# NVIDIA ADA SCIENCE

***How Ada advances the science of graphics with DLSS 3.***

V1.0

# Table of Contents



# List of Figures

# Introduction

Technology is more than mere speeds and feeds. What you do with technology, and the novel applications and innovations it leads to, is what we call science. Science lets us invent things the world has never seen, and it has led to the concept of neural graphics. In this whitepaper, the science of GPUs is illustrated in DLSS—how innovative hardware acceleration techniques have enabled the use of artificial intelligence for game performance.

In late 1999, NVIDIA introduced GeForce 256, the world's first GPU, which included powerful hardware-based transform and lighting engines. A few years later, the programmable shading era was born with the GeForce 3 GPU introducing programmable pixel and vertex shaders. Programmable shading GPUs revolutionized 3D and made possible the beautiful graphics we see in games today. Nearly 20 years after our invention of the GPU, we launched NVIDIA RTX—a new architecture with dedicated processing cores that enabled real-time ray tracing and accelerated artificial intelligence algorithms and applications. RTX—the fusion of programmable shading, ray tracing, and AI—started the next era of graphics: neural graphics.

Neural graphics has quickly become integral to how artists create 3D worlds, and how the world is animated and rendered. One of the most impactful neural graphics inventions is Deep Learning Super Sampling (DLSS)—an AI that has learned to enhance the resolution of motion graphics and improve performance of ray-traced games and creator applications.

## The Era of Neural Graphics

The commitment NVIDIA has made to developing neural graphics strategies in ray tracing and DLSS helps counter the dwindling influence of Moore's Law—the computer science axiom stating that the number of transistors in a dense integrated circuit, like a GPU semiconductor chip, doubles about every two years. The era of Moore's Law is over. As it becomes much harder to wring improvements from hardware, performance strategies from software and algorithms become more important. Neural graphics technologies like DLSS are critical for future performance improvements as AI algorithms leverage dedicated, specialized hardware GPU cores for acceleration.

The following chart shows the rapid efficacy of real-time NVIDIA ray tracing since its debut four years ago during the launch of the Turing architecture. Our investment in dedicated RT cores and neural graphics principles has led to a sharp increase in the number of ray tracing operations per pixel—a 16x improvement, far surpassing the roughly 4x GPU transistor count increase between Turing and Ada GPUs.

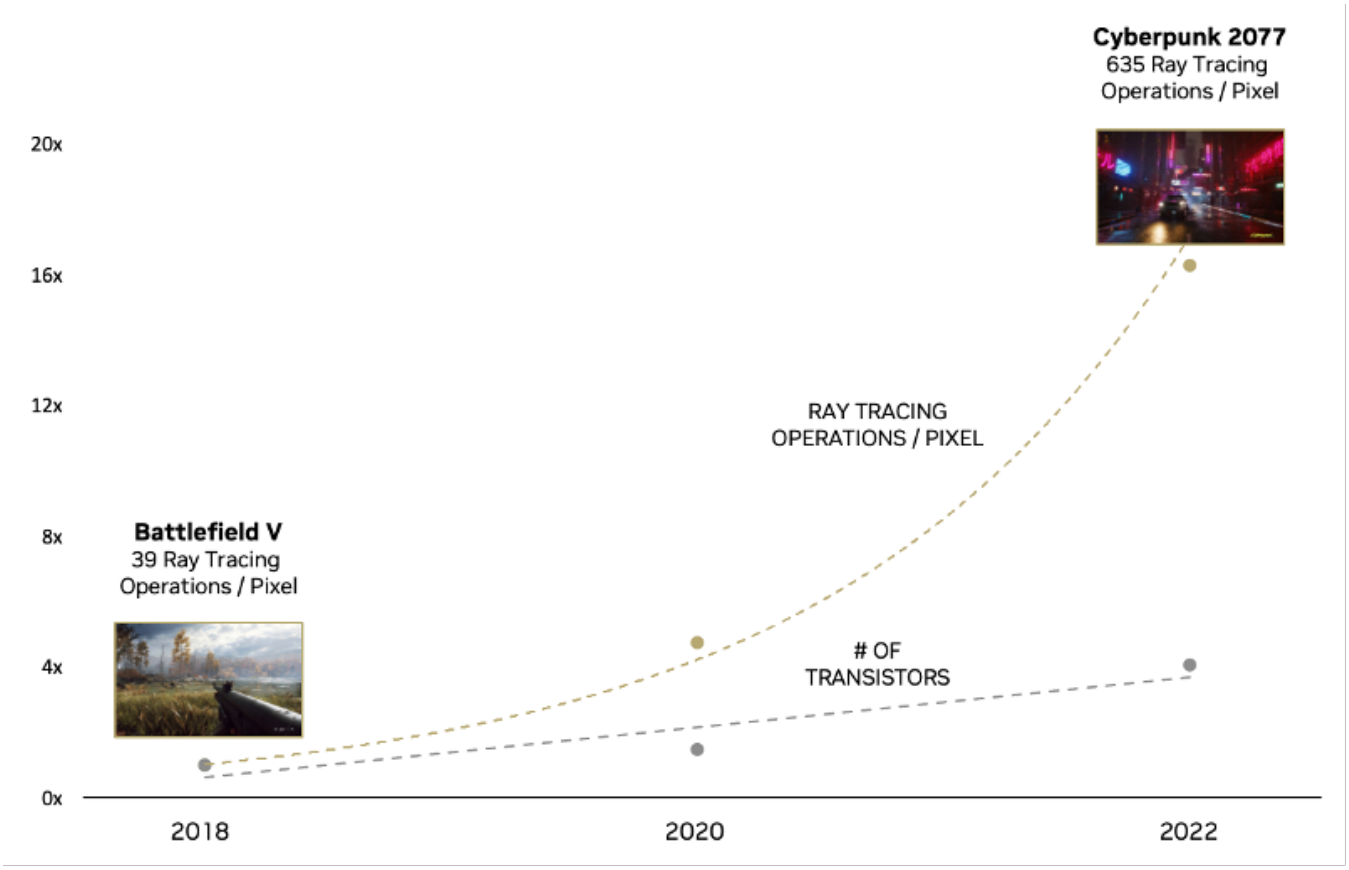
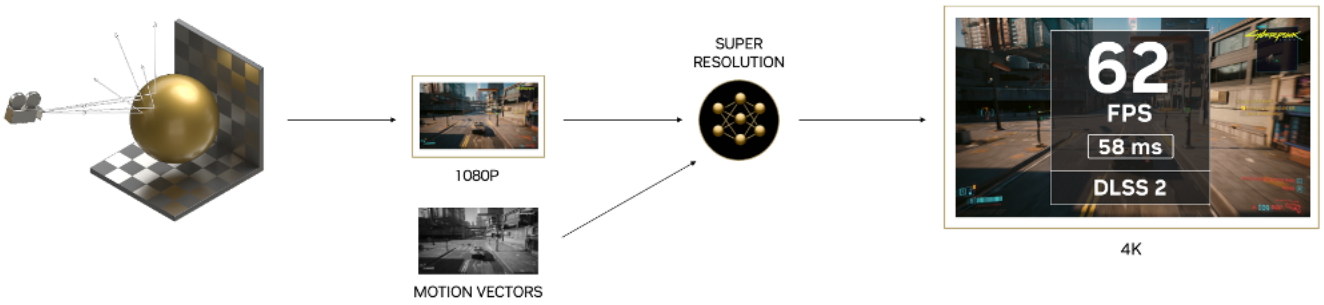



