Best GPU for Deep Learning: Considerations for Large-Scale AI

Best GPU for Deep Learning

Traditionally, the training phase of the deep learning pipeline takes the longest to achieve. This is not only a time-consuming process, but an expensive one. The most valuable part of a deep learning pipeline is the human element – data scientists often wait for hours or days for training to complete, which hurts their productivity and the time to bring new models to market.

To significantly reduce training time, you can use deep learning GPUs, which enable you to perform AI computing operations in parallel. When assessing GPUs, you need to consider the ability to interconnect multiple GPUs, the supporting software available, licensing, data parallelism, GPU memory use and performance.

In this article, you will learn:

- The importance of GPUs in deep learning
- How to choose the best GPU for deep learning
- Using consumer GPUs for deep learning
- Best deep learning GPUs for data centers
- DGX for deep learning at scale

Why are GPUs important for Deep Learning?

The longest and most resource intensive phase of most deep learning implementations is the training phase. This phase can be accomplished in a reasonable amount of time for models with smaller numbers of parameters but as your number increases, your training time does as well. This has a dual cost; your resources are occupied for longer and your team is left waiting, wasting valuable time.

Graphical processing units (GPUs) can reduce these costs, enabling you to run models with massive numbers of parameters quickly and efficiently. This is because GPUs enable you to parallelize your training tasks, distributing tasks over clusters of processors and performing compute operations simultaneously.

GPUs are also optimized to perform target tasks, finishing computations faster than non-specialized hardware. These processors enable you to process the same tasks faster and free your CPUs for other tasks. This eliminates bottlenecks created by compute limitations.

How to Choose the Best GPU for Deep Learning?

Selecting the GPUs for your implementation has significant budget and performance implications. You need to select GPUs that can support your project in the long run and have the ability to scale through integration and clustering. For large-scale projects, this means selecting production-grade or data center GPUs.

GPU Factors to Consider

These factors affect the scalability and ease of use of the GPUs you choose.

Ability to interconnect GPUs

When choosing a GPU, you need to consider which units can be interconnected. Interconnecting GPUs is directly tied to the scalability of your implementation and the ability to use multi-GPU and distributed training strategies.

Typically, consumer GPUs do not support interconnection (NVlink for GPU interconnects within a server, and Infiniband/RoCE for linking GPUs across servers) and NVIDIA has removed interconnections on GPUs below RTX 2080.

Supporting software

NVIDIA GPUs are the best supported in terms of machine learning libraries and integration with common frameworks, such as PyTorch or TensorFlow. The NVIDIA CUDA toolkit includes GPU-accelerated libraries, a C and C++ compiler and runtime, and optimization and debugging tools. It enables you to get started right away without worrying about building custom integrations.

Learn more in our guides about PyTorch GPUs, and NVIDIA deep learning GPUs.

Licensing

Another factor to consider is NVIDIA's guidance regarding the use of certain chips in data centers. As of a licensing update in 2018, there may be restrictions on use of CUDA software with consumer GPUs in a data center. This may require organizations to transition to production-grade GPUs.

3 Algorithm Factors Affecting GPU Use

In our experience helping organizations optimize large-scale deep learning workloads, the following are the three key factors you should consider when scaling up your algorithm across multiple GPUs.

1. Data parallelism - Consider how much data your algorithms need to process. If datasets are going to be large, invest in GPUs capable of performing multi-GPU training efficiently. For very large scale datasets, make sure that servers can communicate very fast with each other and with storage components, using technology like Infiniband/RoCE, to enable efficient distributed training.
2. Memory use - Are you going to deal with large data inputs to model? For example, models processing medical images or long videos have very large training sets, so you'd want to invest in GPUs with relatively large memory. By contrast, tabular data such as text inputs for NLP models are typically small, and you can make do with less GPU memory.
3. Performance of the GPU - Consider if you're going to use GPUs for debugging and development. In this case you won't need the most powerful GPUs. For tuning models in long runs, you need strong GPUs to accelerate training time, to avoid waiting hours or days for models to run.

Using Consumer GPUs for Deep Learning

While consumer GPUs are not suitable for large-scale deep learning projects, these processors can provide a good entry point for deep learning. Consumer GPUs can also be a cheaper supplement for less complex tasks, such as model planning or low-level testing. However, as you scale up, you'll want to consider data center grade GPUs and high-end deep learning systems like NVIDIA's DGX series (learn more in the following sections).

In particular, the Titan V has been shown to provide performance similar to datacenter-grade GPUs when it comes to Word RNNs. Additionally, its performance for CNNs is only slightly below higher tier options. The Titan RTX and RTX 2080 Ti aren't far behind.

NVIDIA Titan V

The Titan V is a PC GPU that was designed for use by scientists and researchers. It is based on NVIDIA's Volta technology and includes Tensor Cores. The Titan V comes in Standard and CEO Editions.

The Standard edition provides 12GB memory, 110 teraflops performance, a 4.5MB L2 cache, and 3,072-bit memory bus. The CEO edition provides 32GB memory and 125 teraflops performance, 6MB cache, and 4,096-bit memory bus. The latter edition also uses the same 8-Hi HBM2 memory stacks that are used in the 32GB Tesla units.

NVIDIA Titan RTX

The Titan RTX is a PC GPU based on NVIDIA's Turing GPU architecture that is designed for creative and machine learning workloads. It includes Tensor Core and RT Core technologies to enable ray tracing and accelerated AI.

Each Titan RTX provides 130 teraflops, 24GB GDDR6 memory, 6MB cache, and 11 GigaRays per second. This is due to 72 Turing RT Cores and 576 multi-precision Turing Tensor Cores.

NVIDIA GeForce RTX 2080 Ti

The GeForce RTX 2080 Ti is a PC GPU designed for enthusiasts. It is based on the TU102 graphics processor. Each GeForce RTX 2080 Ti provides 11GB of memory, a 352-bit memory bus, a 6MB cache, and roughly 120 teraflops of performance.

Best Deep Learning GPUs for Large-Scale Projects and Data Centers

The following are GPUs recommended for use in large-scale AI projects.

NVIDIA Tesla A100

The A100 is a GPU with Tensor Cores that incorporates multi-instance GPU (MIG) technology. It was designed for machine learning, data analytics, and HPC.

The Tesla A100 is meant to be scaled to up to thousands of units and can be partitioned into seven GPU instances for any size workload. Each Tesla A100 provides up to 624 teraflops performance, 40GB memory, 1,555 GB memory bandwidth, and 600GB/s interconnects.

