GameIR: The 1st Challenge on Image Restoration over Gaming Content

Wei Jiang[1], Lebin Zhou[2], and Jinwei Gu[3]

[1] Futurewei Technologies Inc. (USA)
[2] Santa Clara University (USA)
[3] The Chinese University of Hong Kong (HK)

1 Overview and Summary

Modern cloud gaming has become increasingly popular with an expected global market share value reaching over $12 billion by 2025. Generative AI (GAI) technologies are further transforming the gaming industry by enabling fast and accessible high-quality content creation, empowering immeasurable market growth.

Cloud gaming poses tremendous challenges for data compression and transmission. Most current solutions rely on heavy server-side computation and network delivery, where the client device is merely used for display. To avoid input delay and over-consuming bandwidth, high-quality frames need to be heavily compressed. Traditional codecs like H.265/H.266 or recent neural video coding targeting natural videos cannot resolve this transmission bottleneck.

Generative methods like GAN, VQVAE, and diffusion models, when applied to super-resolution and image rendering and synthesis, can largely alleviate the transmission issues. Server-side computation and transmission can be reduced by leveraging the computation power of client devices. For example, the server can render low-resolution (LR) frames to transfer, and high-resolution (HR) frames can be computed on the client side. In multiview (e.g., immersive VR) gaming, the server can render part of the frames or views to transfer, and the remaining frames or views can be computed by client devices. NVIDIA's Deep Learning Super Sampling (DLSS) technology has commercialized this idea, demonstrating the great potential of optimizing the gaming experience by leveraging bandwidth conditions and computation power of client devices.

The key factor of the success of DLSS is the large-scale ground-truth LR-HR paired data or multiview gaming data used for training that matches the test scenarios. In comparison, the research community uses pseudo training data for many restoration tasks. For instance, for super-resolution, the LR data is generated from the HR data by downsampling and adding degradation like noises, blurs, and compression artifacts. Such pseudo training data does not match the real gaming data. For example, true LR gaming frames are high-quality, sharp, and clear without noises or blurs, different from generated pseudo LR data. Also, there are unnatural visual effects and object movements, but with little motion blur, different from captured natural videos. As a result, we have to resort to ground-truth gaming data for effective training.

In this competition, we provide a large-scale, high-quality, computer-synthesized ground-truth dataset, called GameIR. Our hope is to bring the success of commercial-level DLSS to the public research community so that AI-empowered cloud gaming solutions using image restoration techniques can be more effectively investigated in the field. The GameIR dataset was collected using CARLA, an autonomous driving simulator developed based on the UE4 game engine. There are 8 towns, each having a distinct style and environment, including various simulation entities such as weather, roads, buildings, vehicles, pedestrians, and vegetation. Fig. 1 gives example views of these towns. For each town, we collected two types of scenes: static scenes where there were no other moving vehicles; and dynamic scenes, where there were other moving vehicles.

(a) Town_01 **(b)** Town_02 **(c)** Town_03 **(d)** Town_04

(e) Town_05 **(f)** Town_06 **(g)** Town_07 **(h)** Town_08

Fig. 1: A representative view of 8 different towns.

1.1 Track 1: Super Resolution in Deferred Rendering

Track 1 of this challenge aims to restore HR images from LR images along with additional GBuffers during the deferred rendering stage, supporting the gaming solution of rendering and transferring LR images with assistive information using reduced bits and then restoring HR images on the client side. This track uses the GameIR-SR dataset, collected by placing one HR cameras and one LR cameras at the front of the agent vehicle, capturing synchronized RGB images with 1920x1440 and 960x720 resolution, respectively. Each video is 2-second long at 30fps. During capture, the GBuffer data (*i.e.*, segmentation maps and depth maps) from the deferred rendering phase, as well as the cameras's intrinsic parameters and extrinsic 6-DoF parameters, were also collected synchronously. Fig. 2 gives an example of the GameIR-SR dataset. Finally, GameIR-SR has 19200 LR-HR paired ground-truth frames from 320 LR-HR paired videos, along with the corresponding GBuffers and camera parameters. The ground-truth LR frames are clear and sharp, which can better serve as training data for super-resolution methods targeting at gaming content, where the real degradation features can be better learned.

