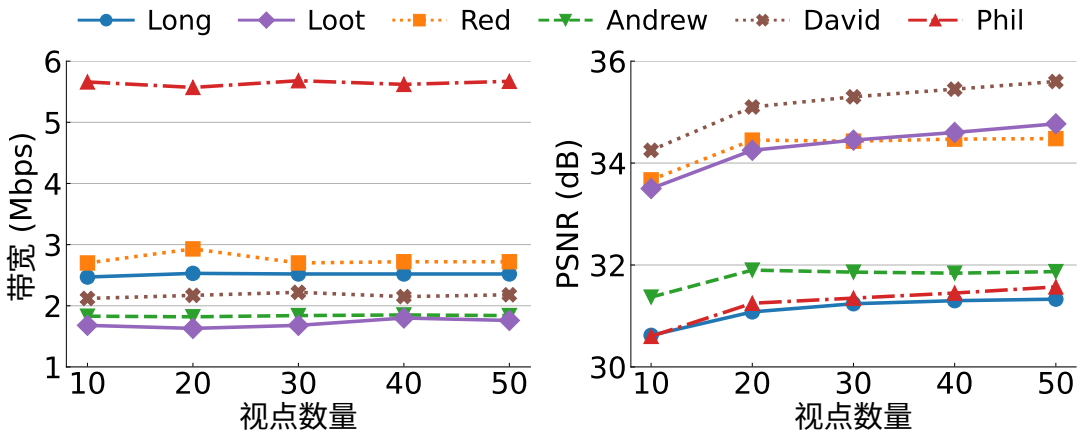


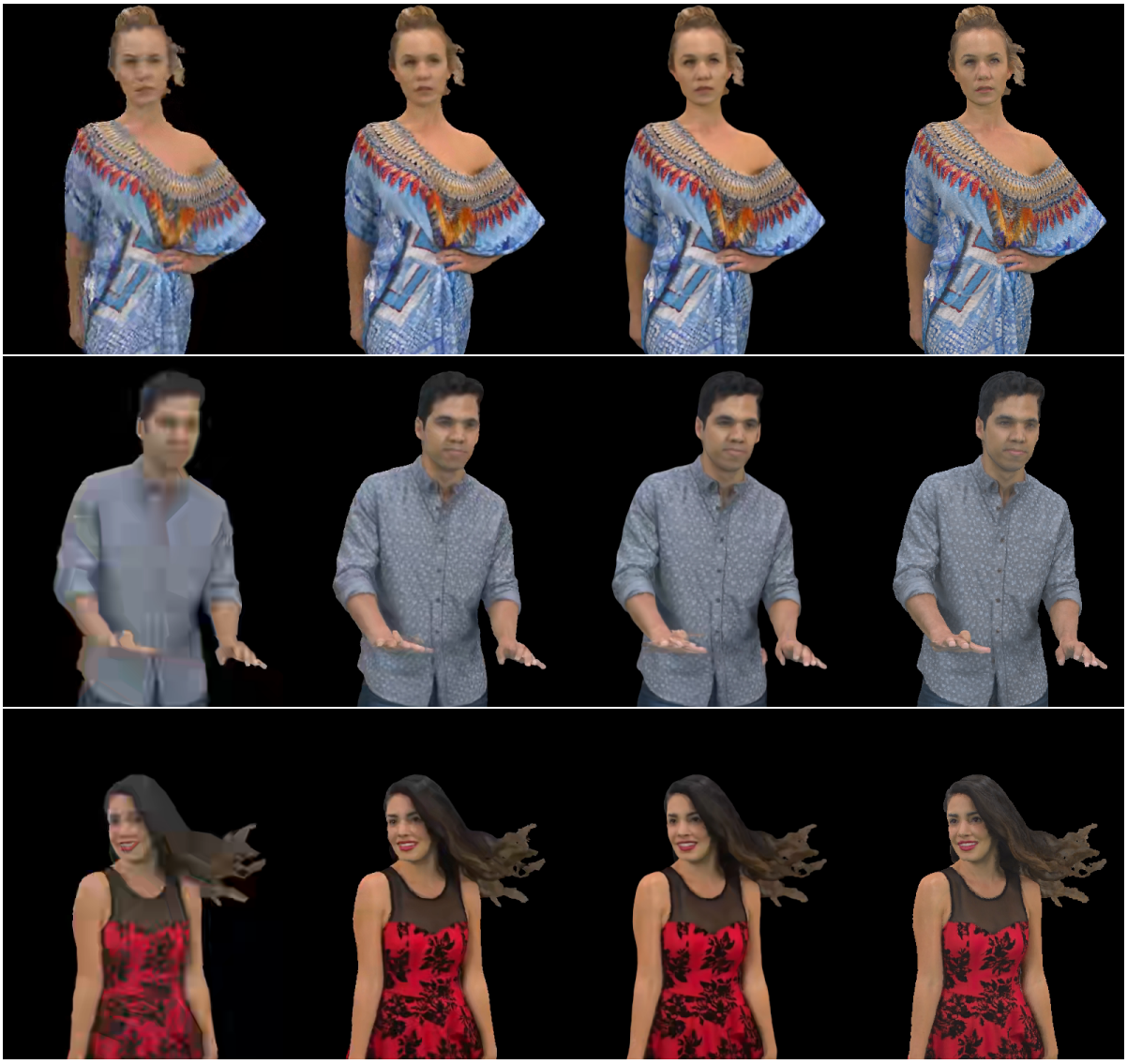
图 4.7　不同视点数量下的带宽和视觉质量评估

度参数更接近的参考视点对，从而减小视点选择误差，提高双目深度估计稳定性以及后续三维高斯重建精度。换言之，尽管所提方法只传输少量关键视点，但原始采集系统提供更高视点密度时，内容自适应视点选择模块能够从中挑选出更具代表性的输入，从而进一步改善重建效果。

因此，从参数分析结果可以看出，所提方法不仅能够在低带宽条件下工作，而且能够随着视点系统规模的扩大获得进一步的质量收益，这为其在大视点数量的裸眼三维显示系统中的应用提供了支持。

