Challenges in Implementing AI

There currently are many challenges in the medical device development process for endoscopy.

- Ingesting high-resolution, high-bandwidth data streams.
- Running AI inference with a low latency budget.
- Requiring flexible sensor and data IO options
- Building a distributed compute platform from edge to data center to cloud.
- Adopting new deep learning algorithms.
- Managing long-life and upgradable hardware and software.
- Repeatedly developing, testing, and commercializing hardware and software that does not add unique value.

To address these common challenges,

- Holoscan has a ConnectX NIC to enable 10/100 GbE data ingestion.
- GPU direct RDMA shortens data transfer time before AI inference; NVIDIA RTX GPUs provide powerful AI inference capabilities.
- The Clara Holoscan ecosystem has data IO partners for different sensor needs.
- NVIDIA offers a distributed compute stack, EGX, as part of the larger product family.
- There are many tools to help developers adopt the latest deep learning algorithms, such as MONAI, an open-source, healthcare-specific, PyTorch-based framework with labeling, training, and deployment capabilities.
- provides 10-year hardware and software long-term support; also provides documentation to support IEC 60601 and IEC 62304 certification.

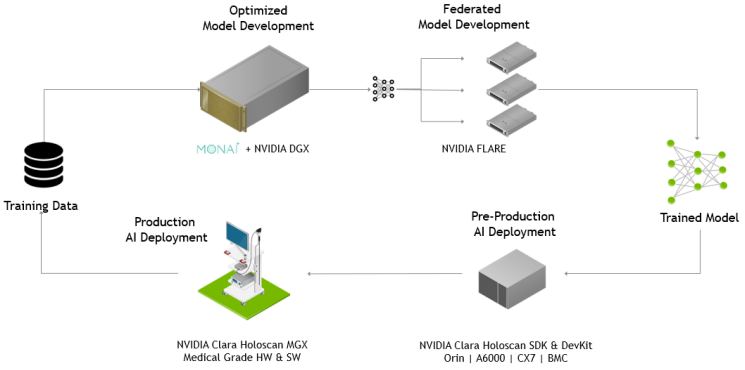
The Solution: Clara Holoscan for Next Generation AI-Enabled Endoscopy

Clara Holoscan is a development platform, comprising both hardware developer kits and the software stack. Clara Holoscan offers compute for AI workloads, such as enhanced visualization and automatic anomaly detection, to better assist doctors performing endoscopies. Clara Holoscan helps you deploy latency-sensitive real-time tasks on the edge and analytic/summarization tasks to the cloud. There are reference applications to give you an easy starting point developing your own applications. To help with that end-to-end application development process, you can use the MONAI ecosystem to label, train, and validate your own models. Use the Holoscan developer kit as a starting point, design our reference architecture into your existing hardware, or go into production with our medical grade hardware.

NVIDIA CLARA HOLOSCAN END-TO-END PLATFORM

Real Time Medical Grade AI Computing Platform

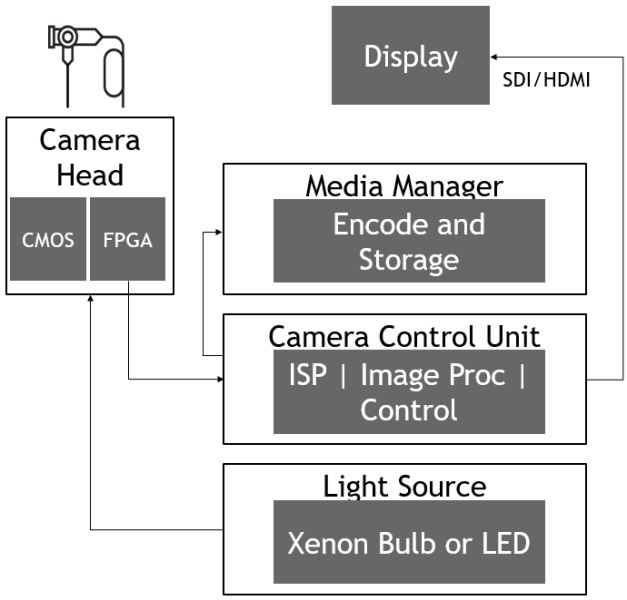
- AI Label, Train and Validate with MONAI
- Improve Accuracy, Generalize with NVIDIA FLARE
- Cloud Natively Deploy with Clara Holoscan
- Accelerate TTM with Medical Grade Full Stack
- Safe, Secure, Manageability for CI/CD