Figure 1. Ray Tracing Operations Per Pixel Over Time

## The Evolution of DLSS

The NVIDIA Turing RTX architecture, first launched in 2018 with the GeForce RTX 2080 Ti, redefined the GPU by adding pioneering real-time ray tracing and AI capabilities. The combination of sophisticated new real-time ray tracing algorithms and real-time AI-powered DLSS brought unprecedented levels of visual fidelity and performance to gamers and creators. DLSS was our breakthrough in AI-powered graphics that massively boosts performance. It does this by using AI to output higher resolution frames from a lower resolution input.

At NVIDIA we are constantly training the DLSS AI model, making full use of our supercomputer. The DLSS super-resolution model is a convolutional autoencoder AI model that takes the low-resolution current frame and the high-resolution previous frame and predicts on a pixel-by-pixel basis a higher resolution current frame. The AI model is trained to predict an ultra-high resolution 16K rendered reference image. The difference between the predicted and reference image is used to train the neural network, and the process is repeated tens of thousands of times until the network can predict a high-quality image. The latest models are delivered to your GeForce RTX PC through NVIDIA Game Ready Drivers. Tensor Cores on the NVIDIA GPU then use their teraflops of dedicated AI horsepower to run the DLSS AI network in real-time.




DLSS analyzes the current and past frame, along with game motion vectors that describe geometry movement, and its deep learning algorithm provides a new supersampled image. This higher resolution output is based on a lower resolution input, and dramatically improves game performance.

Figure 2. DLSS Super Resolution Pipeline

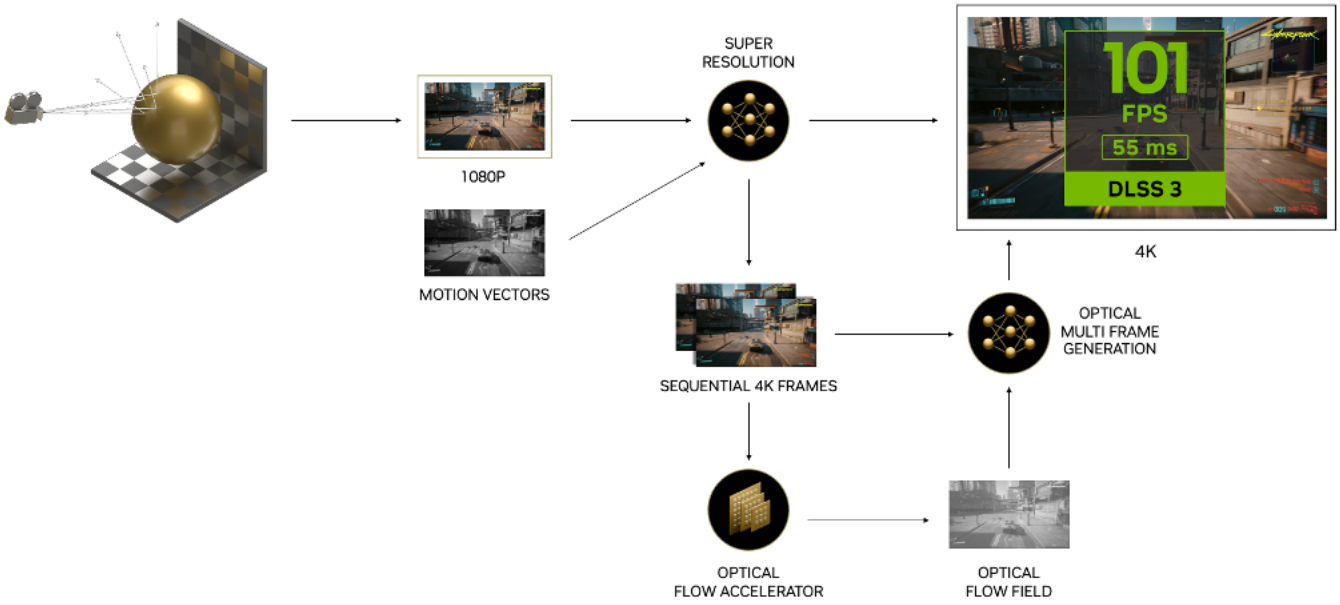
Later, an updated DLSS 2 provided unmatched image quality and performance for game rendering, with up to 2x higher FPS, and image quality comparable to native rendering. DLSS 2 has been adopted into more than 200 games, with plugins in Unreal Engine and Unity, and continues to learn, with new updates continually improving image quality for challenging content such as thin geometry, disoccluded motion, moving objects, and particle effects.

It is now two-and-a-half years later and competing image reconstruction techniques still have not matched the balance of FPS and image quality that DLSS provides. NVIDIA engineers drive continuous progress on the most challenging graphics problems to make games look better and run faster, all by training the DLSS AI to be smarter at the most important problems in image reconstruction.

## DLSS 3 Multiplies Performance by Up to 4X

Alongside the launch of Ada architecture, we are introducing DLSS 3, which takes DLSS to the next level with an innovative capability made possible only by the judicious application of deep learning—intelligent frame generation. DLSS 3 uses AI to generate whole frames in order to massively boost performance.