4.6　小结

本章围绕多视点视频在高带宽开销和实时传输约束下面临的问题，提出了基于三维高斯重建的多视点视频编解码方法。通过在编码端采用内容自适应视点选择策略，仅传输少量关键视点；在解码端结合双目深度估计、像素级三维高斯参数回归与高斯溅射渲染，恢复完整多视点视频输出，从而实现了带宽效率与重建质量的兼顾。实验结果表明，所提方法相较于全视点编码和基于视图插值的方案，能够在显著降低传输码率的同时保持较高的视觉质量。消融实验和参数分析进一步验证了三维高斯重建模块与内容自适应视点选择策略的有效性。上述结果说明，基于三维高斯重建的多视点视频编解码框架能够有效利用场景三维空间冗余，为低带宽条件下的高质量实时多视点视频传输提供了一种可行方案。



(a) *x265Multiview* (b) *ViewInterp* (c) 所提方法 (d) 真实值

图 4.8　多视点解码视觉质量定性对比 1



(a) *x265Multiview*　(b) *ViewInterp*　(c) 所提方法　(d) 真实值

图 4.9　多视点解码视觉质量定性对比 2

第五章　沉浸式自由视点视频通信系统

5.1　系统设计

本章面向沉浸式自由视点视频通信场景，构建了一套兼顾实时交互、带宽约束与裸眼三维显示需求的视频通信系统，并将第三章的快速视点合成方法与第四章的多视点视频编解码方法统一到同一端到端框架中。系统总体上遵循发送端采集与编码、网络传输、接收端解码与视点生成、终端显示的处理链路，通过在不同运行模式下调用不同的核心视点生成模块，实现单用户观看与多用户多视点光场显示两类应用能力。系统总体架构如图 5.1所示。

沉浸式自由视点视频通信系统需要在严格的端到端时延约束下，同时满足用户位置变化所带来的视点动态更新需求和裸眼三维显示对多视点图像的生成需求。与传统二维视频通信仅传输单路视频不同，沉浸式通信不仅涉及多视点视频的采集、组织与传输，还需要在接收端根据观察位置或显示设备要求生成与之对应的目标视图。因此，系统设计必须综合考虑视点生成复杂度、码流规模、传输时延以及终端显示方式之间的耦合关系。

基于上述需求，系统设计遵循以下原则。（1）端到端实时性。系统在采集、编码、传输、解码、视点生成与显示各环节均采用面向实时运行的处理流程，使系统具备交互式通信能力。（2）带宽友好性。对于沉浸式自由视点通信而言，若直接传输全部视点视频，将导致码率和传输开销迅速增长，因此系统应尽可能通过接收端视点生成降低待传输视点数量。（3）显示适配性。系统不仅要支持单用户场景下随观察位置变化而实时更新的双目显示，还应支持面向大尺寸裸眼三维显示终端的多视点并行输出，从而满足不同沉浸式应用场景的需求。

5.1.1　系统总体架构

如图 5.1所示，系统总体上由发送端、传输链路和接收端三部分组成。其中，发送端负责多视点视频源的获取、多视点同步与标定、图像预处理、视点组织以及编码发送；传输链路负责压缩码流和视频数据的实时传输；接收端负责码流解码、目标视图生成以及裸眼三维显示输出。该架构继承了传统视频通信系统采集、编码、传输、解码与显示的基本流程，同时将视点生成模块显式纳入接收端处理链路，使系统具备自由视点显示能力。

在发送端，系统支持两类视频源输入。一类是真实多相机阵列采集的视频流，通过同步触发、相机标定、图像去畸变和视图校正，获得具有已知几何关系的多视点视频；

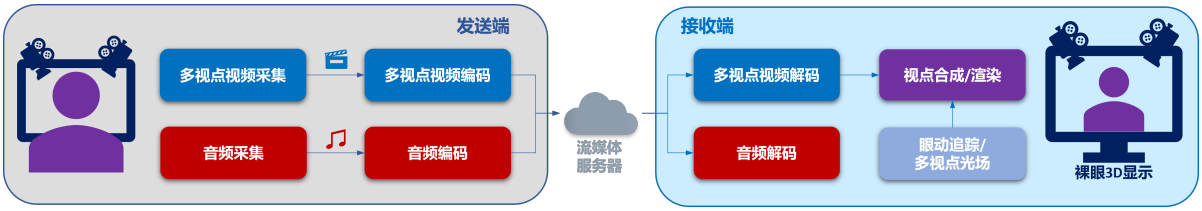


图 5.1　沉浸式自由视点视频通信系统总体架构

另一类是由体视频、点云或神经辐射场等三维内容渲染得到的多视点视频序列。前者用于真实系统验证，后者则能够通过灵活设置相机外参、相机间距和视角分布，为系统提供可控的数据输入与测试环境。经过预处理后，视频数据根据不同显示模式进入相应的视点组织与编码流程。

在传输链路上，系统依据运行模式采用不同的数据组织方式。对于单用户眼动驱动模式，系统以较少数量的参考视图作为输入，在保证交互连续性的前提下降低传输开销；对于多用户多视点光场显示模式，系统首先依据内容自适应原则选取稀疏视点进行编码与发送，再由接收端恢复和生成剩余目标视图。通过将部分视点生成任务后移到接收端，系统避免了全视点直接传输所带来的高码率问题，从而提高了整体传输效率。

在接收端，系统首先完成码流接收与解码，然后依据当前运行模式进入不同的视点生成分支。若系统处于单用户眼动驱动模式，则利用用户眼位或观察位置确定目标视点，并调用第三章提出的快速视点合成方法生成当前时刻所需的双目图像；若系统处于多用户多视点光场显示模式，则调用第四章提出的稀疏视点解码与三维高斯重建方法，在接收端恢复场景的中间三维表示，并进一步渲染出显示设备所需的稠密多视点图像集合。最终，生成的目标视图被送入裸眼三维显示终端，实现与观察位置相匹配的沉浸式观看体验。

