2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)

2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA) | 979-8-3503-2658-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISCA59077.2024.00097

# GameStreamSR: Enabling Neural-Augmented Game Streaming on Commodity Mobile Platforms

Sandeepa Bhuyan, Ziyu Ying, Mahmut T. Kandemir, Mahanth Gowda and Chita R. Das
*Department of Computer Science and Engineering, The Pennsylvania State University, State College, USA*
{sxb392, zjy5087, mtk2, mahanth.gowda, cxd12} @psu.edu

***Abstract*—Cloud gaming (also referred to as Game Streaming) is a rapidly emerging application that is changing the way people enjoy video games. However, if the user demands a high-resolution (e.g., 2K or 4K) stream, the game frames require high bandwidth and the stream often suffers from a significant number of frame drops due to network congestion degrading the Quality of Experience (QoE). Recently, the DNN-based Super Resolution (SR) technique has gained prominence as a practical alternative for streaming low-resolution frames and upscaling them at the client for enhanced video quality. However, performing such DNN-based tasks on resource-constrained and battery-operated mobile platforms is very expensive and also fails to meet the real-time requirement (60 frames per second (FPS)). Unlike traditional video streaming, where the frames can be downloaded and buffered, and then upscaled by their playback turn, Game Streaming is real-time and interactive, where the frames are generated on the fly and cannot tolerate high latency/lags for frame upscaling. Thus, state-of-the-art (SOTA) DNN-based SR cannot satisfy the mobile Game Streaming requirements.**

**Towards this, we propose *GameStreamSR*, a framework for enabling real-time Super Resolution for Game Streaming applications on mobile platforms. We take visual perception nature into consideration and propose to only apply DNN-based SR to the regions with high visual importance and upscale the remaining regions using traditional solutions such as bilinear interpolation. Especially, we leverage the depth data from the game rendering pipeline to intelligently localize the important regions, called regions of importance (RoI), in the rendered game frames. Our evaluation of ten popular games on commodity mobile platforms shows that our proposal can enable real-time (60 FPS) neurally-augmented SR. Our design achieves a 13× frame rate speedup (and ≈4× Motion-to-Photon latency improvement) for the reference frames and a 1.6× frame rate speedup for the non-reference frames, which translates to, on average 2× FPS performance improvement and 26-33% energy savings over the SOTA DNN-based SR execution, while achieving about 2dB PSNR gain and better perceptual quality than the current SOTA.**

***Index Terms*—games, mobile computing, interactive systems, real-time systems, client/server systems, emerging technologies**

## I. Introduction

Video games have evolved tremendously since their inception. Advances in computer graphics techniques and graphics hardware have elevated video games from their humble beginnings as a dot on an oscilloscope [12] to highly immersive and realistic experiences of the present day. Currently, with a worldwide user base of approximately 3.2 billion [98], the video game market is a rapidly expanding business, with the mobile gaming revenue accounting for the majority of the global gaming market in 2022 [99]. The advent of 5G era as well as the recently emergent cloud gaming[1] services such as Microsoft xCloud [66], Sony PlayStation Now [83], NVIDIA GeForce Now [71], etc., encompassing a wide collection of games, which offer Games-as-a-Service (GaaS) [69], have made video games more accessible. Especially, cloud gaming services have enabled experiencing graphics-rich desktop and console games on mobiles, which was previously not possible owing to power constraints and resource limitations of mobile SoCs for supporting high-fidelity games.

However, cloud gaming on mobile platforms, demands a high network bandwidth for streaming high-resolution game frames (2K or 4K) and often suffers from significant frame drops even when utilizing a 5G or WiFi network [8]. To mitigate this, the emerging technique of *Super Resolution* (SR), which involves upscaling low-resolution frames to a higher resolution, can be leveraged by streaming low-resolution frames and upscaling them at the client. A significant number of SR techniques from both industry (NVIDIA DLSS [70], Intel XeSS [44], Topaz Video AI [88], etc.) as well as academia ( [9], [17], [22], [54], [95]) showcase the **impressive quality improvements** of DNN-based SR models.

Unfortunately, due to their compute-heavy nature, these models are mostly restricted to desktops/servers and lack widespread adoption on mobile platforms. Even with the acceleration techniques [43], [51], [100], [108] or hardware accelerators (TPU and NPU), they are yet to meet 60 FPS frame rate for producing high-resolution frames.

Towards achieving real-time Super Resolution on mobile platforms, NEMO [100] leverages the temporal similarity in video frames by reusing the high-quality super-resoluted reference frame to reconstruct the subsequent non-reference frames, thus amortizing SR processing over a video stream. While this works for video streaming, where the frames are already available and hence can be downloaded and buffered, such a technique is not amenable for Game Streaming owing to its real-time and interactive nature, which requires on the fly game frame generation based on user input. Also, due to the compute-intensive nature of DNN-based SR, this technique incurs a significant lag (> 200 ms) during the reference frame upscaling, thus violating the stringent real-time requirements (16.66 ms) for games, which could even result in “Game over” scenarios in multiplayer games. Observing this, we want

[1]We use the terms “cloud gaming” and “game streaming” interchangeably throughout this paper.