Over the past four years, the NVIDIA Research team has been developing a frame generation technique that combines AI-powered optical flow estimation with DLSS to improve frame rates and the gaming experience. As a component of DLSS 3, the Optical Multi Frame Generation convolutional autoencoder takes four inputs—current and prior game frames, an optical flow field generated by Ada's Optical Flow Accelerator, and game engine data such as motion vectors and depth. Optical Multi Frame Generation then compares a newly rendered frame to the prior rendered frame, along with motion vectors and optical flow field information to understand how the scene is changing, and from this generates an entirely new, high-quality frame in between each DLSS Super Resolution frame. These generated frames are interleaved between the standard game-rendered frames, enhancing motion fluidity just as any highly performant frame rate does.




DLSS 3 combines previous DLSS techniques, such as motion vectors and Super Resolution, with new Frame Generation, leveraging the Ada architecture's Optical Flow Accelerator.

Figure 3. DLSS 3 With Frame Generation Pipeline

DLSS 3 is a full stack innovation, using a host of software and hardware solutions together in order to deliver up to 4x game performance and up to 2x responsiveness. On the software side, DLSS 3 is three pieces of software: Super Resolution intelligently reconstructs high resolution supersampled images from low resolution inputs, Frame Generation uses AI to generate entirely new frames, and NVIDIA Reflex is integrated to reduce input latency and increase responsiveness—these three DLSS 3 technologies all improve game and player performance, by increasing FPS and lowering input latency. However, these technologies would not be possible without hardware: 4th Gen Tensor Cores run the DLSS algorithm in real time, the newly improved Ada Optical Flow Accelerator helps DLSS measure the scene on a pixel basis, and the NVIDIA supercomputers continue to train and improve image quality and feed advancements through Game Ready Drivers to your GPU. This total gaming performance is a marriage of software and hardware, of AI and dedicated, specialized GPU cores, and is only possible on NVIDIA Ada architecture GPUs.

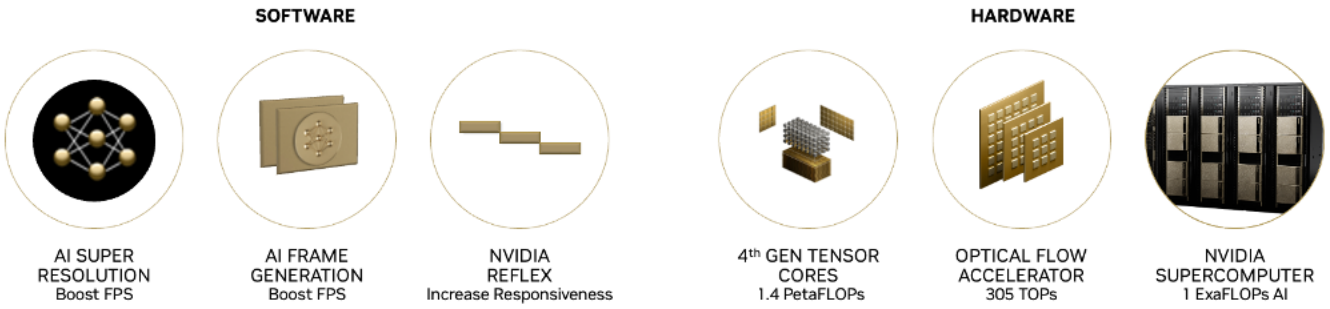



Figure 4. DLSS Software and Hardware Components

The quality of DLSS 3 would have been impossible without a full stack solution. While deep learning algorithms are always learning and improving, they rely on the hardware used to train

and analyze images. Some of the hardest challenges in frame generation require hardware, such as the Optical Flow Accelerator, to overcome. The Optical Flow Accelerator is able to capture pixel-level information such as particles, reflections, shadows, and lighting, which are not included in game engine motion vector calculations. Post processing effects and ray tracing effects are difficult to reconstruct accurately because they lack the same type of motion vectors associated with geometry in a game. The standard game engine motion vectors—used in temporal anti-aliasing algorithms to blend pixels from the prior frame with the current frame—may account for the positioning of moving geometry, but they do not help determine the position of many types of post processing effects, and this causes artifacts like ghosting, stutter, and blurriness.

For example, the following game communicates to DLSS the motion vectors for geometry, such as the street, but has no motion vectors to pass along for the ray-traced shadow that is also moving while cast against the street. Without motion vectors for ray tracing effects, the shadow would be reconstructed inaccurately, causing stuttering or ghosting as the motion vectors of the street drag the shadow in the wrong direction.

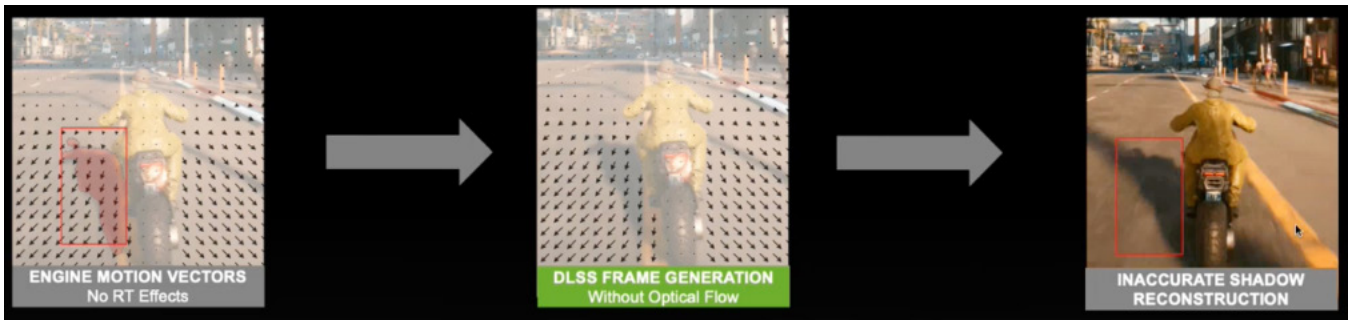
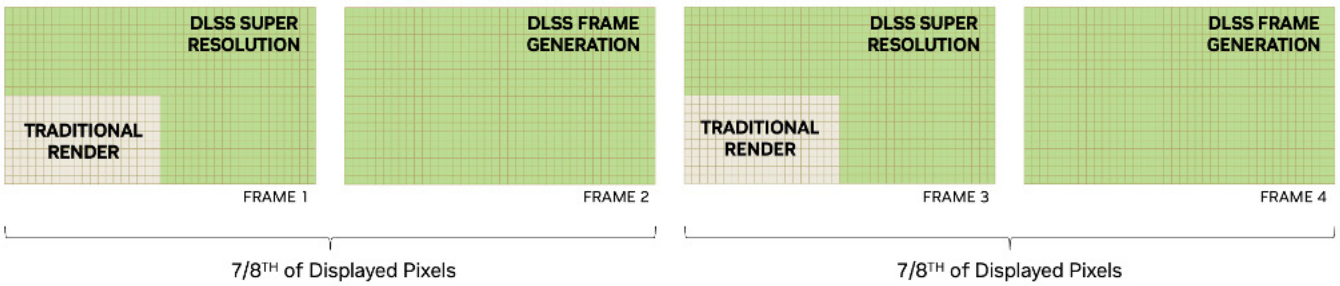