5.1.2　运行模式设计

为兼顾交互式单人观看与多人同时观看两类典型应用场景，在统一系统架构下设计了单用户眼动驱动模式和多用户多视点光场显示模式两种运行方式，其差异如表 5.1所示。两类模式共享发送端采集、预处理与基础传输链路，但在接收端的视点生成目标、数据调度方式和显示输出形式上存在明显差异。其核心差异体现在接收端视点生成策略的选择：单用户模式强调基于用户位置的按需生成，以最小计算代价满足交互连续性；多用户模式强调基于多视点输入的批量重建与密集视点输出，以较低带宽支撑大范围观察覆盖。

表 5.1　沉浸式视频通信系统的两类运行模式设计

运行模式	单用户眼动驱动模式	多用户多视点光场显示模式
目标场景	单人观看	多人同时观看
显示终端	眼动跟踪式裸眼 3D 显示器	光场裸眼 3D 显示器
系统侧重点	高帧率、低时延	多视点一致性、低带宽

5.1.2.1　单用户眼动驱动模式

单用户眼动驱动模式面向交互式沉浸式通信场景，目标是在用户观察位置连续变化的情况下，实时生成与当前视点一致的双目图像。该模式以第三章提出的快速视点合成方法为核心，在接收端引入眼位跟踪模块，实时获取用户双眼位置或等效观察位置参数，并据此确定当前所需的目标视点。随后，系统从已解码的多视点视频中选取与目标视点相邻的参考视图，调用轻量化视点合成网络直接生成用于裸眼三维显示的双目图像。

该模式的关键特点在于，系统无需为所有可能观察方向同时生成全部视图，而仅需围绕当前用户实际观察位置生成少量目标图像，从而显著降低接收端计算负载和传输带宽压力。与此同时，由于第三章方法采用直接视图到视图的合成路径，避免了复杂三维重建与显式深度建模过程，因此更适合强调低时延和高帧率的实时交互场景。对于单人视频通信、远程交互展示等应用，该模式能够在有限算力条件下提供较强的视点连续性和临场感。

5.1.2.2　多用户多视点光场显示模式

多用户多视点光场显示模式面向多人同时观看或大屏裸眼三维展示场景，目标是在接收端输出覆盖不同观察方向的稠密多视点图像，以支撑无需佩戴设备的多人共享观看。该模式以第四章提出的基于高斯重建的多视点编解码方法为核心，在发送端首先根据内容自适应策略从原始多视点数据中选取少量关键视点进行编码和传输；接收端在完成稀疏视点解码后，进一步利用深度估计、三维高斯参数回归和高斯溅射渲染等模块，恢复中间三维高斯表示并合成显示所需的稠密视图集合。

与单用户模式相比，该模式不再围绕特定用户的瞬时观察位置进行按需生成，而是面向显示终端所需的全部或大部分观察方向统一生成视图，因此更强调带宽压缩能力、多视点覆盖能力以及大尺寸裸眼显示终端上的空间一致性。通过将三维高斯重建和多视点渲染置于接收端完成，系统能够在不显著增加发送端负担的前提下，以稀疏视图传输替代全视图传输，从而有效缓解多视点视频通信中的带宽瓶颈。对于数字展陈、远程汇报、多人会议等应用场景，该模式更具实用价值。

5.2　系统实现

在 5.1 节给出的总体架构基础上，本节介绍沉浸式自由视点视频通信系统的工程实现。系统整体采用发送端与接收端协同工作的双端部署方式，其中发送端负责多视点视频同步采集、预处理、编码与发送，接收端负责码流接收、解码、目标视点生成以及裸眼三维显示。考虑到单用户眼动驱动模式与多用户多视点光场显示模式在处理目标和算力需求上的差异，本章在统一系统框架下对双端处理流程进行了针对性实现。图 5.2 展示了该系统发送端和接收端的实现形态。



(a) 发送端多视点采集



(b) 接收端多视角呈现

图 5.2　沉浸式自由视点通信系统双端实现图

5.2.1　系统硬件平台配置

为同时满足单用户交互式自由视点显示与多用户多视点光场显示两类应用需求，系统在采集端、发送端、接收端和显示端均进行了面向实时处理的软硬件部署。总体而言，系统硬件由多视点采集设备、同步与标定设备、发送端计算平台、接收端计算平台、眼位跟踪设备以及裸眼三维显示终端构成；软件部分则包括相机标定与图像校正模块、视频编码与传输模块、视点生成与三维重建模块以及显示控制模块。

在多视点采集方面，系统采用由多台线性排布的 Basler acA2440-75uc 工业相机组成的相机阵列获取多视点视频，每台相机均配备 8 mm 焦距镜头。为实现多相机同步采集，系统采用硬件级同步触发方式：由信号发生器产生稳定的晶体管—晶体管逻辑

（TTL）方波信号，触发各相机同步曝光。系统的触发频率与采集帧率均设置为 30 Hz，单帧曝光时间为 20 ms，采集分辨率为 2448×2048，原始图像格式为 BayerRG8。为获得准确的相机参数，系统采用基于平面标定板的张氏相机标定方法[78] 标定各相机的内参与外参，具体标定过程通过 MATLAB 计算机视觉工具箱实现。

在计算平台方面，发送端主要负责图像预处理、视点组织、视频压缩与码流发送；接收端除承担视频解码任务外，还需要执行快速视点合成模块或三维高斯重建与多视点渲染模块。因此，在多用户多视点光场显示模式下，系统配置了更高的并行计算资源以支撑稠密视图生成。整个系统部署于两台工作站之上，每台工作站均配备 Intel Core i7-13700K CPU、64 GB 内存以及 3 块 NVIDIA RTX 4090 显卡。两台工作站位于同一局域网环境下，用于模拟两位用户进行远程视频通信的应用场景。每台工作站均可同时运行发送端与接收端程序。其中，单用户模式下系统仅使用 1 块显卡，多用户模式下则使用 3 块显卡协同完成计算任务。

在裸眼 3D 显示方面，系统根据不同运行模式采用两类显示终端。单用户模式下，系统采用带有眼位跟踪能力的 27 英寸 4K 裸眼 3D 显示设备，通过屏幕下方的内置摄像头实时检测用户双目位置，从而实现由观察位置驱动的双目视图显示。多用户模式下，系统采用 110 英寸 8K 多视点光场显示设备，该设备能够在 25° 视场角范围内同时渲染 50 个不同视点，从而支持多观察方向下的多人同时观看。