NVIDIA Tesla V100

The NVIDIA Tesla V100 is a Tensor Core enabled GPU that was designed for machine learning, deep learning, and high performance computing (HPC). It is powered by NVIDIA Volta technology, which supports tensor core technology, specialized for accelerating common tensor operations in deep learning. Each Tesla V100 provides 149 teraflops of performance, up to 32GB memory, and a 4,096-bit memory bus.

NVIDIA Tesla P100

The Tesla P100 is a GPU based on an NVIDIA Pascal architecture that is designed for machine learning and HPC. Each P100 provides up to 21

teraflops of performance, 16GB of memory, and a 4,096-bit memory bus.

NVIDIA Tesla K80

The Tesla K80 is a GPU based on the NVIDIA Kepler architecture that is designed to accelerate scientific computing and data analytics. It includes 4,992 NVIDIA CUDA cores and GPU Boost™ technology. Each K80 provides up to 8.73 teraflops of performance, 24GB of GDDR5 memory, and 480GB of memory bandwidth.

Google TPU

Slightly different are Google's tensor processing units (TPUs). TPUs are chip or cloud-based, application-specific integrated circuits (ASIC) for deep learning. These units are specifically designed for use with TensorFlow and are available only on Google Cloud Platform.

Each TPU can provide up to 420 teraflops of performance and 128 GB high bandwidth memory (HBM). There are also pod versions available that can provide over 100 petaflops of performance, 32TB HBM, and a 2D toroidal mesh network.

Learn more in our guide about TensorFlow GPUs.

NVIDIA DGX for Deep Learning at Scale

The NVIDIA DGX systems are full stack solutions designed for enterprise-grade machine learning. These systems are based on a software stack that is optimized for AI, multi-node scalability, and enterprise-grade support.

You can implement the DGX stack in containers or on bare metal. This technology is meant to be plug-n-play and is fully integrated with NVIDIA deep learning libraries and software solutions. DGX is available for server-class workstations, servers, or pods. Below, the server options are introduced.

DGX-1

The DGX-1 is a GPU server based on the Ubuntu Linux Host OS. It integrates with Red Hat solutions and includes the DIGITS deep learning training application, the NVIDIA Deep Learning SDK, the CUDA toolkit, and the Docker Engine Utility for NVIDIA GPU.

Each DGX-1 provides:

- Two Intel Xeon CPUs for deep learning framework coordination, boot, and storage management
- Up to 8 Tesla V100 Tensor Cores GPUs with 32GB of memory
- 300Gb/s NVLink interconnects
- 800GB/s communication with low-latency
- Single 480GB boot OS SSD and four 1.92 TB SAS SSDs (7.6 TB total) configured as a RAID 0 striped volume

DGX-2

The DGX-2 is the next level up from the DGX-1. It is based on the NVSwitch networking fabric for greater parallelism and scalability.

Each DGX-2 provides:

- Two petaflops of performance
- 2X 960GB NVME SSDs for OS storage and 30TB of SSD storage
- 16 Tesla V100 Tensor Core GPUs with 32GB of memory
- 12 NVSwitches for 2.4TB/s of bisection bandwidth
- 1.6TB/s low-latency, bi-directional bandwidth
- 1.5TB system memory
- Two Xeon Platinum CPUs for deep learning framework coordination, boot, and storage
- Two high I/O ethernet cards

DGX A100

The DGX A100 is designed to be a universal system for machine learning workloads, including analytics, training, and inference. It is fully optimized for CUDA-X. The DGX A100 can be stacked with other A100 units to create massive AI clusters, including the NVIDIA DGX SuperPOD.

Each DGX A100 provides:

- Five petaflops of performance
- Eight A100 Tensor Core GPUs with 40GB memory
- Six NVSwitches for 4.8TB bi-directional bandwidth
- Nine Mellanox Connectx-6 network interfaces with 450GB/s bi-directional bandwidth
- Two 64-core AMD CPUs for deep learning framework coordination, boot, and storage

- 1TB system memory
- 2x 1.92TB M.2 NVME drives for OS storage and 15TB SSD storage

Which is the best GPU for Deep Learning?

Unfortunately, there is no easy answer. The best GPU for your project will depend on the maturity of your AI operation, the scale at which you operate, and the specific algorithms and models you work with. In the preceding sections we provided many considerations that can help you select a GPU or set of GPUs that is best suited for your needs.

Automated Deep Learning GPU Management With Run:ai

Run:AI automates resource management and workload orchestration for machine learning infrastructure. With Run:AI, you can automatically run as many compute intensive experiments as needed.

Here are some of the capabilities you gain when using Run:AI:

- Advanced visibility—create an efficient pipeline of resource sharing by pooling GPU compute resources.
- No more bottlenecks—you can set up guaranteed quotas of GPU resources, to avoid bottlenecks and optimize billing.
- A higher level of control—Run:AI enables you to dynamically change resource allocation, ensuring each job gets the resources it needs at any given time.

Run:AI accelerates deep learning on GPU by, helping data scientists optimize expensive compute resources and improve the quality of their models.

Learn more about the Run.ai GPU virtualization platform.

See Our Additional Guides on Key Artificial Intelligence Infrastructure Topics

We have authored in-depth guides on several other artificial intelligence infrastructure topics that can also be useful as you explore the world of deep learning GPUs.

MLOps

In today's highly competitive economy, enterprises are looking to Artificial Intelligence in general and Machine and Deep Learning in particular to transform big data into actionable insights that can help them better address their target audiences, improve their decision-making processes, and streamline their supply chains and production processes, to mention just a few of the many use cases out there. In order to stay ahead of the curve and capture the full value of ML, however, companies must strategically embrace MLOps.

See top articles in our MLOps guide:

- Machine Learning Ops: What is it and Why We Need It
- Machine Learning Automation: Speeding Up the Data Science Pipeline
- Machine Learning Workflow: Streamlining Your ML Pipeline

Kubernetes and AI

This guide explains the Kubernetes Architecture for AI workloads and how K8s came to be used inside many companies. There are specific considerations implementing Kubernetes to orchestrate AI workloads. Finally, the guide addresses the shortcomings of Kubernetes when it comes to scheduling and orchestration of Deep Learning workloads and how you can address those shortfalls.