Figure 5. Engine Motion Vectors and Optical Flow

DLSS 3 overcomes this problem by running optical flow calculations using the updated Ada architecture Optical Flow Accelerator, which significantly speeds up optical flow processing compared to prior generation OFA units. By analyzing the optical flow field, DLSS 3 can calculate how everything in the scene is moving from pixel to pixel, including effects like shadows, reflections, and particles. Quality frame interpolation has long been a goal in computer graphics, but without custom hardware acceleration to assist analysis in the absence of motion vectors, the algorithm would fail to discern that the shadow and the street are independent of each other, and that the motion of the road is different from the motion of the shadow. DLSS 3 and Ada architecture solve this problem by employing a neural network to decide how to use information provided by the game motion vectors, the data from the optical flow field, and the game input frames to generate accurate frames that also support ray tracing and post processing effects.

DLSS 3 can reconstruct high resolution image detail and entirely new frames even though it has fewer game-rendered pixels to work with. With DLSS 3 enabled, AI is reconstructing three-fourths of the first frame with DLSS Super Resolution, and reconstructing the entire second frame using DLSS Frame Generation. In total, DLSS 3 reconstructs seven-eighths of the total displayed pixels, increasing performance significantly.




DLSS 3 reconstructs 7/8th of the total displayed pixels using AI. Frame 1 uses DLSS Super Resolution to reconstruct a higher resolution frame, and frame 2 then uses DLSS Frame Generation to entirely generate a new frame before resuming DLSS Super Resolution in frame 3, and so on.

Figure 6. AI Pixels Provide Most of the Image

## DLSS 3 Can Double CPU Bound Performance

DLSS 3 also reduces CPU load during high-FPS gaming, helping to transform CPU-bound games and scenarios into GPU-bound ones. CPU boundedness is a problem with some games that happens due to CPU-heavy operations, such as physics simulations, or generally CPU-intensive draw calls such as in games with large open worlds. Essentially, the GPU is idle waiting for instructions due to a CPU performance bottleneck. CPU-bound scenarios are difficult to overcome because the GPU must wait for the CPU to serially issue calls and cannot contribute to help ease the bottleneck. DLSS 3, however, does alleviate this problem.

Firstly, DLSS 3 can transform CPU-bound games into GPU-bound games because the CPU does not need to call the GPU to render every frame. DLSS Frame Generation executes as a post-process on the GPU, allowing the AI network to boost frame rates even when the game is bottlenecked by the CPU. For CPU-limited games, such as those that are physics-heavy or involve large worlds, DLSS 3 allows GeForce RTX 40 Series graphics cards to render at up to twice the frame rate over what the CPU is able to compute.

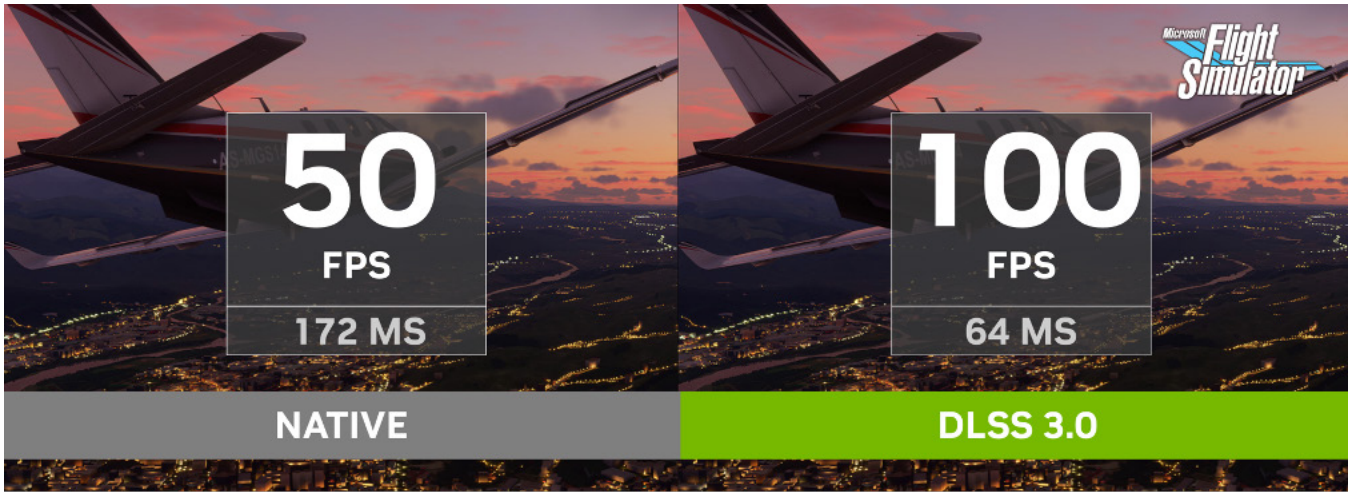
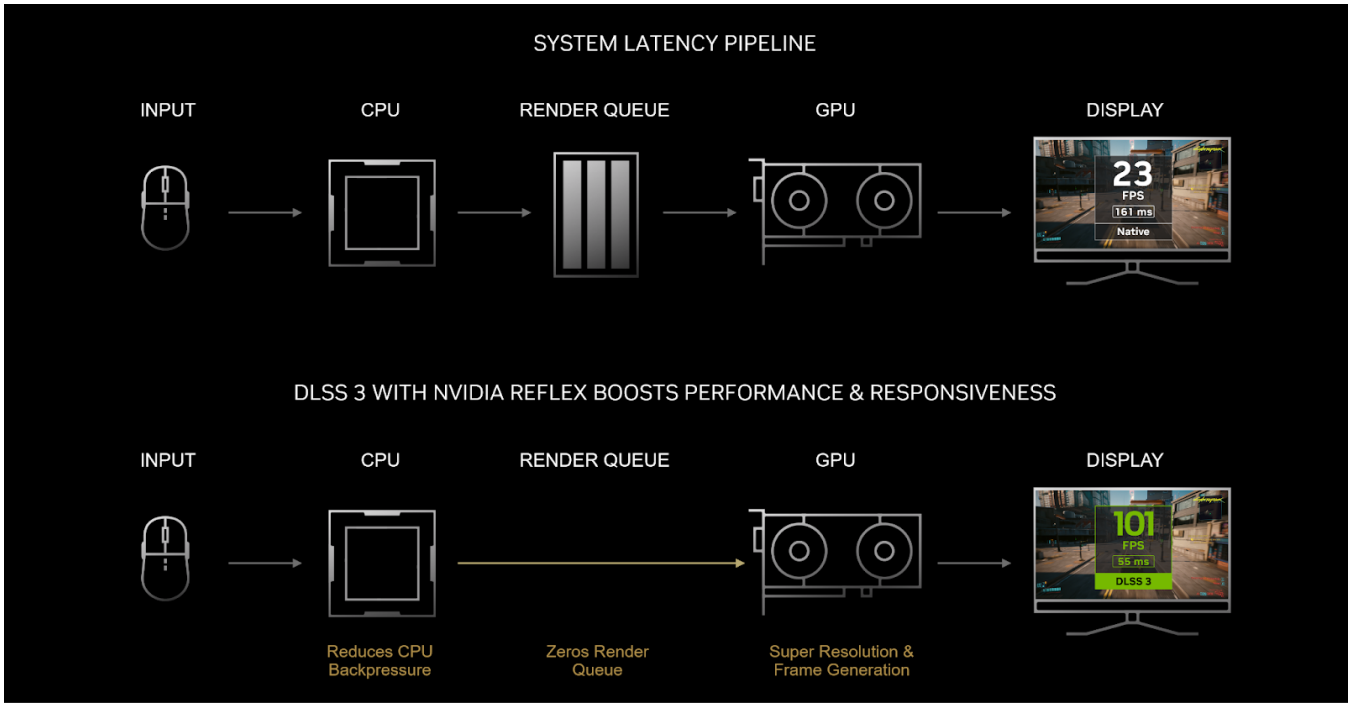