从工程实现角度看，两类运行模式共用采集、预处理和基础传输框架，但在显示终端和接收端算力配置上存在明显区别。单用户模式更强调低时延和交互连续性，因此接收端聚焦于少量目标视图的按需生成；多用户模式更强调多视点覆盖能力和显示一致性，因此接收端需要承担更大规模的重建与渲染任务。

5.2.2 发送端实现

发送端是沉浸式自由视点视频通信系统的数据入口，其主要功能包括多视点视频采集、视频预处理以及多视点编码与发送。为兼顾单用户眼动驱动模式与多用户多视点光场显示模式，发送端采用统一的采集与预处理框架，并根据不同运行模式在视频组织方式和编码策略上进行相应配置。

5.2.2.1 多视点视频采集

发送端首先完成多视点视频的获取。系统支持两类输入源：一类是真实场景下的多相机同步采集视频流，另一类是由体积视频、点云或神经辐射场渲染得到的多视点视频序列。前者主要用于真实系统部署与端到端通信验证，后者则能够通过灵活设置虚拟相机参数，获得可控的视点数量、视角范围和空间分布，从而为系统测试提供标准化输入。

对于真实采集场景，系统根据不同运行模式采用不同规模的相机阵列。在单用户眼动驱动模式下，发送端采用由 5 台相机构成的线性阵列进行多视点视频采集；在多用户多视点光场显示模式下，发送端采用由 9 台相机构成的线性阵列进行采集。两种模式下，相机阵列的整体长度均与显示终端的有效显示宽度相对应，以使采集空间与显示空间在水平视差范围上保持一致，从而更好地支撑后续视点合成与裸眼 3D 显示。

5.2.2.2　视频预处理

在完成多视点视频采集后，发送端需要对原始图像进行预处理，以提高后续编码压缩效率以及接收端视点生成的稳定性。预处理过程主要包括图像去畸变、视图校正、分辨率统一和颜色空间转换等步骤。

具体而言，系统首先对各视图执行去畸变与几何校正，使不同相机采集的图像尽可能对齐到统一参考平面，从而减小相机成像差异对后续处理的影响。随后，根据不同显示模式的需求，对图像进行裁剪与缩放，并统一至固定分辨率。其中，单用户眼动驱动模式下的图像分辨率设置为 1920×1080，多用户多视点光场显示模式下的图像分辨率设置为 1440×810。

在颜色空间转换方面，两种模式采用不同的处理流程。对于单用户眼动驱动模式，采集得到的 BayerRG8 原始图像直接转换为 NV12 格式，以满足后续实时编码与传输的需求。对于多用户多视点光场显示模式，考虑到发送端需要执行基于深度估计的内容自适应视点选择，系统首先将 BayerRG8 图像转换为 RGB 格式，以支持深度估计与视图分析等前端处理；在完成视点选择后，再将图像进一步转换为 NV12 格式用于编码与传输。

5.2.2.3　多视点编码

经过预处理后，发送端将多视点视频组织为待传输码流，并根据不同运行模式采用不同的编码策略，以在视觉质量、传输时延和带宽开销之间取得平衡。

在单用户眼动驱动模式下，系统将 5 路视频帧拼接为一帧更大的图像，拼接后分辨率为 3840×3240，并采用 HEVC 标准[9] 结合 NVENC 硬件编码器进行实时编码。编码完成后，压缩视频被进一步封装，并通过实时流媒体传输协议（RTSP）协议发送至流媒体服务器。

在多用户多视点光场显示模式下，系统采用稀疏关键视图传输策略。具体而言，发送端首先依据 4.4 节提出的内容自适应视点选择方法，将 9 个视点构成一个视点组，并从中选取一对关于中心对称的相机视图作为待编码对象。对于选定视图，系统采用 x265[76] 进行压缩编码，其中集成了实时 MV-HEVC[7] 实现。编码完成后，压缩视频同样通过 RTSP 发送至流媒体服务器。

5.2.3 接收端实现

接收端承担系统自由视点显示能力的核心实现任务，其主要功能包括多视点解码、视点合成和多视点重建以及 3D 显示控制。接收端首先对接收到的视频码流进行解码，然后根据系统当前运行模式进入不同的目标视图生成分支，最终将生成结果输出至相应的裸眼 3D 显示终端。

5.2.3.1 多视点解码

接收端首先完成网络码流的接收与多视点视频解码。系统从传输链路中接收发送端压缩后的多视点视频流，经过解码后恢复参考视图帧，并将其由 NV12 格式转换为 RGB 格式，以供后续视点生成模块使用。

在单用户眼动驱动模式下，接收端采用 NVDEC 硬件解码器对拼接后的多视点大图进行实时解码。解码完成后，系统进一步将大图拆分为 5 路参考视图，作为后续双目视点合成的输入。

在多用户多视点光场显示模式下，接收端采用 MV-HEVC 对发送端传输的稀疏关键视图进行解码，相关功能由 FFmpeg 7.1[77] 实现。解码得到的一对关键视图将作为后续完整多视点生成与场景重建的输入。

5.2.3.2 视点合成和多视点重建

在完成多视点解码后，接收端根据当前运行模式执行不同的目标视图生成策略。其中，单用户眼动驱动模式采用快速视点合成路径，多用户多视点光场显示模式采用三维高斯重建与多视点渲染路径，两者共同构成了系统接收端的核心计算模块。

对于单用户眼动驱动模式，接收端通过眼位跟踪模块实时获取用户双眼位置，并将其映射到显示坐标系下的目标观察位置。随后，系统依据该观察位置，在已解码的多视点视频中选取最邻近的参考视图，计算目标视点相对于参考视图的插值关系，并调用第 三 章提出的快速视点合成方法，生成当前时刻所需的左眼视图和右眼视图。该过程不依赖显式三维建模，而是直接基于相邻参考图像实现视图到视图的快速映射，因此具有较低的计算复杂度和较高的实时性，适用于单人交互式观看场景。