See top articles in our Kubernetes for AI guide:

- Kubernetes Architecture -Understanding Kubernetes Architecture for Data Science Workloads
- The Challenges of Scheduling AI Workloads on Kubernetes

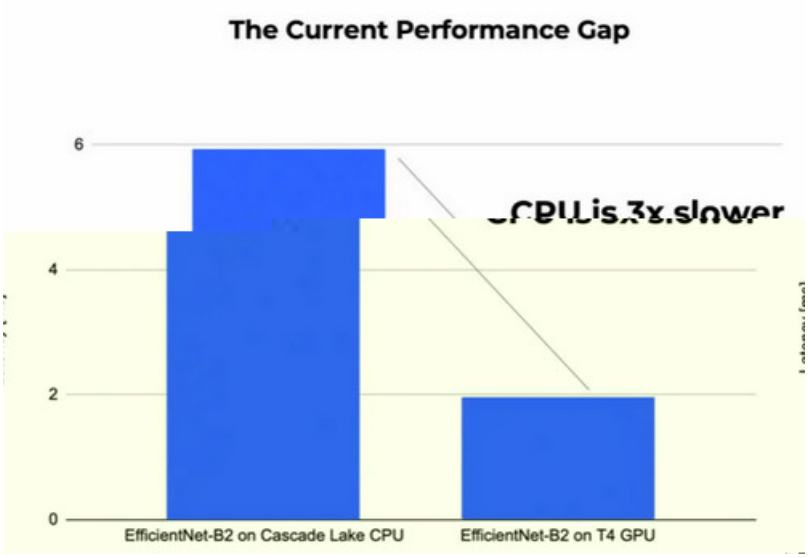
Which Graphic Card is Best for AI

Besides the technological development of A.I., the computer system also needs to handle it. Since GPU technology is advancing at a remarkable rate, GPUs have become very common for machine learning purposes. Especially when you enhance or upscale the old video using Video Enhancer AI, a strong GPU will accelerate the speed. So which graphic card is best for AI?

Why do we need a strong graphic card for AI?

CPUs are everywhere and can serve as more cost-effective options for running AI-based solutions compared to GPUs. However, finding models that are both accurate and can run efficiently on CPUs can be a challenge.

Generally speaking, GPUs are 3X faster than CPUs. Here's an example to serve as a reference for the rest of this post. The next graph shows the latency of an EfficientNet-B2 model on two different hardware: T4 GPU and Intel Cascade Lake CPU. As you can see, a forward pass on a CPU is 3X slower than a forward pass on a GPU.



Why graphic card is important for AI?

The longest and most resource-intensive phase of most deep learning implementations is the training phase. This phase can be accomplished in a reasonable amount of time for models with smaller numbers of parameters but as your number increases, your training time does as well. This has a dual cost; your resources are occupied for longer and your team is left waiting, wasting valuable time. Graphical processing units (GPUs) can reduce these costs, enabling you to run models with massive numbers of parameters quickly and efficiently. This is because GPUs enable you to parallelize your training tasks, distributing tasks over clusters of processors, and performing compute operations simultaneously. GPUs are also optimized to perform target tasks, finishing computations faster than non-specialized hardware. These processors enable you to process the same tasks faster and free your CPUs for other tasks. This eliminates bottlenecks created by computing limitations.

How to choose a suitable graphic card?

We need to consider the following factors when choosing a suitable graphic card:

Ability to interconnect GPUs

When choosing a GPU, you need to consider which units can be interconnected. Interconnecting GPUs is directly tied to the scalability of your implementation and the ability to use multi-GPU and distributed training strategies. Typically, consumer GPUs do not support interconnection (NVlink for GPU interconnects within a server and Infiniband/RoCE for linking GPUs across servers) and NVIDIA has removed interconnections

on GPUs below RTX 2080.

Supporting software

NVIDIA GPUs are the best supported in terms of machine learning libraries and integration with common frameworks, such as PyTorch or TensorFlow. The NVIDIA CUDA toolkit includes GPU-accelerated libraries, a C and C++ compiler and runtime, and optimization and debugging tools. It lets you start immediately without worrying about building custom integrations.

Recommendations for Graphic Card

The following are GPUs recommended for use in large-scale AI projects.

* NVIDIA Tesla A100

The A100 is a GPU with Tensor Cores that incorporates multi-instance GPU (MIG) technology. It was designed for machine learning, data analytics, and HPC. The Tesla A100 is meant to be scaled to up to thousands of units and can be partitioned into seven GPU instances for any size workload. Each Tesla A100 provides up to 624 teraflops performance, 40GB memory, 1,555 GB memory bandwidth, and 600GB/s interconnects.

* NVIDIA Tesla V100

The NVIDIA Tesla V100 is a Tensor Core enabled GPU that was designed for machine learning, deep learning, and high performance computing (HPC). It is powered by NVIDIA Volta technology, which supports tensor core technology, specialized for accelerating common tensor operations in deep learning. Each Tesla V100 provides 149 teraflops of performance, up to 32GB memory, and a 4,096-bit memory bus.

* NVIDIA Tesla P100

The Tesla P100 is a GPU based on an NVIDIA Pascal architecture that is designed for machine learning and HPC. Each P100 provides up to 21 teraflops of performance, 16GB of memory, and a 4,096-bit memory bus.