Figure 7. DLSS Can Double CPU Bound Performance

## DLSS 3 and Responsiveness — NVIDIA Reflex

DLSS 3 also incorporates NVIDIA Reflex, which synchronizes the GPU and CPU, ensuring optimum responsiveness and low system latency.




DLSS 3 reconstructs 7/8th of the total displayed pixels using AI. Frame 1 uses DLSS Super Resolution to reconstruct a higher resolution frame, and frame 2 then uses DLSS Frame Generation to entirely generate a new frame before resuming DLSS Super Resolution in frame 3, and so on.

Figure 8. System Latency Pipeline

Traditionally in the graphics pipeline, there is a render queue, which schedules draw calls for the GPU. And to ensure that the GPU is running at full utilization and prevent stalling, this render queue is filled. This means that the CPU has to submit in advance, adding more latency to the graphics pipeline. With NVIDIA Reflex, which is a core part of DLSS 3, we're able to remove the render queue without sacrificing frame rates with just-in-time rendering methods. This optimized pipeline enables the CPU to generate draw calls right before the GPU needs it, removing the need for the render queue. As a result, NVIDIA Reflex is able to remove latency from the overall system. And lower system latency makes game controls more responsive, and ensures on-screen actions occur almost instantaneously once you click your mouse or other control input. When compared to native, DLSS 3 can reduce latency by up to 2X.

## DLSS 3 and the Three Pillars of Performance

Latency is critical for performance. The holistic view of game performance sees three pillars that define the most important facets of how humans feel performance: smoothness, responsiveness, and image quality. Gamers are always assessing performance—the performance of how fast the game is rendering, their own player performance and the feel of the game, and the performance of how the GPU renders quality images. Each facet is crucial alone but when combined, begin to

influence or modulate each other: a fast game with poor image quality is not a great experience, and a game with punishing latency will feel sluggish despite high FPS performance.