对于多用户多视点光场显示模式，接收端不再围绕单一用户的瞬时观察位置按需生成目标视图，而是利用解码后的稀疏关键视图恢复场景的三维高斯表示，并进一步渲染面向多个观察方向的目标视图。具体而言，系统首先通过第 四 章提出的三维高斯参数回归方法构建场景的三维高斯表示，随后利用高斯溅射渲染模块生成显示终端所需的稠密多视点图像集合。相较于单用户模式，该模式具有更高的计算复杂度，但能够以较少的传输视图数支撑更大范围的视点覆盖，并满足多人同时观看的显示需求。

表 5.2　单用户眼动驱动模式系统端到端延迟

模块	延迟（ms）	模块	延迟（ms）
视频采集	33	解码	5
编码	8	视点合成	8
传输	80	显示	8
		总延迟	142

5.2.3.3　3D 显示

完成视点合成和多视点重建后，接收端首先根据显示终端的输入格式要求，对生成的目标图像进行分辨率调整和像素排布处理，随后输出至对应的 3D 显示终端，从而实现沉浸式自由视点观看。

综上，接收端在统一框架下，通过快速视点合成和基于三维高斯重建的多视点渲染两条处理路径，实现了面向单用户交互式观看和多用户共享观看的两类显示功能。

5.3　系统测试与分析

本节从端到端时延、实时处理能力以及带宽占用三个方面，对所构建的沉浸式自由视点视频通信系统进行测试与分析。考虑到单用户眼动驱动模式与多用户多视点光场显示模式在目标视图生成方式、接收端计算规模以及传输数据组织上的差异，本节分别对两类模式进行讨论。

5.3.1　端到端时延与实时性

(1) 单用户眼动驱动模式

本节对单用户眼动驱动模式下通信系统的端到端时延进行了测量。实验中，通过拍摄计时器并比较其在显示端呈现时的时间差，对用户实际感知到的系统时延进行估计。实验结果表明，该模式下系统的端到端时延约为 142 ms。根据 Chen 等[79] 的结论，该时延处于远程视频通信可接受的范围之内。系统各模块的时延分解结果如表 5.2 所示。在完成多视点视频采集、预处理与编码后，接收端能够稳定输出分辨率为 1080p、帧率为 30 FPS 的双目显示结果，从而保证用户观察位置变化过程中的视觉连续性。

从接收端视点合成模块的运行效率来看，第三章所采用的加速实现对系统实时性起到了关键作用。通过输入降采样、参数量化以及 TensorRT 部署等优化手段，双目视点合成的平均推理时间由原始实现的 34.9 ms 降低至 7.7 ms，该推理时间仅占 30 FPS 帧周期缓冲时间的 23%，说明视点合成模块能够在留有一定时延余量的情况下完成实时处理。因此，在单用户模式下，系统整体时延主要由采集、编码、传输与显示链路共同决定，而视点合成本身已不再构成主要瓶颈。

表 5.3　多用户多视点光场显示模式系统端到端延迟

模块	延迟（ms）	模块	延迟（ms）
视频采集	35	3DGS 重建	16
编码	33	3DGS 渲染	23
传输	80	显示	8
解码	3	总延迟	198

(2) 多用户多视点光场显示模式

本节对多用户多视点光场显示模式下通信系统的端到端时延进行了测量。实验结果表明，该模式下系统的平均端到端时延为 198 ms。根据 Chen 等[79] 的结论，该时延同样能够满足沉浸式多视点视频通信的实时显示需求。系统各模块的时延分解结果如表 5.3 所示。

具体来看，系统各环节的平均时延分别为：视频采集 35 ms、编码 33 ms、传输 80 ms、解码 3 ms、三维高斯重建 16 ms、高斯溅射渲染 23 ms、显示 8 ms，总时延为 198 ms。可以看出，传输链路以及前端采集仍然占据了较大的时延开销，而接收端中的三维高斯重建与渲染部分已被压缩到较低水平，说明本章提出的前馈式三维高斯参数回归模型具有较高的在线推理效率。

在实时性方面，系统发送端平均每帧编码耗时约为 33 ms，对应运行速度约为 30 FPS；在接收端，借助两块 GPU 的并行计算，整个解码与重建过程平均每帧耗时约为 21 ms。其中，经过 TensorRT 优化后的三维高斯重建网络耗时约为 16 ms，而 3DGS 光栅化渲染平均每视点仅需 0.76 ms，能够满足实时多视点视频通信的基本要求。

需要指出的是，多用户模式面向的是 50 个视点的并行显示需求，因此上述实时性结果是在较大视图生成规模下取得的。与单用户模式相比，多用户模式虽然具有更高的计算开销，但仍能够维持 30 FPS 的运行速度，说明第四章提出的基于稀疏视图传输与三维高斯重建的多视点生成方法具有较好的系统部署能力。

5.3.2　带宽占用

(1) 单用户眼动驱动模式

单用户眼动驱动模式下，系统仅需传输用于双目视点合成的少量参考视图，因此整体传输负担相对较低。实验中，系统采用 15 Mbps 的恒定码率控制策略对多视点视频进行压缩传输，在保证实时交互能力的同时能够维持较好的显示质量。

与已有沉浸式系统通常需要 30-100 Mbps 带宽相比，该模式在保持实时双目显示能力的同时，显著降低了带宽占用。这表明，基于相邻参考视图的快速视点合成路径能够在不引入显式三维表示传输开销的前提下，实现较为高效的沉浸式自由视点视频通信，因而更适用于强调低时延、低带宽和交互连续性的单用户应用场景。

(2) 多用户多视点光场显示模式

多用户多视点光场显示模式面向稠密多视点输出需求，对系统带宽效率提出了更高要求。与传统全视点直接传输方案相比，采用稀疏关键视图传输与接收端多视点重建相结合的策略，从而有效降低了通信链路中的数据传输开销。

实验结果表明，在 4.5.2 节所采用的测试视频上，系统仅需约 1.76-5.67 Mbps 即可支撑多视点显示结果的生成。结合实验中发送端与接收端在 15 Mbps 以太网环境下的稳定运行情况可以看出，该模式已经能够在较为常见的网络条件下实现多视点视频的实时传输与显示。

上述结果说明，通过将部分视点生成任务后移至接收端，系统能够以较少的传输视图数支撑稠密多视点输出，从而在保证显示能力的同时显著降低带宽占用。这也进一步验证了第四章所提出方法在带宽受限场景中的应用潜力。