* NVIDIA Tesla K80

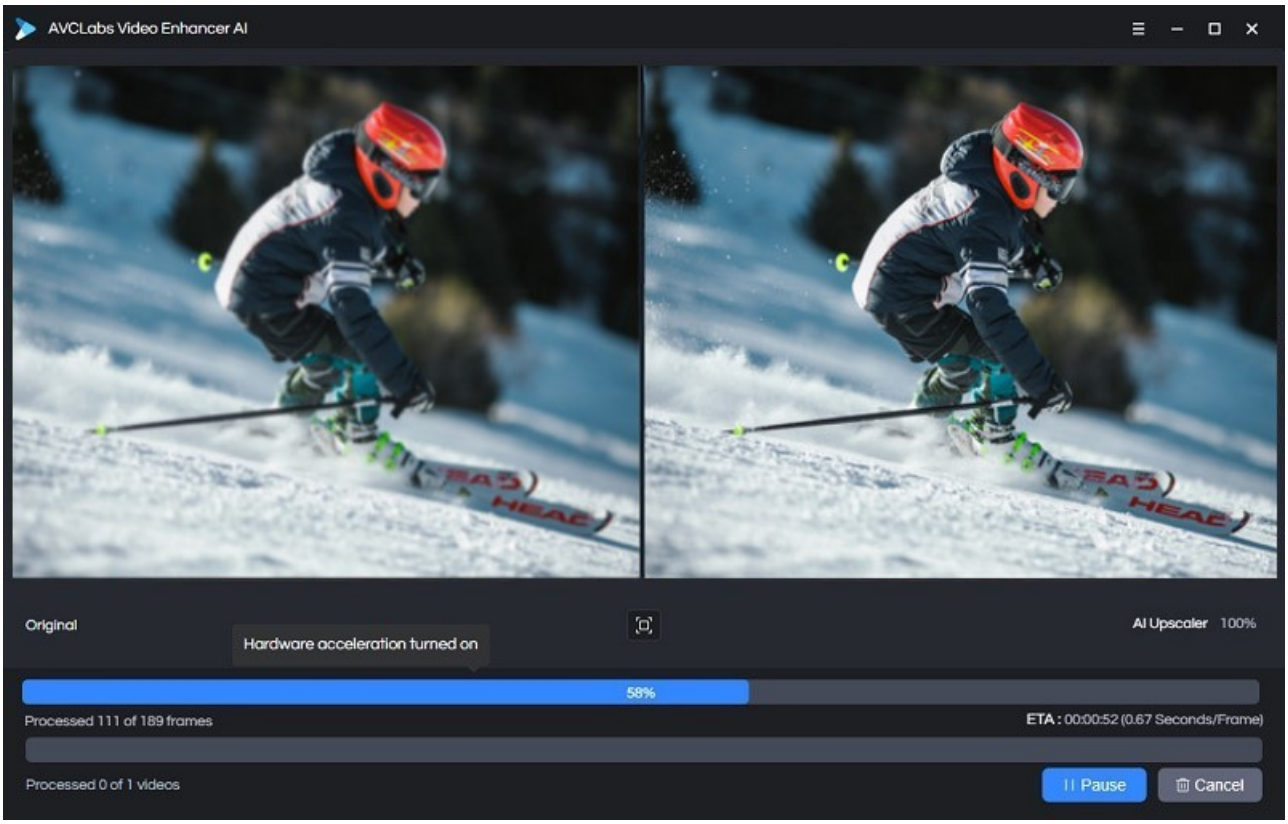
The Tesla K80 is a GPU based on the NVIDIA Kepler architecture that is designed to accelerate scientific computing and data analytics. It includes 4,992 NVIDIA CUDA cores and GPU Boost™ technology. Each K80 provides up to 8.73 teraflops of performance, 24GB of GDDR5 memory, and 480GB of memory bandwidth.

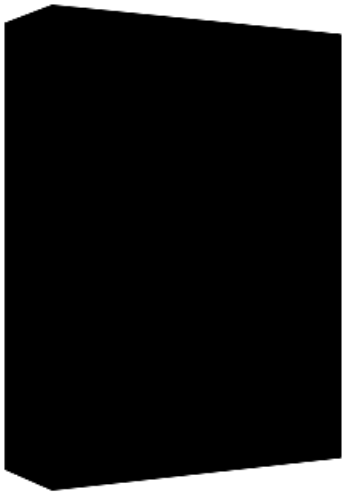
* Google TPU

Slightly different are Google’s tensor processing units (TPUs). TPUs are chip or cloud-based, application-specific integrated circuits (ASIC) for deep learning. These units are specifically designed for use with TensorFlow and are available only on Google Cloud Platform.Each TPU can provide up to 420 teraflops of performance and 128 GB high bandwidth memory (HBM). There are also pod versions available that can provide over 100 petaflops of performance, 32TB HBM, and a 2D toroidal mesh network.

Conclusion

Which graphic card is best for AI? There isn't really any definitive answer to this question as it depends on a number of factors, including what type of AI you are using and what your budget is. And if you are a user who wants to enhance or upscale video, during our testing of using AVCLabs Video Enhancer AI, the Nvidia's GeForce GTX 1080 Ti is excellent in the processing. You can also download AVCLabs Video Enhancer AI for a try!





AVCLabs Video Enhancer AI

Use Multi-frame enhancement to improve the visual quality

Upscale video from SD to HD, HD to 4K, or 8K

Sharpen blurry faces and enhance facial details

Remove noise and grain from the your noisy footage

Support lossless MP4, MOV, MKV, AVI as the output format

The 5 Best GPUs for Deep Learning to Consider in 2022

The editors at Solutions Review have compiled this list of the best GPUs for deep learning based on advice from experts in the field.

The best GPUs for deep learning and data science are becoming an increasingly vital hardware requirement as practitioners scale analytics and machine learning. The challenge of finding the right graphics processing unit for your use case can be difficult for this very reason. So project your current and future needs carefully because GPU selection will hinge mainly on your workload. You will also need to take into account that different products are better served for personal versus professional use.

With these things in mind, our editors assembled this list of the best GPUs for deep learning based on expert advice from some of the top contributors on Quora. We also consulted A 2021-Ready Deep Learning Hardware Guide from the folks at Towards Data Science, which is another excellent resource. Each of the best GPUs for deep learning featured in this listing are featured under Amazon's Computer Graphics Cards department. Only products with verified customer reviews are included.

Note: The best GPUs for Deep Learning are listed in order based on the total number of Amazon user reviews at the time of publication.

NVIDIA Tesla K80

SUMMARY: The NVIDIA Tesla K80 has been dubbed "the world's most popular GPU" and delivers exceptional performance. The GPU is engineered to boost throughput in real-world applications while also saving data center energy compared to a CPU-only system. The increased throughput means improved performance. The K80 features 4992 NVIDIA CUDA cores with a dual-GPU design, 24GB of GDDR5 memory, 480 GB/s aggregate memory bandwidth, ECC protection for increased reliability and server-optimization.

PROS

- Supports DirectX 12
- Ideal for an array off AI use cases
- Proven, reliable performance

CONS

- No integrated fans
- Requires passive cooling
- Not ideal for very large workloads

OUR TAKE: The NVIDIA Tesla K80 combines two graphics processors to increase performance. Being a dual-slot card, the NVIDIA Tesla K80 draws power from a 1x 8-pin power connector, with power draw rated at 300 W maximum. This device has no display connectivity, as it is not designed to have monitors connected to it. Tesla K80 is connected to the rest of the system using a PCI-Express 3.0 x16 interface.

NVIDIA GeForce GTX 1080 Ti

SUMMARY: The NVIDIA GeForce GTX 1080 is powered by NVIDIA's popular Pascal architecture and offers top-notch performance and power efficiency. According to NVIDIA, Pascal can deliver up to "3x the performance of previous-generation graphics cards, plus innovative new gaming technologies and breakthrough VR experiences." The GTX 1080 also touts upgraded heat dissipation from previous generations, as well as a vapor chamber cooling technology. This GPU is made out of premium materials as well.

