Poster: Mobile Volumetric Video Streaming Enhanced by Super Resolution

Anlan Zhang Chendong Wang Xing Liu Bo Han* Feng Qian
University of Minnesota, Twin Cities *AT&T Labs – Research

ABSTRACT

Volumetric videos allow viewers to exercise 6-DoF (degrees of freedom) movement when watching them. Due to their true 3D nature, streaming volumetric videos is highly bandwidth demanding. In this work, we present to our knowledge a first volumetric video streaming system that leverages deep super resolution (SR) to boost the video quality on commodity mobile devices. We propose a series of judicious optimizations to make SR efficient on mobile devices.

CCS CONCEPTS

• Information systems → Multimedia streaming; **• Human-centered computing → Mobile computing**.

ACM Reference Format:
Anlan Zhang, Chendong Wang, Xing Liu, Bo Han, and Feng Qian. 2020. Poster: Mobile Volumetric Video Streaming Enhanced by Super Resolution. In *The 18th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '20), June 15–19, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3386901.3396598

1 INTRODUCTION

Volumetric video is an emerging type of video content that allows its viewers to have 6-DoF (degrees of freedom) movement during playback, where viewers can change their positions (X, Y, Z) and orientations (yaw, pitch, roll) freely when watching a video. Unlike regular videos or 360-degree videos, volumetric videos consist of 3D points or meshes, making them highly immersive and interactive.

Volumetric videos can be captured using RGB-D cameras with depth sensors (Figure 1). They can be stored in different ways including 3D meshes and point clouds. In this project, we focus on the most popular *Point Cloud* (PtCl) based representation where each video frame is represented as a collection of points. Due to their true 3D nature, volumetric videos have numerous applications in, for example, healthcare, education, and military training. However, streaming them is extremely challenging from the perspective of bandwidth consumption. For a high-resolution PtCl footage, its data rate can be as high as 6 Gbps. Even after compression, the required bandwidth may still be prohibitively high.

In this poster, we present our ongoing work on developing VoluSR, a novel system for streaming high-quality volumetric content wirelessly to commodity mobile devices. The key contributions of VoluSR consist of the following.

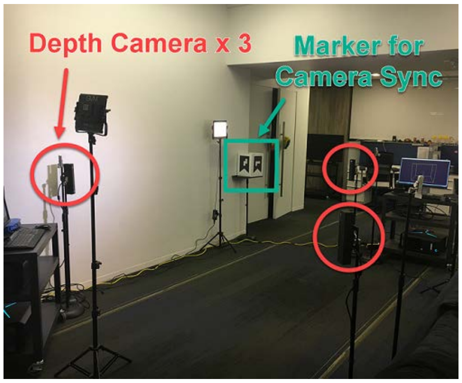

Figure 1: The volumetric video capturing system in our lab.

- We apply super resolution (SR), a deep-learning-powered technique that can significantly boost the visual quality. SR leverages model overfitting to achieve intelligent content compression. In a typical SR scheme, we train a DNN model to learn how to reconstruct high-resolution details from low-resolution content. Then in the online inference stage, low-resolution images or frames of the *same* content, which can be efficiently transmitted over the bandwidth-constrained networks, are fed into the model to derive the high-resolution content (called *upsampling*). Although SR has been applied to regular 2D videos [3], VoluSR is to our knowledge the first volumetric video streaming system enhanced by SR.
- Applying SR to upsample a video frame involves performing an inference on a heavy-weight DNN model, making its performance on mobile devices (and even on PCs with state-of-the-art GPUs) falling far short of supporting smooth playback at 30 FPS or higher. To address this critical challenge, VoluSR applies a serious of optimizations, at both the system and the algorithmic levels, to make SR feasible on commodity mobile devices. They include strategically simplifying SR models, aggressively caching/reusing inference results, and judiciously adapting to users' perception and mobile devices' heterogeneous computation capabilities.
- We are integrating the above design into a holistic system, and plan to evaluate it against real volumetric videos. We are also in the progress of recruiting voluntary participants to assess our system through an IRB-approved study.

2 SR FOR 3D POINT CLOUD VIDEOS

Compared to SR for 2D content (images and regular videos), SR for 3D content remains an emerging research topic. We first demonstrate that 3D SR can effectively improve the volumetric video quality. We apply PU-GAN [2], a very recently developed 3D SR model on a PtCl video captured at our lab (Figure 1). The raw footage of the

MobiSys '20, June 15–19, 2020, Toronto, ON, Canada
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7954-0/20/06.
https://doi.org/10.1145/3386901.3396598

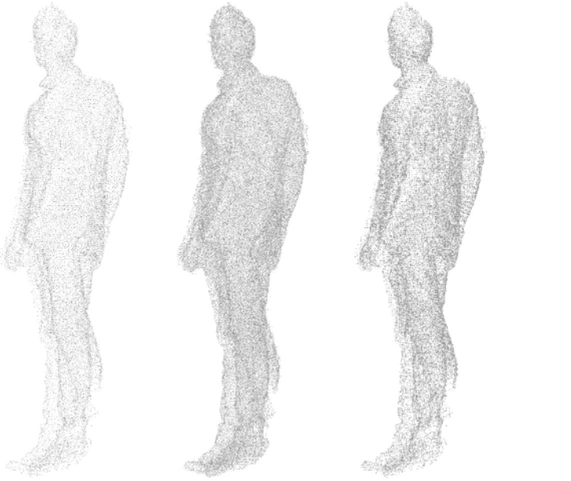

Figure 2: Left: a low-resolution frame (input to PU-GAN), Middle: the high-resolution frame inferred by PU-GAN, Right: the ground truth high-resolution frame.