5.4　小结

本章围绕沉浸式自由视点视频通信系统的构建与验证，完成了系统设计、系统实现以及系统测试与分析。本章面向单用户眼动驱动显示和多用户多视点光场显示两类应用场景，建立了从多视点视频采集、预处理、编码传输，到接收端解码、视点合成或多视点重建，再到裸眼 3D 显示输出的完整处理链路。实验结果表明，单用户模式能够以较低时延实现稳定的双目实时显示，多用户模式能够在较低带宽条件下实现稠密多视点重建与实时显示。上述结果验证了第三章和第四章所提出方法在统一系统框架下的有效集成，也说明所提出的关键技术具备面向实际沉浸式自由视点视频通信场景的系统部署能力。

第六章　总结与展望

6.1　本文总结

随着多媒体通信、计算机视觉与三维显示技术的不断发展，视频通信正逐步由传统二维视频向具备空间感知能力与自由观看特性的沉浸式形态演进。在实际系统部署中，受限于带宽、算力以及端到端时延约束，如何在保证实时性的前提下实现高质量视点生成与高效多视点视频传输，仍然是亟待解决的关键问题。

针对沉浸式自由视点视频通信中的关键挑战，本文从接收端视点生成与多视点视频高效组织与传输两个方面展开研究。面向稀疏视点输入与实时交互需求，研究低复杂度、高质量的视点合成方法；针对多视点视频数据规模庞大带来的带宽压力，探索高效的编解码策略。本文主要工作总结如下：

（1）针对视点生成问题，提出了一种基于多尺度光流预测的快速视点合成方法。通过对参考视图进行几何校正与对齐，降低跨视点像素运动复杂度，并构建多尺度光流预测网络，实现目标视点的高效重建。实验结果表明，该方法在多个公开数据集及真实采集场景中均能够获得稳定的重建质量，同时具备较低的计算复杂度。

（2）针对多视点视频编解码问题，提出了一种基于三维高斯重建的多视点视频解码方法，并设计了内容自适应视点选择与编码策略。该方法通过双目深度估计与像素级三维高斯参数回归，实现从稀疏视点到目标视点的前馈式重建，避免了传统方法中依赖逐场景优化带来的高时延问题。实验结果表明，所提出的视点选择与编码策略能够显著降低传输带宽，并保持较高的重建质量，在率失真性能方面表现出良好的优势。

（3）在系统应用方面，设计并实现了沉浸式自由视点视频通信系统。系统支持单用户眼动驱动模式与多用户多视点显示模式，并在真实网络环境下进行了测试。实验结果表明，整体帧率满足实时应用需求，验证了所提出方法在实际沉浸式视频通信场景中的可行性与有效性。

6.2　研究展望

尽管本文围绕沉浸式自由视点视频通信的关键问题进行了探索，但在更复杂场景与更高性能需求下，仍有进一步研究空间，主要体现在以下几个方面：

（1）复杂动态场景下的稳健视点生成。当前方法在一般动态场景中已能够实现稳定运行，但在存在大尺度运动、低相邻视图相关性或严重遮挡的情况下，视点生成的稳定性与细节一致性仍有提升空间。未来可结合更强的时序建模与运动约束，提高方

法对复杂动态变化的适应能力。

（2）面向通信场景的三维表示优化。虽然三维高斯表示在重建效率与质量之间取得了较好平衡，但在大规模、高分辨率多视点视频场景中，其参数规模与更新开销仍可能成为系统瓶颈。未来可进一步研究更加紧致的三维表示方式，以及与视频编码机制深度融合的压缩策略。

（3）面向系统落地的扩展性问题。当前系统在接收端仍依赖高性能计算设备以支持三维重建与实时渲染过程，在多用户通信与多终端协同场景下，系统的计算与传输压力将进一步增加，单节点处理模式难以满足大规模应用需求。未来研究可围绕轻量化部署与分布式计算展开，通过探索移动端适配、分布式渲染以及边缘计算等技术路径，结合任务划分与跨端协同机制，降低单节点计算负担，从而提升系统在实际应用场景中的扩展能力。

参考文献

[1] LAWRENCE J, OVERBECK R, PRIVES T, et al. Project starline: A high-fidelity telepresence system[G]//ACM SIGGRAPH 2024 emerging technologies. 2024: 1-2.

[2] ZHANG Y, YANG J, LIU Z, et al. Virtualcube: An immersive 3d video communication system[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(5): 2146-2156.

[3] TU H, SHAO R, DONG X, et al. Tele-Aloha: A telepresence system with low-budget and high-authenticity using sparse rgb cameras[C]//ACM SIGGRAPH 2024 Conference Papers. 2024: 1-12.

[4] 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.

[5] KERBL B, KOPANAS G, LEIMKÜHLER T, et al. 3D Gaussian splatting for real-time radiance field rendering.[J]. ACM Trans. Graph., 2023, 42(4): 139-1.

[6] WIEGAND T, SULLIVAN G J, BJONTEGAARD G, et al. Overview of the H. 264/AVC video coding standard[J]. IEEE Transactions on circuits and systems for video technology, 2003, 13(7): 560-576.

[7] TECH G, CHEN Y, MÜLLER K, et al. Overview of the multiview and 3D extensions of high efficiency video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 26(1): 35-49.

[8] BOYCE J M, DORÉ R, DZIEMBOWSKI A, et al. MPEG immersive video coding standard[J]. Proceedings of the IEEE, 2021, 109(9): 1521-1536.

[9] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on circuits and systems for video technology, 2012, 22(12): 1649-1668.

[10] GAO W, MA S, ZHANG L, et al. AVS video coding standard[G]//Intelligent multimedia communication: techniques and applications. Springer, 2010: 125-166.

[11] GAO W, MA S. An overview of AVS2 standard[J]. Advanced Video Coding Systems, 2015: 35-49.

[12] ZHANG J, JIA C, LEI M, et al. Recent development of AVS video coding standard: AVS3[C]//2019 picture coding symposium (PCS). 2019: 1-5.