PROS

- Pascal architecture
- Upgraded heat dissipation
- Component protection system

CONS

- Unsecure fan cables
- Runs hot at factory settings
- Expensive for the functionality

OUR TAKE: The NVIDIA GeForce GTX 1080 supports DirectX 12 and features a large chip with a die area of 314 mm² and 7,200 million transistors. It offers major upgrades from its GeForce GTX 980 predecessor like a new architecture framework, double the frame buffer RAM, 30 percent faster memory speed, and more juice out of the boost clock. Display outputs include 1x DVI, 1x HDMI, 3x DisplayPort. GeForce GTX 1080 is connected to the rest of the system using a PCI-Express 3.0 x16 interface.

GeForce RTX 2080 Founders Edition

SUMMARY: The NVIDIA GeForce RTX 2080 is powered by NVIDIA's next-generation Turing architecture which, according to the company "gives you up to 6X the performance of previous-generation graphics cards." The Turning architecture brings AI-processing horsepower that hastens performance with NVIDIA DLSS 2.0. as well. Simultaneous floating-point and integer processing enables the GPU to more efficiently process and compute-heavy workloads.

PROS

- Turing architecture
- Factory overclocked
- DLSS 2.0 graphics

CONS

- Minor air exhausting issues
- Resolution output learning curve
- Value lacking compared to other NVIDIA GPUs

OUR TAKE: The Founders Edition of the NVIDIA GeForce RTX 2080 is factory overclocked and offers an 8-phase power supply for overclocking. It also boasts a dual-axel 13-blade fan coupled with a vapor chamber for cooler and quieter performance. Compared to the base model NVIDIA RTX 2080, this version offers small but notable improvements. NVIDIA has paired 8 GB GDDR6 memory with the GeForce RTX 2080, which are connected using a 256-bit memory interface. The GPU is operating at a frequency of 1515 MHz and can be boosted up to 1710 MHz.

NVIDIA GeForce RTX 3060 XC

SUMMARY: The NVIDIA GeForce RTX 3060 takes advantage of NVIDIA's Ampere architecture, the company's second-generation RTX framework. The GPU, according to the company, offers "Ray Tracing Cores and Tensor Cores, new streaming multiprocessors, and high-speed G6 memory." The GeForce RTX 3060 also touts NVIDIA's Deep Learning Super Sampling, an AI rendering that boosts frame rates with uncompromised image quality using a dedicated Tensor Core AI processing framework.

PROS

- Ampere architecture
- DLSS AI acceleration
- Great for modern use cases

CONS

- Some shading is disabled
- No DVI ports
- Current price point

OUR TAKE: The NVIDIA GeForce RTX 3060 is a premium GPU that supports DirectX 12 Ultimate. Unlike the fully unlocked GeForce RTX 3070, which uses the same GPU but has all 6144 shaders enabled, NVIDIA has disabled some shading units on the GeForce RTX 3060. The 3060 also includes 152 tensor cores which help to increase the speed of machine learning applications. The product has 38 raytracing acceleration cores as well. The card measures 242 mm in length, 112 mm in width, and features a dual-slot cooling solution.

NVIDIA TITAN RTX

SUMMARY: The NVIDIA Titan RTX is designed for researchers, developers and creators. The GPU is powered by NVIDIA's Turning architecture and touts 130 Tensor TFLOPs of performance, 576 tensor cores, and 24GB of GDDR6 memory. The Titan RTX is supported by NVIDIA drivers and SDKs as well. According to NVIDIA, the Titan RTX works with "all popular deep learning frameworks and is compatible with NVIDIA GPU Cloud (NGC)."

PROS

- Turing architecture
- Designed for AI and machine learning
- Great for large models and neural networks

CONS

- Coil whine under heavy stress
- Additional cooling sometimes needed
- Use case dependant; compare to NVIDIA RTX 2080

OUR TAKE: The NVIDIA Titan RTX is a dual-slot card with a power draw rated at 280 maximum watts. Display outputs include 1x HDMI, 3x DisplayPort, 1x USB Type-C and is connected to the rest of the system using a PCI-Express 3.0 x16 interface. The DirectX 12 Ultimate capability ensures support for hardware-raytracing, variable-rate shading and more. The graphics processor is on the larger side.

NOW READ: The Best Data Analytics Laptops for Data Science

- Author
- Recent Posts

Tim is Solutions Review's Editorial Director and leads coverage on big data, business intelligence, and data analytics. A 2017 and 2018 Most Influential Business Journalist and 2021 "Who's Who" in data management and data integration, Tim is a recognized influencer and thought leader in enterprise business software. Reach him via tking at solutionsreview dot com.

Best Graphics Cards 2022 - IGN

Between Nvidia's powerful Ampere GPUs and AMD's designs using a 7nm process, the graphics card market may be the most competitive it has been in some time. That competition means good value for you as a consumer, especially considering there are options to fit just about every need. So, if you're looking to outfit your gaming PC with one of its most crucial parts, you've come to the right place.

You can find one for a little over $300 that will make short work of 1080p gaming, and you won't break the bank to enjoy 1440p at high frame rates or even a perfectly playable 4K. Thanks to so many variants, like Nivida's Ti and AMD's XT cards, you can find something that'll perfectly pair with your gaming monitor. Just keep in mind there is a global chip shortage making graphics cards harder to find and keep in stock, so be sure to keep checking back if your top choice is not available. We'll walk you through the best options, so you can pick out exactly what you need to make your computer hum – and click here to see them in the UK.

TL;DR – These are the Best Graphics Cards:

- Nvidia RTX 3070
- EVGA RTX 3060 XC Black
- Nvidia GeForce RTX 3080 Ti Founders Edition
- XFX Speedster SWFT319 AMD Radeon RX 6800
- Nvidia RTX 3060 Ti Founders Edition
- Asus TUF Gaming RTX 3070
- MSI Gaming Radeon RX 6800 XT
- MSI RTX 3080 Gaming X Trio
- Nvidia GeForce RTX 4090
- AMD Ryzen 7 5700G

Nvidia RTX 3070

Best Graphics Card

For an incredible marriage of performance and price, you can't do better than the new Nvidia RTX 3070. This card costs just $500, but it is capable of offering performance levels exceeding even the Nvidia RTX 2080 Ti, which had been retailing often for more than three times the price of the RTX 3070.