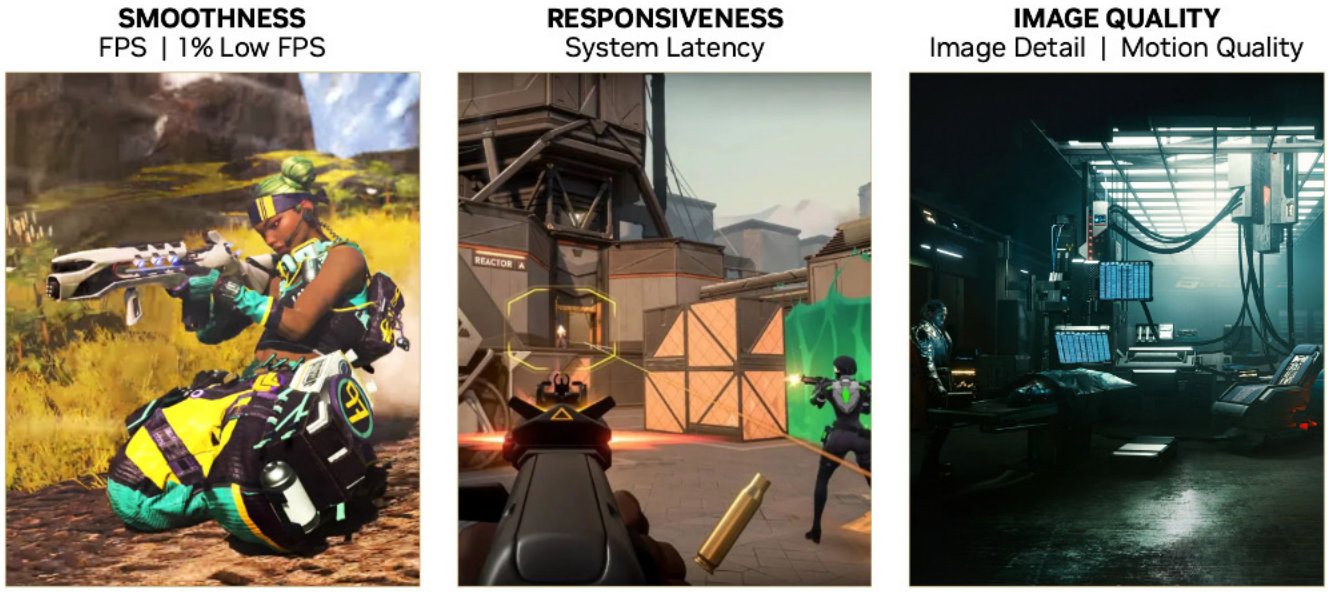



Figure 9. The Three Pillars of Performance

DLSS 3 is dedicated to all aspects of performance, boosting smoothness, responsiveness, and image quality. Smoothness, measured in FPS and 1% Lows, describes the experience of speed and fluidity, where high FPS refreshes the display with more information faster, and the regular delivery of these frames avoids stutter; if FPS is low or irregularly paced, the game does not feel smooth.

Responsiveness is system latency, the number of milliseconds it takes for a player to click the mouse and see input reflected in the game; if latency is too high, the game will feel unresponsive and lead to tangible skill decreases in the player.

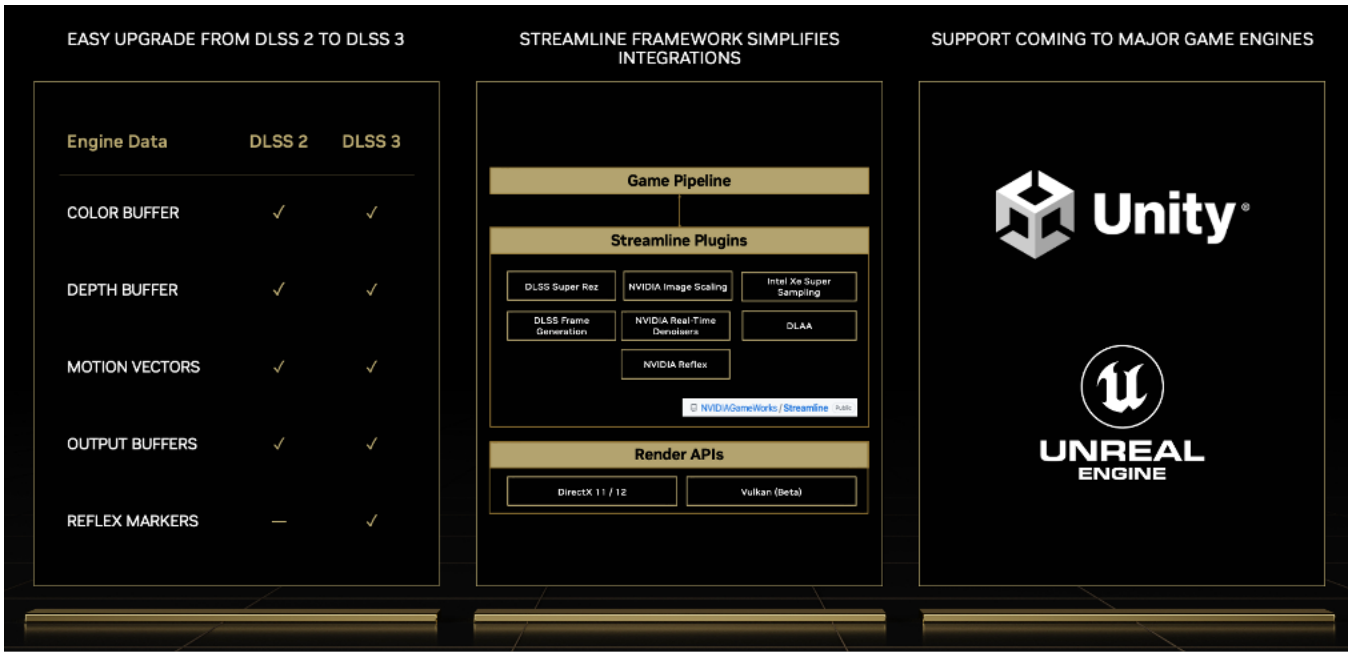
Image quality rounds out the performance picture and assesses the performance of fidelity, how true the GPU is to quality images, while static and in motion.

## DLSS 3 and NVIDIA Streamline

DLSS 3 was built for speedy integration into games. Game developers that have integrated DLSS 2 using the NVIDIA Streamline SDK have already done the work necessary to integrate DLSS 3 and its AI Super Resolution, AI Frame Generation, and NVIDIA Reflex components. There are no multiple code-paths to complicate DLSS versions because Streamline handles all three aspects of DLSS. If the game is played on an Ampere or Turing GPU, and the DLSS 3 game will effectively run DLSS 2 without Frame Generation. This makes DLSS easy for developers to integrate and simple for gamers to enable, no matter what generation the GPU.

NVIDIA Streamline is an open-sourced cross-platform solution that simplifies integration of the latest NVIDIA and other independent hardware vendors' super resolution technologies into applications and games. This framework allows developers to easily implement one single integration and enable multiple super resolution technologies and other graphics effects supported by the hardware vendor. Streamline supports Deep Learning Anti-Aliasing (DLAA),

NVIDIA Image Scaling, NVIDIA Real-Time Denoisers (NRD), NVIDIA Reflex, and Optical Multi Frame Generation and DLSS technologies.




Streamline offers a single integration with a plug-and-play framework, including DLSS 2 and 3.

Figure 10. NVIDIA Streamline and DLSS

DLSS 3 games are backwards compatible with DLSS 2 technology. DLSS 3 technology is supported on GeForce RTX 40 Series GPUs. It includes 3 features: our new Frame Generation tech, Super Resolution (the key innovation of DLSS 2), and Reflex. Developers simply integrate DLSS 3, and DLSS 2 is supported by default. NVIDIA continues to improve DLSS 2 by researching and training the AI for DLSS Super Resolution, and will provide model updates for all GeForce RTX gamers, as we've been doing since the initial release of DLSS.