[13] SCHWARZ S, PREDA M, BARONCINI V, et al. Emerging MPEG standards for point cloud compression[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2018, 9(1): 133-148.

[14] HEDMAN P, SRINIVASAN P P, MILDENHALL B, et al. Baking neural radiance fields for real-time view synthesis[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 5875-5884.

[15] LI L, SHEN Z, WANG Z, et al. Compressing volumetric radiance fields to 1 mb[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 4222-4231.

[16] LI S, LI H, LIAO Y, et al. Nerfcodec: Neural feature compression meets neural radiance fields for

memory-efficient scene representation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 21274-21283.

[17] LI L, SHEN Z, WANG Z, et al. Streaming radiance fields for 3d video synthesis[J]. Advances in Neural Information Processing Systems, 2022, 35: 13485-13498.

[18] WANG L, HU Q, HE Q, et al. Neural residual radiance fields for streamably free-viewpoint videos [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 76-87.

[19] ZHANG Z, LU G, LIANG H, et al. Rate-aware compression for nerf-based volumetric video[C]// Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 3974-3983.

[20] LU T, YU M, XU L, et al. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 20654-20664.

[21] CHEN Y, WU Q, LIN W, et al. Hac: Hash-grid assisted context for 3d gaussian splatting compression [C]//European Conference on Computer Vision. 2024: 422-438.

[22] SUN J, JIAO H, LI G, et al. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 20675-20685.

[23] DAI P, ZHANG P, DONG Z, et al. 4d gaussian videos with motion layering[J]. ACM Transactions on Graphics (TOG), 2025, 44(4): 1-14.

[24] LI H, LI S, GAO X, et al. Gifstream: 4d gaussian-based immersive video with feature stream[C]// Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 21761-21770.

[25] FEHN C. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV[C]//Stereoscopic displays and virtual reality systems XI: vol. 5291. 2004: 93-104.

[26] ZITNICK C L, KANG S B, UYTTENDAELE M, et al. High-quality video view interpolation using a layered representation[J]. ACM transactions on graphics (TOG), 2004, 23(3): 600-608.

[27] OH K J, YEA S, HO Y S. Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video[C]//2009 Picture coding symposium. 2009: 1-4.

[28] DO L, ZINGER S, et al. Quality improving techniques for free-viewpoint DIBR[C]//Stereoscopic Displays and Applications XXI: vol. 7524. 2010: 159-168.

[29] SCHARSTEIN D. View synthesis using stereo vision[M]. Springer, 2001.

[30] MUELLER K, SMOLIC A, DIX K, et al. View synthesis for advanced 3D video systems[J]. EURASIP Journal on image and video processing, 2009, 2008(1): 438148.

[31] MCMILLAN L, BISHOP G. Plenoptic modeling: An image-based rendering system[G]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 433-440.

[32] FITZGIBBON A, WEXLER Y, ZISSERMAN A. Image-based rendering using image-based priors [J]. International Journal of Computer Vision, 2005, 63(2): 141-151.

[33] FLYNN J, NEULANDER I, PHILBIN J, et al. Deepstereo: Learning to predict new views from the world's imagery[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.

2016: 5515-5524.

[34] ZHOU T, TUCKER R, FLYNN J, et al. Stereo magnification: Learning view synthesis using multiplane images[J]. arXiv preprint arXiv:1805.09817, 2018.

[35] WILES O, GKIOXARI G, SZELISKI R, et al. Synsin: End-to-end view synthesis from a single image[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7467-7477.

[36] CURLESS B, LEVOY M. A volumetric method for building complex models from range images[C]//Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. 1996: 303-312.

[37] NEWCOMBE R A, IZADI S, HILLIGES O, et al. Kinectfusion: Real-time dense surface mapping and tracking[C]//2011 10th IEEE international symposium on mixed and augmented reality. 2011: 127-136.

[38] 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.

[39] YAO Y, LUO Z, LI S, et al. Recurrent mvsnet for high-resolution multi-view stereo depth inference [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5525-5534.

[40] YU Z, GAO S. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1949-1958.

[41] GU X, FAN Z, ZHU S, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2495-2504.

[42] HEDMAN P, PHILIP J, PRICE T, et al. Deep blending for free-viewpoint image-based rendering[J]. ACM Transactions on Graphics (ToG), 2018, 37(6): 1-15.

[43] JIANG H, SUN D, JAMPANI V, et al. Super slomo: High quality estimation of multiple intermediate frames for video interpolation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9000-9008.

[44] NIKLAUS S, LIU F. Softmax splatting for video frame interpolation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5437-5446.

[45] HUANG Z, ZHANG T, HENG W, et al. Real-time intermediate flow estimation for video frame interpolation[C]//European Conference on Computer Vision. 2022: 624-642.

[46] YU A, LI R, TANCIK M, et al. Plenoctrees for real-time rendering of neural radiance fields[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 5752-5761.

[47] GARBIN S J, KOWALSKI M, JOHNSON M, et al. Fastnerf: High-fidelity neural rendering at 200fps [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 14346-14355.

[48] WANG Q, WANG Z, GENOVA K, et al. Ibrnet: Learning multi-view image-based rendering[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 4690-4699.

[49] LIN H, PENG S, XU Z, et al. Efficient neural radiance fields for interactive free-viewpoint video[C]//SIGGRAPH Asia 2022 Conference Papers. 2022: 1-9.

[50] MÜLLER T, EVANS A, SCHIED C, et al. Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM transactions on graphics (TOG), 2022, 41(4): 1-15.

[51] KEETHA N, KARHADE J, JATAVALLABHULA K M, et al. Splatam: Splat track & map 3d gaussians for dense rgb-d slam[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 21357-21366.

[52] DAI P, XU J, XIE W, et al. High-quality surface reconstruction using gaussian surfels[C]//ACM SIGGRAPH 2024 conference papers. 2024: 1-11.

[53] CHARATAN D, LI S L, TAGLIASACCHI A, et al. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 19457-19467.

[54] SZYMANOWICZ S, RUPPRECHT C, VEDALDI A. Splatter image: Ultra-fast single-view 3d reconstruction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 10208-10217.

[55] LIU T, WANG G, HU S, et al. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo[C]//European Conference on Computer Vision. 2024: 37-53.