This latest model offers up 5,888 CUDA cores that can run at a decent clip with a 1,730MHz boost clock. The ray tracing performance will fully immerse you in your games thanks to its RT cores. It also includes 8GB of GDDR6, so you'll have plenty of memory for game textures and frame buffer. This card can do some 4K if that's your aim, but it's best suited for maxing out on 1440p. We're not talking 1440p/60Hz either, but rather high-speed and high-resolution gaming.

EVGA RTX 3060 XC Black

Best Budget Graphics Card

- See it on EVGA

If you’re shopping for components for your gaming PC build on a budget and lucky enough to come across the EVGA RTX 3060 XC Black at its retail price of $349, it’s a very worthwhile component. You’ll be getting a ton of power for your dollar with 3,585 CUDA Cores offering speeds and capabilities more in line with the previous generation’s RTX 2070 than with the RTX 2060 or GTX 1660 Super that it succeeds.

The EVGA RTX 3060 XC Black also dials things up above the stock RTX 3060. That comes in the form of a ramped up clock speed that we saw exceed the default boost clock of 1,777MHz, even going as high as 1.9GHz in our testing. EVGA has also built the RTX 3060 XC Black to run fairly cool and quiet, as it peaked at 69C and 36dB in our testing. The 1080p graphics card works great, but if you want to do any 1440p gaming, you may have to turn down your settings. As a bonus for those who like a subtle build, the card is also all black, including the PCB.

Nvidia GeForce RTX 3080 Ti Founders Edition

Best 4K Graphics Card

Nvidia GeForce RTX 3080 Ti Founders Edition

Nvidia has stepped things up with the Nvidia RTX 3080 Ti Founders Edition graphics card. This new card is more than a small upgrade to the original RTX 3080 and its 4k performance is faster falling just a fraction behind the RTX 3090. It has a whopping 10,240 CUDA cores, giving it over a 17% increase over the RTX 3080's 8,704 CUDA cores and bringing it close to the RTX 3090's 10,496 CUA cores. The RTX 3080 Ti may be running them at slightly lower clocks, but you'll still see heaps of performance out of this card.

That's not the only upgrade in store. The RTX 3080 Ti includes 2GB of extra GDDR6X memory and boosts the memory bus to 384-bit, giving it an edge when handling large game assets. Despite the increase in performance, the RTX 3080 Ti maintains the same proportions as the RTX 3080, so you should have no trouble fitting it into your system if you're upgrading. And, at 350W, it's only drawing 30W more power than the RTX 3080 before it. If you're in the market, this card really is one the most powerful that you can find on the market

XFX Speedster SWFT319 AMD Radeon RX 6800

Best 1440p Graphics Card

XFX Speedster SWFT319 AMD Radeon RX 6800

AMD came out swinging with the Radeon RX 6800. It isn't the most powerful around, but it goes toe to toe with Nvidia's RTX 3070 in pretty much any match-up that doesn't pull in ray-tracing and DLSS, and it usually comes out ahead. This performance sees it especially well suited for 1440p gameplay, where you can enjoy high frame rates.

The AMD Radeon RX 6800 needs a modest 250 watts of power, which should make it easier to slot into existing computers without needing to upgrade your power supply to something beefier. Plus, its new design with a 3-fan cooler system makes it quieter and more stylish. And, even if you're only gaming at 1440p, the RX 6800's 16GB of GDDR6 memory will serve as a solid buffer for your frames as well as home to extra-high resolution textures — the last thing you want is sharp and smooth visuals to show off low-res game textures because your card didn't have enough VRAM.

Nvidia RTX 3060 Ti Founders Edition

Best 1080p Graphics Card

Nvidia RTX 3060 Ti Founders Edition

4K and 1440p may be exciting, but they can also be an entry into the stressful world of constantly trying to optimize your gaming rig to run at either high quality or high frame rates. Playing at 1080p still provides clear visuals and makes it so much easier to just crank everything to the max without worrying too much about low frame rates. And, the new Nvidia RTX 3060 Ti Founders Edition is geared for epic performance in 1080p.

This graphics card comes with the slick styling of its RTX 30-Series siblings, but tones down the performance for a corresponding reduction in price. We're talking just $400. That'll land you a card that's got more than enough VRAM to handle high-quality game assets and enough muscle to spit out high frame rates a 1080p. And, unlike the GTX 1660 Ti, the RTX 3060 Ti is built with dedicated hardware for ray tracing and Deep Learning Super Sampling. In many cases, the RTX 3060 Ti can even outperform the RTX 2080 Super. If your want to be gaming in 4k save up and look elsewhere, but if you want solid 1080p performance and an unbelievable price, look no further than this graphics card.

Asus TUF Gaming RTX 3070

Best Nvidia RTX Graphics Card

One of the best ways to enjoy Nvidia's excellent new RTX 3070 is through the options available from Nvidia's board partners. Asus's TUF Gaming RTX 3070 is an excellent option. You'll still get the same 5.888 CUDA cores and boost clock, but you'll be getting a different design that comes a few perks of its own.

There are a couple of chief differences between this card and Nvidia's reference model. For one, you'll get multiple HDMI 2.1 ports, letting you take advantage of the supported 4K/120Hz signal on multiple displays that lack DisplayPort. You'll also find two 8-pin power connectors instead of the new 12-pin connector that Nvidia has implemented. And then there's the triple-fan design, which should have no trouble keeping this card cool and quiet, just as the Asus TUF Gaming RTX 3080 did in our review.

MSI Gaming Radeon RX 6800 XT

Best AMD Graphics Card

MSI Gaming Radeon RX 6800 XT

At $649, AMD's Radeon RX 6800 XT clears a niche for itself in the market of recently released graphic cards. It's more affordable than Nvidia's RTX 3080, and it's only a bit more expensive than the non-XT RX 6800 while increasing the number of compute units and clock speeds. The 16GB of GDDR6 memory is unmatched by most competitors. All of this makes it an especially potent choice for anyone that's optimizing for value.

The RX 6800 XT can hold its own against the RTX 3080 in a lot of cases, with exceptional performance at 1080p and 1440p plus decent chops in 4K. It may not have much to offer when it comes to ray tracing, but that's still not a widely implemented feature in games, and the RX 6800 XT may yet regain some ground when AMD eventually launches its FidelityFX Super Resolution feature. So, if you're not overly concerned about uncertain performance in ray-tracing, the Radeon RX 6800 XT offers a compelling alternative to Nvidia's RTX 3080 while costing less and drawing less power from your wall.