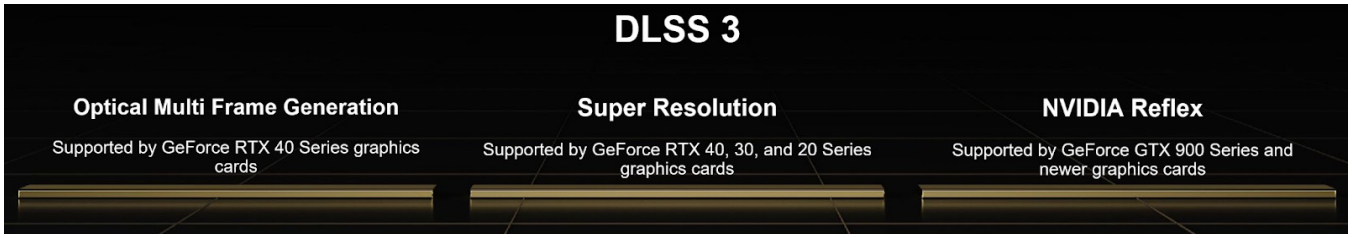



Figure 11. DLSS 3 is Optical Multi Frame Generation, Super Resolution, and NVIDIA Reflex

## NVIDIA DLSS 3: Available this October

DLSS 3 has already received support from many of the world's leading game developers and engines, with more than 35 games and applications announcing support, the first of which launches in October.


Figure 12. DLSS 3 Momentum

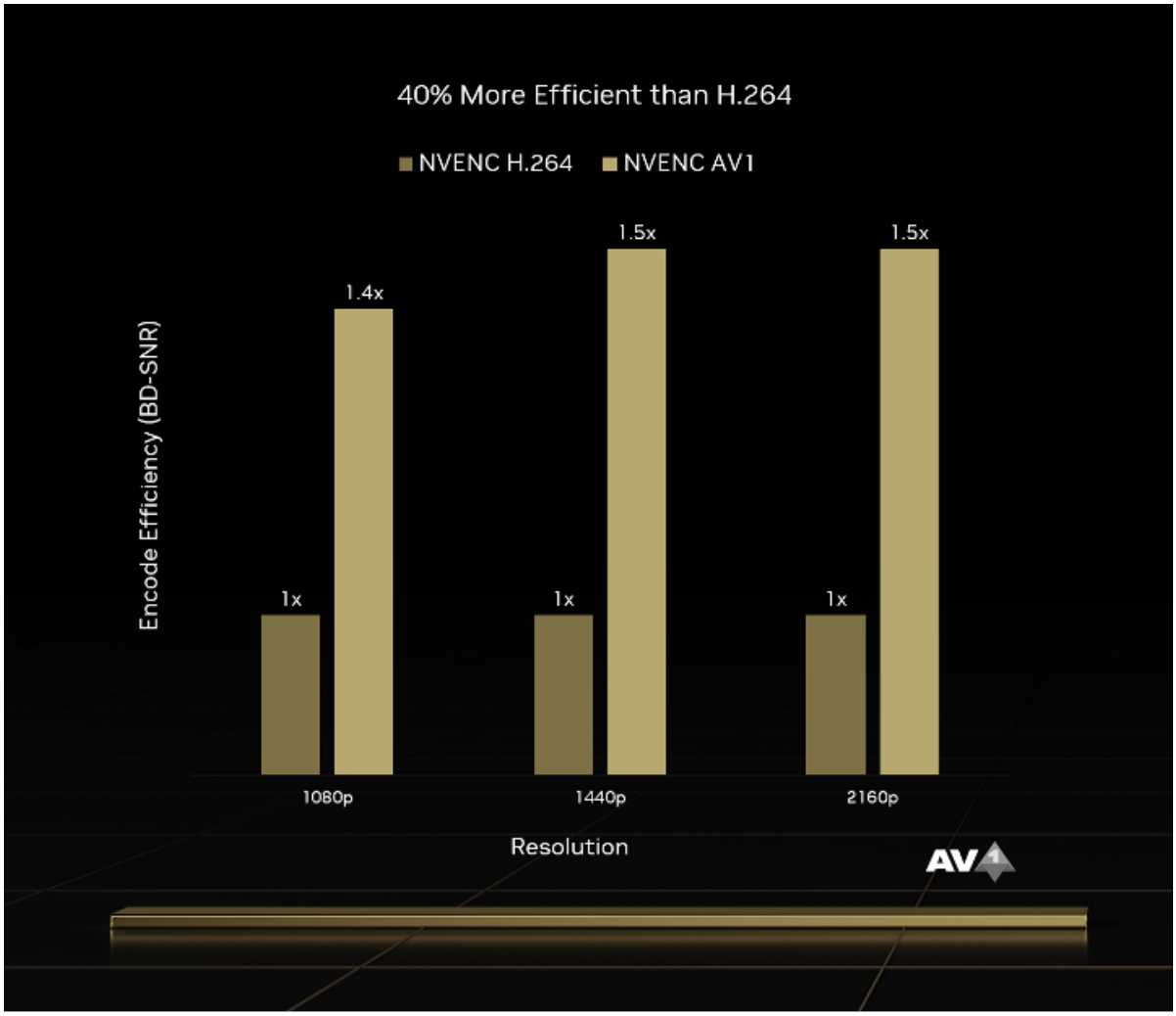
# NVIDIA Encoder and Broadcast

Personal computers now more than ever help us connect with others, whether to chat and share your screen with your friends while you play, videoconference for remote work or studying, communicate with friends and family online, or as a stage for self-expression through livestreaming. NVIDIA is the hardware of choice for live streamers thanks to the hardware encoder available in our GPUs, and our AI technologies to enhance audio and video—available in the NVIDIA Broadcast app and its SDKs.

During the launch of Ampere architecture GPUs, NVIDIA introduced a free suite of AI-enhanced audio and video effects for your microphone, speakers, and webcam called NVIDIA Broadcast—ideal for those wanting to create content, or work and study from home. These Broadcast tools enable live streamers to look and sound more professional without adding additional cost. And thanks to the tensor cores on RTX GPUs, they can be run at the same time as you are gaming.

## AV1 Encoder

The dedicated hardware encoders on our GPUs receive a major update with the 8th Gen NVIDIA Encoder (also known as NVENC), with support for AV1. This new codec developed by AOM—a consortium of the leading video companies in the world, including NVIDIA—massively increases encoding efficiency, allowing users to dramatically improve their video streaming quality. With AV1 support, our new encoder is 40% more efficient at encoding than H.264. This improvement makes streams look much better, as if using 40% more bandwidth. It also unlocks new experiences for end users who face bandwidth limitations. For example, Twitch and Discord limits streams to 8 Mbps, which was not enough to stream 1440p resolution with H.264. Using AV1, streams look amazing at 8 Mbps and 1440p resolution. Incidentally, we have worked with Discord to enable AV1 end-to-end streaming; Discord plans on releasing AV1 streaming to their users later this year.