[56] ZHENG S, ZHOU B, SHAO R, et al. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 19680-19690.

[57] WANG J, CHEN M, KARAEV N, et al. Vggt: Visual geometry grounded transformer[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 5294-5306.

[58] JIANG L, MAO Y, XU L, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views[J]. ACM Transactions on Graphics (TOG), 2025, 44(6): 1-16.

[59] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[60] CARBALLEIRA P, CARMONA C, DÍAZ C, et al. FVV live: A real-time free-viewpoint video system with consumer electronics hardware[J]. IEEE Transactions on Multimedia, 2021, 24: 2378-2391.

[61] ZHOU Y, WANG S, LI W, et al. Live4D: a real-time capture system for streamable volumetric video [G]//SIGGRAPH Asia 2023 technical communications. 2023: 1-4.

[62] STENGEL M, NAGANO K, LIU C, et al. Ai-mediated 3d video conferencing[G]//ACM SIGGRAPH 2023 Emerging Technologies. 2023: 1-2.

[63] NOVER H, ACHAR S, GOLDMAN D. ESPReSSo: efficient slanted PatchMatch for real-time space-time stereo[C]//2018 International Conference on 3D Vision (3DV). 2018: 578-586.

[64] ZHENG Y, PARK S, ZHANG X, et al. Self-learning transformations for improving gaze and head redirection[J]. Advances in Neural Information Processing Systems, 2020, 33: 13127-13138.

[65] XUE T, CHEN B, WU J, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127(8): 1106-1125.

[66] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.

[67] DZIEMBOWSKI A, MIELOCH D, STANKOWSKI J, et al. IV-PSNR—The objective quality metric for immersive video applications[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7575-7591.

[68] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.

[69] KARA Y E, GENC G, ARAN O, et al. Actively estimating crowd annotation consensus[J]. Journal of Artificial Intelligence Research, 2018, 61: 363-405.

[70] HU J, GUO S, DONG Y, et al. A Multi-User Oriented Live Free-Viewpoint Video Streaming System Based on View Interpolation[C/OL]//2022 IEEE International Conference on Multimedia and Expo (ICME). 2022: 1-6. DOI: 10.1109/ICME52920.2022.9859922.

[71] TEED Z, DENG J. Raft: Recurrent all-pairs field transforms for optical flow[C]//European conference on computer vision. 2020: 402-419.

[72] D' EON E, HARRISON B, MYERS T, et al. 8i voxelized full bodies-a voxelized point cloud dataset [J]. ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006, 2017, 7(8): 11.

[73] LOOP C, CAI Q, ESCOLANO S O, et al. Microsoft voxelized upper bodies-a voxelized point cloud dataset[J]. ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document m38673 M, 2016, 72012(42): 2016.

[74] MAYER N, ILG E, HAUSSER P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4040-4048.

[75] YU T, ZHENG Z, GUO K, et al. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 5746-5756.

[76] x265[EB/OL]. 2024. https://x265.readthedocs.io/en/master/releasenotes.html.

[77] FFmpeg[EB/OL]. 2024. https://www.ffmpeg.org/.

[78] ZHANG Z. A flexible new technique for camera calibration[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 22(11): 1330-1334.

[79] CHEN Y, FARLEY T, YE N. QoS requirements of network applications on the Internet[J]. Information Knowledge Systems Management, 2004, 4(1): 55-76.

攻读硕士期间发表的论文及其他成果

[1] **Hongbo Ning**, Zihao Zhou, Siwei Ma. High-Performance and Low-Cost Immersive Telepresence System with 3D Gaussian Reconstruction. Accepted by 2026 5th International Conference on Intelligent Systems, Communications and Computer Networks, 2026.

[2] Yizong Wang, **Hongbo Ning**, Haohua Wang, Yutao Yuan, Yue Ling, Dong Zhao, Siwei Ma, Wen Gao. Argus: Bandwidth-Efficient Live Multiview Video Streaming via Sparse-View Gaussian Reconstruction. Accepted by IEEE INFOCOM 2026 - IEEE Conference on Computer Communications.[CCF-A]

[3] Hongyue Huang, Xilong Zhou, **Hongbo Ning**, Haopeng Lu, Qi Zhang, Yanpeng Liang, Wanjun Lyu, Chuanmin Jia, Xinfeng Zhang, Liuxin Zhang, Siwei Ma. Low-Complexity 3D-Vision Conferencing System based on Accelerated RIFE Model[C/OL]//2024 Picture Coding Symposium (PCS), 2024, pp. 1-5, doi: 10.1109/PCS60826.2024.10566458.

致谢

三年时光白驹过隙，学生生涯的旅程即将到站。值此论文完成之际，谨向所有在学习、科研和生活中给予我帮助、支持与陪伴的人致以诚挚的感谢。

首先，我要郑重感谢我的导师马思伟教授。有幸师从马老师，是我读研期间最为珍贵的经历。马老师平易近人，为人谦和，学识深厚，治学严谨，是我学术道路上的引路人。从课题选择、论文开题到撰写、修改，马老师始终给予我耐心细致的指导。马老师严谨的治学态度和温和的处事方式，都让我受益匪浅。

感谢张新峰老师、贾川民老师和王义总老师在科研与生活中给予我的指导、关心与帮助。感谢师兄师姐和同门伙伴们的陪伴与帮助。做实验、写论文、赶项目时并肩作战的日夜，组会中分享经验、共同讨论、彼此鼓励的时刻，论文和项目完成后的狂欢，还有那些被投喂的零食和奶茶……这些都是我读研生涯里珍贵而温暖的回忆。能和一群有趣、靠谱又优秀的人一起学习和共事，真的很幸运。

最想感谢的，是我的家人。他们虽然不懂我的研究方向也不懂我的论文，但他们始终是我最坚实的后盾，默默支持我的每一个决定，包容我的焦虑和疲惫，让我能够安心前行。

最后，要感谢那个不曾放弃的自己，谢谢自己在无数个想摆烂的时刻咬牙坚持，谢谢自己在迷茫时没有停下脚步，谢谢自己跌跌撞撞，终于走到了这里。

谨以此文，献给所有陪伴我走过这段旅程的人。