video consists of about 3000 frames each consisting of around 100K points. We use this video to train a vanilla PU-GAN model, and then use it to upsample 100 low-resolution frames at about 25K points per frames. In other words, the upsampling ratio is set to 4, which can achieve a bandwidth saving of ~75% under the same quality level, or a video resolution increase of 300% under the same bandwidth usage. We conduct our experiments on a desktop PC with an NVIDIA 2080Ti GPU.

We find that the model achieves good accuracy as visualized in Figure 2. Quantitatively, the average Chamfer Distance (CD) between the inference result and the ground truth across all frames is measured to be $0.33\times10^{-3}m^2$, indicating that PU-GAN can generate upsampled PtCls whose geometric structures are fairly close to those of the ground truth [2]. However, the downside is the poor runtime performance. Even on a desktop PC with a state-of-the-art GPU, the inference speed is less than 2 FPS, not to mention executing the inference task on mobile devices with weaker computational capabilities. Our pilot results indicate the need for heavy optimizations that should strike a tradeoff between video quality improvement and runtime performance.

3 PROPOSED OPTIMIZATIONS

VoluSR employs several critical optimizations to achieve the goal of supporting SR for 3D PtCl content on mobile devices, as to be detailed below.

Speeding up Model Inference. We accelerate the inference process by judiciously modifying the PU-GAN model. First, we employ general DNN model simplification methods such as weight pruning and quantization to reduce the computational footprint for inference. Second, we profile the performance of different stages of the PU-GAN model. Table 1 indicates that the feature extraction stage remains the performance bottleneck. We thus also explore reducing the model complexity for feature extraction using, for example, a more efficient spherical kernel for 3D PtCl convolution [1].

Table 1: Profile the running time of the PU-GAN model.

	Feature Extraction	Feature Expansion	Point Set Generation
% Time	78.3%	19.3%	2.4%

Caching and Reusing Inference Results. Regular videos oftentimes exhibit significant similarities across frames. We find that volumetric videos make no exception. As a result, there is no need to run SR for all the points in all frames. We thus propose to cache the inference results and reuse them aggressively. We spatially segment each frame into what we call *3D tiles*, which are the basic units for SR inference and caching. If any tile in the current frame has the same geometric structure to a cached tile, the cached inference can be directly used. To further facilitate reusing cached results, VoluSR allows approximating a tile using a geometrically similar cached tile and a lightweight patch that *delta-encodes* the difference. Note that the computationally intensive tasks of tile segmentation, patch generation, and similarity measurement can all be done offline for on-demand volumetric videos.

Adapting to User's Perception. VoluSR also leverages human users' perception to reduce the computational workload. Specifically, it performs 6-DoF prediction of the user's viewport movement. Based on that, VoluSR only conducts SR for the tiles that (1) fall into the predicted viewport, (2) are not blocked by other tiles, (3) bear a close physical distance to the viewpoint (if a tile is too far, using a high content resolution does not bring much benefit, in terms of visual quality improvement), and (4) have sufficiently high brightness (this can be determined offline if a viewport-independent lighting model is used). Through such viewport adaptation, a large fraction of content can be skipped for SR with a small impact on the user's quality-of-experience (QoE).

Adapting to Devices' Computation Capabilities. VoluSR takes into account the heterogeneity of mobile devices. It assesses a new device's computation capability through one-time profiling, and builds a model that dictates the upsampling time given a tile. At runtime, leveraging this model and according to the available computation resources, VoluSR dynamically adjusts the upsampling ratio. When processing the tiles under a tight resource budget, it also ranks them based on their visual importance (*e.g.,* a closer tile may take a higher priority than a tile that is far from the viewpoint).

System-level Optimization and Integration. We are working on developing the above components and integrating them into a holistic system. System modules consuming different resources (CPU, GPU, network) will be pipelined to maximize the overall resource efficiency. In our current design, all key runtime logic resides on the client side, but we will also consider offloading certain functions such as viewport adaptation and prediction to the server (edge). We will thoroughly evaluate our prototype using real PtCl videos, real users' viewport traces, and off-the-shelf mobile devices.

REFERENCES

[1] H. Lei, N. Akhtar, and A. Mian. Spherical kernel for efficient graph convolution on 3d point clouds. *arXiv preprint arXiv:1909.09287*, 2019.

[2] R. Li, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng. Pu-gan: a point cloud upsampling adversarial network. In *ICCV*, pages 7203–7212, 2019.

[3] H. Yeo, Y. Jung, J. Kim, J. Shin, and D. Han. Neural adaptive content-aware internet video delivery. In *OSDI*, pages 645–661, 2018.