NVENC AV1 encoder provides massive encoding speedups at all major resolutions compared to the previous H.264 encoder.

## Figure 13. AV1 Encoder Performance

AV1 makes further use of motion vectors to improve compression, making it vastly more demanding than prior codecs, and thus requiring good hardware encoders to run optimally. AV1 also allows for high resolution capture of up to 8K. And finally, AV1 enables more encoding options, such as captures in HDR.

The RTX 4090, RTX 4080 16 GB, and RTX 4080 12 GB will offer added flexibility of two separate encoders. Two AV1 encoders provide powerful options for creators, such as the ability to encode 8K60 HDR in real time, or the ability to parallelize work in video editing apps for faster video exports. To simplify the use of dual encoders, we have added a new split frame encode feature that takes the upper and lower part of each frame and sends them to each encoder, respectively, in order to process them in parallel, while retaining temporal information. The frame information can then be combined on the driver for final encode, generating a single bitstream. This mode is completely automated and offered to apps as part of the fastest profiles (P1 and P2 in our SDK) for resolutions of 4K and above and can reduce export times by up to 2x compared to encoding with Ampere on a single encoder. This new split frame encode feature will be available in October

in the next build of DaVinci Resolve, Jianying—China's top video editing app—and Voukoder, a popular encoder plugin for Adobe Premiere Pro.


Figure 14. AV1 Image Quality

## Broadcast SDK Features

Shortly after the release of the NVIDIA Broadcast app in October 2020, we released the SDKs that power the app for developers to use, hoping to extend these benefits to many more workflows. Since then, the SDKs have been integrated in 15+ apps. These integrations can simplify the use of these technologies, like Logitech G Hub, that adds the effects directly in the audio driver for a seamless experience—or enable new advanced workflows, like OBS Studio, which denoises each audio source independently according to its needs.

With Ada architecture, the Maxine SDK adds three exciting new features: an improved Virtual Background, Face Expression Estimation, and Eye Contact.

The Virtual Background, which replaces the video background behind the subject, has received a generational upgrade: the new version now utilizes a temporal network that uses temporal information to increase the segmentation quality. By using temporal information, the network can better determine if a section of the frame is a person or not, thus avoiding issues with elements popping in and out.

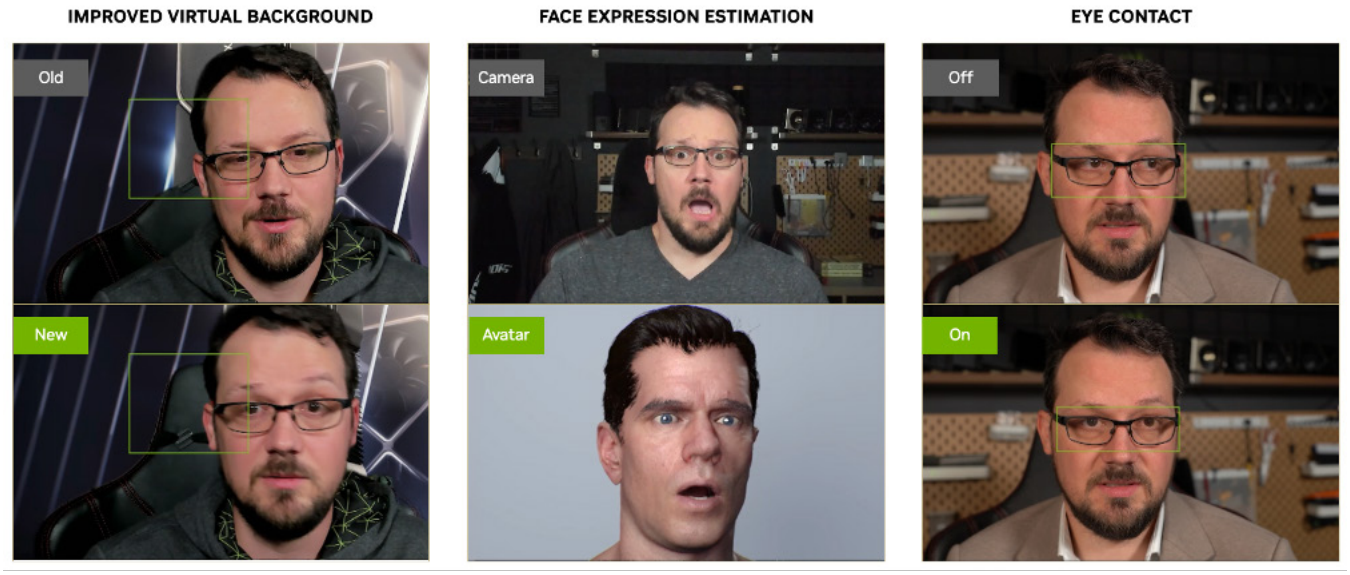



Figure 15. Virtual Background, Face Expression Estimation, and Eye Contact

Face Expression Estimation is a new feature in our Augmented Reality toolkit. The feature tracks face expressions to animate face meshes in digital avatars, providing better transfer of emotions to the avatar. This feature will help developers create more realistic avatars. It will be available in VTube Studio, one of the top VTubing apps, in October.

Eye Contact is a feature that tracks the position of the subject's eyes and morphs their position to directly view the camera, no matter where the eyes might be looking. The effect looks extremely natural because Eye Contact maintains your natural movements, like blinking. Eye Contact helps users chatting over the internet maintain direct eye contact even though the webcam is not positioned in the exact center of the screen. It is also immensely useful for removing eye movements while reading from screens or teleprompters.

All these features are available for our partners, free of charge. They are shipping in more than 10 applications. And, the top device manufacturers are using these effects to enhance their devices, including Logitech, Corsair, and Elgato.

# Conclusion

The science of NVIDIA technology is introducing the next era of rapid performance improvements with help from DLSS 3: using the boundless opportunities of artificial intelligence, paired with dedicated hardware acceleration, to advance performance. The future will continue to see cutting edge technological advancements married to powerful neural graphics concepts, and we will continue investing to invent the hardware and software that powers the neural graphics revolution.