MSI RTX 3080 Gaming X Trio

Best for High-End Gaming for Most Gamers

MSI RTX 3080 Gaming X Trio

- See it on Newegg

If you're looking at the RTX 3080, you definitely are looking for speed. The MSI RTX 3080 Gaming X Trio takes the already excellent GPU card, and makes it even faster. In our testing, it was one of the faster RTX 3080 models. Those speeds come right out of the box as well, so with some tweaking, you could potentially see even more.

The MSI RTX 3080 Gaming X Trio is fairly beefy, as it's stacking on a triple-fan cooler to keep temperatures in check, and it'll require you to have three 8-pin connectors. That extra power may come in handy if you're trying to overclock this card for even more performance than the stock boost clock speed of 1815MHz. MSI tops it all off with a bit of RGB lighting. All that's on offer here makes up for the $50 price hike over the Nvidia reference RTX 3080.

Nvidia GeForce RTX 4090

The Out of Your Mind Graphics Card

If you need a machine that won't sweat in even the most demanding situations, then you'll want the latest and greatest technology. The Nvidia GeForce RTX 4090 happens to be that, as it’s part of the newest generation of Nvidia GPUs. It includes 24GB of GDDR6 memory and an insane 16,384 CUDA cores. This card even manages to run those cores at a high base clock, and it can be overclocked much higher. That means this rig increases the shader teraflops of computing power, tensor teraflops, and ray tracing teraflops by a wild number compared to the RTX 3090 Ti.

You'll need a beefy power supply and some major cooling power to run the hefty RTX 4090 in your system, but you'll get that energy back in the form of serious performance. It can achieve smooth 4K gaming, and depending on the type of game you throw at it or whether it has any enhancements like DLSS, it can even make 8K playable. Speaking of DLSS, with DLSS 3 on board, there’s the potential to bring 4x the FPS with AI machine learning. You also get AV1 encoding, so you can render or stream higher-quality video without requiring more bandwidth. And when you're not gaming, this card puts other GPUs to shame in 3D modeling or content creator tasks. If you don’t mind dropping a ton of dough, this is the best graphics card out there.

AMD Ryzen 7 5700G

Best HTPC Graphics Card

OK, hear us out. No, this isn’t a graphics card, but it can handle the graphics. The Ryzen 7 5700G is an APU from AMD that combines both the

CPU and GPU on a single chip. While that may not be the ideal setup for a high-power gaming rig, it’s almost perfect for a home theater PC setup.

Going with an APU will let you stick with a smaller build. You don’t need to consider the space for a dedicated graphics card, nor do you need to allot for one in the power budget. The AMD Ryzen 7 5700G actually doesn’t require a beefy power supply to run, and with a 65W TDP, it doesn’t need all that serious cooling either. The eight Radeon Graphics cores on this APU will handle your HTPC needs nicely and can even spit out some playable frame rates in a wide variety of games. Just slap this baby onto a Mini ITX motherboard, feed it with some fast RAM (don’t slouch on memory for an APU, as both parts of the chip share it), throw it into one of the best Mini ITX cases, and your home theater will be ready to rock.

Where to Get the Best Graphics Cards in the UK

There aren't too many differences when it comes to the graphics cards you can pick up in the UK, but the main takeaway is where you can purchase them. All of the following links have been updated with UK vendors, saving you some time and money if you're interested in picking up any of the graphics cards we've mentioned. Don't see the graphic cards below? Click here.

Best Graphics Card

Nvidia RTX 3070

Often low in stock. Check back for updates.

Best 4K Graphics Card

Nvidia GeForce RTX 3080 Founders Edition

Often low in stock. Check back for updates.

Best 1080p Graphics Card

Nvidia RTX 3060 Ti Founders Edition

Often low in stock. Check back for updates.

Best 1440p Graphics Card

AMD Radeon RX 6800

Often low in stock. Check back for updates.

Best Nvidia RTX Graphics Card

Asus TUF Gaming RTX 3070

Best AMD Graphics Card

AMD Radeon RX 6800 XT

Best for High-End Gaming for Most Gamers

MSI RTX 3080 Gaming X Trio

Kick-Off Your Esports Career with this Graphics Card

MSI RX 5500 XT Mech 8G OC

The Out of Your Mind Graphics Card

Nvidia RTX 3090 Founders Edition

Best HTPC Graphic Card

PNY GeForce GTX 1650 Super

What to Look for in a Graphics Card

Below we explain how to pick the right GPU for your display, why there are so many variants of the same Nvidia and AMD graphics cards, and a few factors you should consider when buying. Above all, you should buy the graphics card you need for the display you're using.

If you're gaming on a Full HD monitor, it would be a huge waste to buy a component designed to play games at 2160p or 1440p. Likewise, you'll want a powerful graphics card to drive games playing on that premium 4K gaming monitor or 4K TV.

We've laid out what are the best graphics cards to play games at 1080p, 1440p, and 2160p resolutions above, but here are some more general rules. For a decent to high-frame-rate Full HD experience, you should look at parts ranging from the GTX 1650 to the GTX 1660 Ti on Nvidia's end. If you're looking at AMD's family, you'll want a Radeon RX 5500 or up.

Jumping up to QHD resolutions will require a more capable graphics card, ideally an Nvidia GTX 1660 Ti or AMD Radeon RX 5600 and up. 4K gaming using a single card is still a tough proposition, but thanks to recent developments it's actually approachable with the latest graphics cards like the Nvidia RTX 2080 Super and AMD Radeon VII.

Another Variable

Another thing to keep in mind when choosing the right graphics card for your gaming monitor (or vice versa) is what kind of variable refresh rate technology can you take the most advantage of. For the uninitiated, variable refresh rate (VRR) technology basically syncs the number of frames shooting out of your GPU to the frame rate of your display.

This way it isn't overworking itself for nothing while also helping to eliminate screen tearing on your monitor. Without this VRR tech, your GPU might end up clogging the frame bugger with two or more frames, which your display might then try and display two different shots of gameplay at the same time. If you have a TV and gaming monitor that supports FreeSync, you should get an AMD graphics card.

Alternatively, if you happen to be playing primarily a G-Sync gaming monitor or one of the latest LG C1 OLED TVs then you'll want an Nvidia card. Luckily for you, the line separating G-Sync and Freesync is quickly disappearing as more and more displays that offer the latter are adding support for the former.