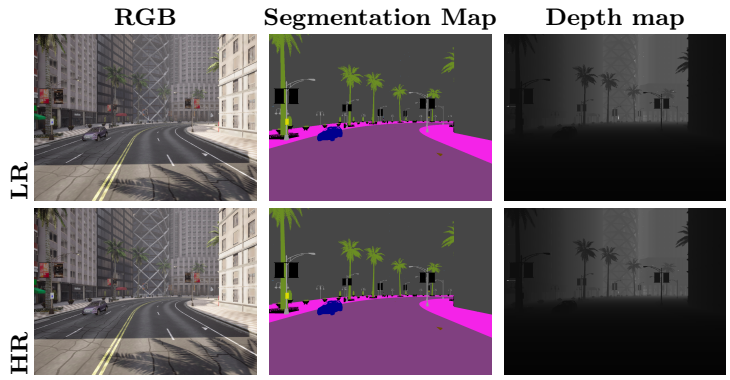

Fig. 2: An example of GameIR-SR dataset

1.2 Track 2: Novel View Synthesis

Track 2 of this challenge aims to synthesize intermediate frames from a sparse set of input frames in multiview videos, along with camera intrinsic and extrinsic parameters and additional segmentation maps and depth maps, supporting the multiview gaming solution of rendering and transferring part of multiview frames with assistive information using reduced bits and then generating the remaining frames on client side. For novel view synthesis (NVS), this track uses the GameIR-NVS dataset, which contains only static autonomous driving scenes. We randomly selected 20 spawn points in each of the 8 towns and recorded 160 scenes in total. To provide ground truth for NVS, we placed 6 sets of cameras in 6 directions around the agent vehicle: front view, left 60° view, right 60° view, left 120° view, right 120° view, and back view. Each set captured the RGB images, semantic segmentation maps, and depth maps at the resolution of 1920x1440 when the vehicle drove through different parts of the towns. Adjacent cameras have some overlapping field-of-view. For each scene, the video is 2-second long at 30fpsframes. Fig. 4 gives an example of the GameIR-NVS dataset. The camera intrinsic parameters and the 6-DoF camera extrinsic parameters for each frame are also recorded. Finally, the GameIR-NVS dataset comprises 960 videos from 160 scenes, totaling 57,600 HR frames. These 360-degree scene-level synthetic data are suitable for training and evaluating NVS methods over gaming content.

2 Challenge and Paper Submissions

Please see the competition website: https://richmediagai.github.io/#challenges for details about data download, important dates, how to participate, *etc.*

This GameIR challenge is hosted alongside with the 2nd Workshop on Rich Media with Generative AI (RichMediaGAI) at ACCV 2024. The winners will be announced at the workshop, and the top 3 non-corporate winners of each track will be rewarded a prize. The winners are invited to submit a paper to the RichMediaGAI workshop through the paper submission system (please

see https://richmediagai.github.io/#papers for details). For the paper to be accepted, each paper must be a self-contained description of the method, and be detailed enough to reproduce the results. This would typically mean at least 8 pages (not including references) following the regular paper format of ACCV 2024 Author Guidelines.

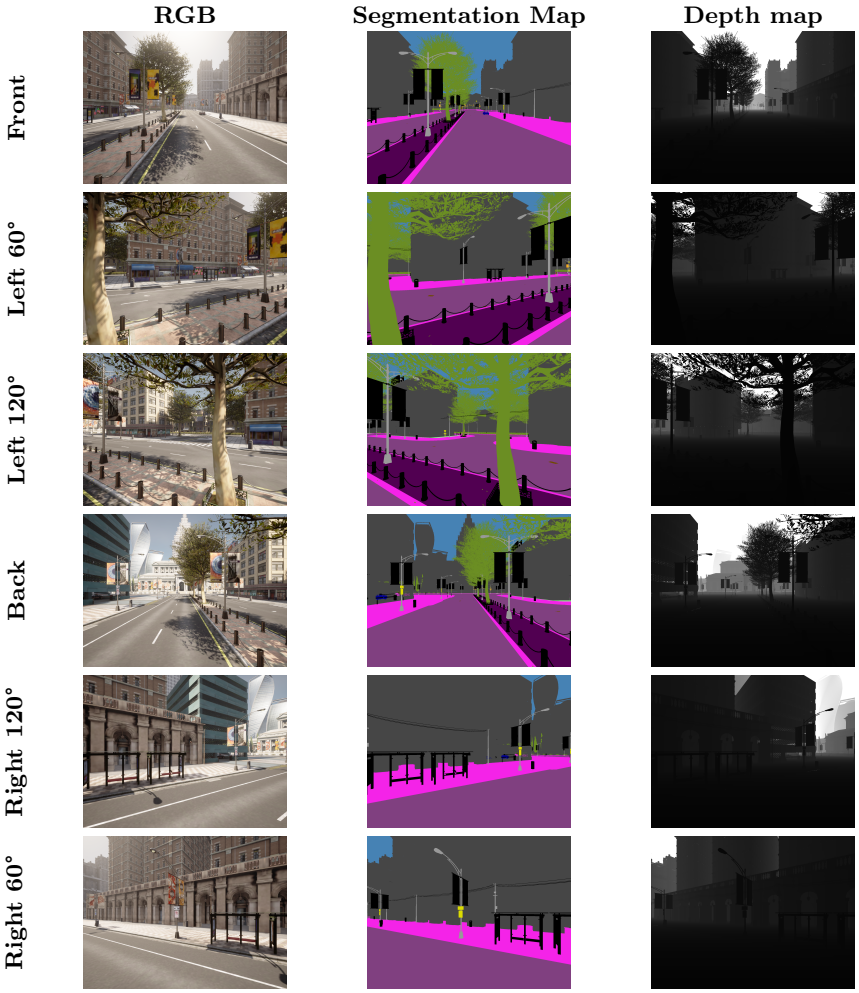

Fig. 3: An example of our GameIR-NVS dataset.