Hybrid Guided AI Frame Interpolation using lightweight intermediate renders

The Idea to help frame generation AI to create a fake Image, by letting it not just look at the previous and following picture, but a third one, wich is traditionally rendered. This third Image only contains the renderd textures, but no other information like shadows or pbr materials. For the GPU it is an easy task to just render the textures, but for an AI this is key information, to know, how the interpolated Image should look like. The goal is, to reduce artifacts, ghosting, motion blur and temporal inconsistencies and even additional frame time.

The Problem with todays AI frame generation is ghosting, wrong pixel movement and other issues from AI based frame generation, wich leads to less precise images and a bad user experience. This frame generation approach coud be usefull for modern gaming and VR, where quality and perfomance matters a lot.

The Idea is to feed the AI with more crucial Information (the unlit render), wich would look imprecise and be performance intensive if the AI has to create this. But this is very cheap and more exakt, if it‘s be rendered by the GPU.

Nvidia with DLSS and AMD with FSR 3 are working on similar technologies, providing the AI with more information like Motion vectors, optical flow and depth awarenes. But all of them are getting to their limits if there is fast camera movements or high fidelity geometry in the scene, potencially occluding pixels, wich give information, where some pixels came from or will travel to. One similar approach to mine is proxy geometrie / low detail reprojection from Nvidia. It reprojects coarse geometrie from the G Buffer to assist the AI. But no approach collects data from the actual moment, when the AI has to insert its image.

The rendering pipeline gets changed, so that every second Frame is a simple frame, without any ligthing. Every frame is an input to the ai and the task is to add the lighting from the surrounding frames to the simple frame. What the AI now should do, is not to create a whole new image out of nothing, wich is very performance heavy, but just editing the pixels hue, brightness and saturation, based on the other two images. These do contain

the detailed effects, like all types of shadows, specular reflection, refection, refraction and other heavy shaders.

The advantages are, that we have the most coarsed but important information of the AI image already renderd, such as sharp textures and precise motion. Especially motion is the problem with todays frame generation solutions, as these frames don’t contain new user input. This approach solved that, by just rendering a new image with these inputs. As the base image is rendered, there is no way, the AI can create ghosting or artifacts. It just needs to add the special effects (wich don’t need to be so precise) from the other images to the simple image and it is done.
Ghosting and artifacts are not a thing to worry about, as the AI no longer has to guess object‘s position. Temporal incoherence is no longer a thing, as the AI‘s Image contains user input and real motions (no approximations based on motion vectors). There is no motion blur, because all the textures visible on screen are directly rendered. To top it all, this solution is more performant than other solutions, as the simple image rendering can be parallelized, and the AI has less work, as it just needs to sometimes adjust pixels a little and not create a whole new image.

To sum it up, this is a new approach to improve AI based frame generation. Quality and performance should in theory be better, as crucial image information is continued to be rendered traditionally and other just visually appealing effects are created and overlayed by an interpolation AI.

This is just an Idea of how future frame generation by AI could look like. I didn‘t test this, because I don‘t have the knowlege to implement this well enough and compare performance and quality against other approaches. Maybe there is someone interested in implementing this and also can try different versions of the third image, with more, less or different additional data (normals, depth, motion vectors, etc.)