G-Sync-compatible gaming monitors are all the rage now because they offer a tear-free and smooth gameplay experience when connecting to either an AMD and Nvidia graphics card.

Graphics card variants

Ok, you've decided which one you want, great! However, even with this monumental decision out of the way, the world of GPUs isn't done being confusing and daunting just yet. Although there are only two companies—Nvidia and AMD—that actually manufacture the underlying chip, there are dozens of different variants of the same graphics card

For example, when the most recent graphics card launched, the Nvidia GTX 1650, there was a multitude of different versions from Asus, Gigabyte, MSI, EVGA and the list goes on. In this case, while Nvidia may have introduced only one new model, vendors or board partners will introduce their own versions featuring different overclock settings, cooling systems, and other differentiating factors we will explain below.

Length: One of the number one factors you should consider before plopping down cash for that shiny new component is whether it will actually fit. If you're building your PC in a Mini ITX case, you should be looking at the smallest or mini graphics cards that will actually fit inside.

Overclocking: Most third-party cards—and even Nvidia's own Founders Edition cards—will often come factory overclocked, and this means the card has been tuned to operate above its rated maximum clock speed. As you might expect, the higher the number the faster it will perform.

At this point, you won't find many, including the entry-level cards, without some amount of 'overclocking from the factory.' However, even without a factory overclock, it's easy enough do it yourself using software such as EVGA Precision X or MSI Afterburner.

Cooling solutions:

In your quest for the best graphics card, you might have noticed that some models come with one, two, or up to three fans. As you might expect, more fans equal better cooling, but there are also two distinct ways of keeping it chilled. GPUs equipped with a single fan often use a blower-style cooler, which means the card sucks in air and blows it out the back like a leaf blower.

Dual and triple fan setups are often used in conjunction with 'open-air cooling systems,' which are designed to move cool air through the open heatsinks and exhaust heat in every direction.

Blower style coolers are typically most useful for PCs built into small Mini ITX cases because they help exhaust heat out of a compact chassis with

restricted airflow. If the system you're building is in a Micro ATX PC case or a larger Mid tower chassis, you'd be better off with an open-air cooled graphics card, as there are more mounting points for multiple case fans to do the brunt of cooling while the card's own two (or three) fans blow heat off the card itself.

RTX vs GTX: With Turing, Nvidia didn't just introduce better, faster graphics cards it also debuted the RTX architecture with hardware designed to support real-time ray tracing, and AI-powered supersampling and anti-aliasing (known as Deep Learning Super Sampling).

Thankfully, Nvidia decreed in early April 2019 that you don't need an RTX card with dedicated RT Cores to process real-time ray tracing. So any of the GTX 16-series cards and (most) older 10-series cards can run games with ray tracing turned on. DLSS is still an RTX exclusive since it requires Tensor cores to function, but it's a niche performance smoothing feature compared to the strikingly realistic reflections and complex shadows effects that ray tracing produces.

Bargain your way to getting a graphics card

Strangely, one of the more affordable ways to get yourself the latest graphics card is to buy a gaming PC while it's on sale. Gaming PCs from brands like Asus, Dell, MSI, Acer, and HP will often see discounts for hundreds of dollars off, so not only are you saving a ton of money, you're also avoiding potential headaches that can accompany a DIY build—and you also get a warranty.

Prebuilt PCs have come a long way, too. They aren't proprietary machines with randomly soldered-on components. They're mostly as upgradeable as anything you might put together on your own.

Another way of enjoying the latest graphics cards is through gaming laptops. There are plenty of Nvidia RTX 20- and GTX 16-series gaming laptops out there right now. New GTX gaming laptops have also hit the streets and they're far more affordable than the RTX-equipped models thanks to the laptops introduced during IFA 2019 like the new Acer Predator Triton 300.

Kevin Lee is IGN's Hardware and Roundups Editor. Follow him on Twitter @baggingspam

Danielle Abraham is a freelance writer and unpaid music historian.

Best Video Card For Ai | 10 Top Picks Of 2022

After analyzing all the features, specifications, functionality and exploring the detail of more than 1,652 customer opinions about the best video card for ai, We have come up with the top 10 video card for ai for you. We ranked the products by considering brands, models, price, durability, consumer reviews, and more. We also taking help from Artificial Intelligent and Big Data, as you see.

A company or manufacturer may claim that its product has the most upgraded features but at the end of the day what matters most is consumer satisfaction i.e., real user's experience during using the product in real life. So, we have spent a couple of hours analyzing the reviews of all consumers so that we can give you a realistic idea about the product.

Let's see our top 10 Best video card for ai of 2022

1

9.8

2

9.3

3

9.3

4

9

5

8.1

6

8

7

7.8

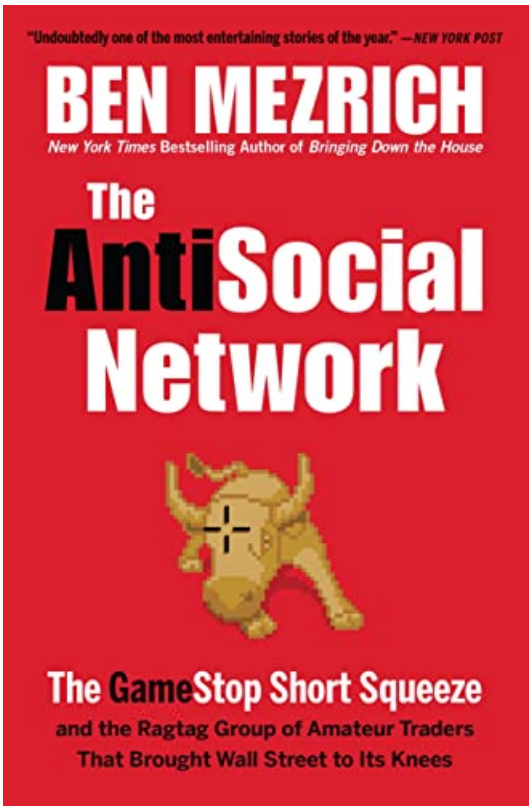
8

7.5

9

7.2

10



7.2

You can get an exact idea about the good and bad sides of the product from the real user review and also from a forum discussion. But if you have less time to find out the details of the product from social media or forum discussion, you can quickly get one from our best video card for ai list. It will save your time and energy.

When you have enough information about the product you won't feel overwhelmed by the variety, attractive designs, and features, rather you will feel confident during making a purchase – no matter whether you purchase it online or offline.

Happy Shopping!!