Using AI to Assist Doctors in Abnormality Detection

The Traditional Endoscopy Hardware Setup

The traditional endoscopy hardware setup has a camera head, a light source, a Camera Control Unit, a Media Manager, and a connected display. The camera head is handheld, with a limited power and size budget, and has a short-reach MIPI or SLVDS interface to the sensor. The CCU typically includes multiple FPGAs, has a custom Image Signal Processor (ISP) of the camera feed, and also controls light source at low latency based on ISP output. The Media Manager encodes and stores the video feed.



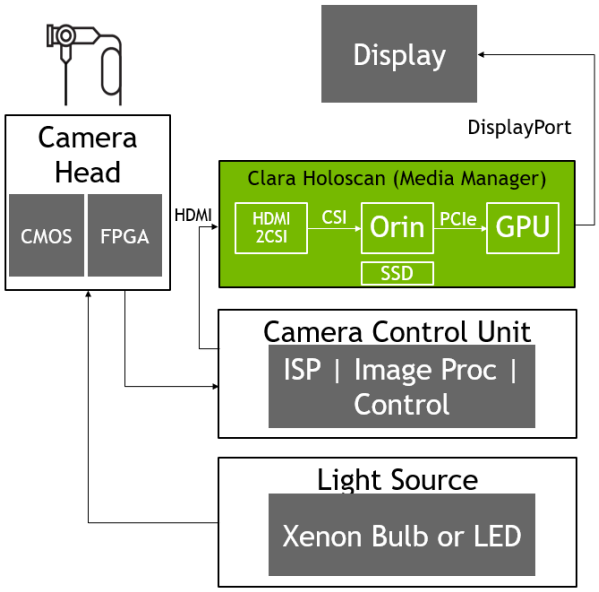
The Role of the Clara Holoscan Platform in an Endoscopy Hardware Setup

You can adopt either of the two cases we have available today.

Case 1: Clara Holoscan replaces the Media Manager

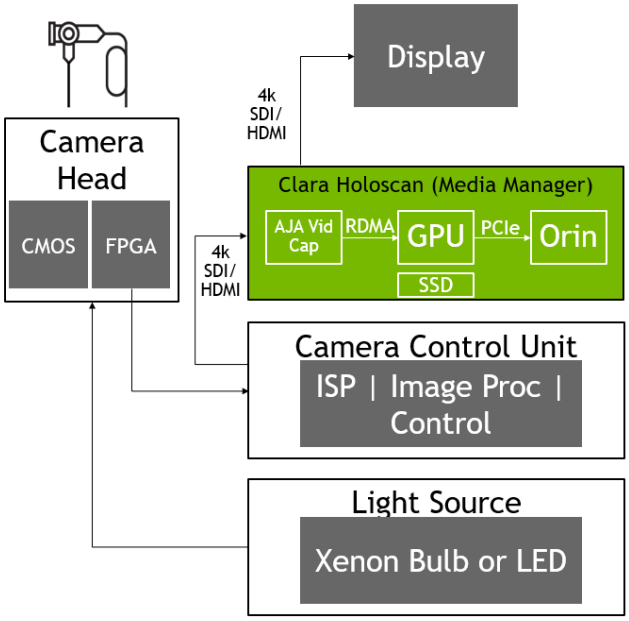
Clara Holoscan can replace the traditional Media Manager for encoding and storing the video feed by connecting to CCU via an SDI to HDMI Adapter and using the onboard HDMI2CSI

interface to directly feed the video stream to an Orin SoC. The Orin SoC then provides standard encode and storage functions of the Media Manager, but additionally also connects to the discrete GPU to run assisting AI tasks such as anatomy recognition, tool identification, tool tracking, etc, before rendering to a display.



Case 2: Clara Holoscan replaces the Media Manager with 4K support

Instead of going through the onboard HDMI2CSI interface combined with a SDI to HDMI adapter, Clara Holoscan can receive video from the CCU through two to four SDI or HDMI channels supporting 4k 10bit YUV422 video on an AJA video capture device. AJA video capture devices have established GPU Direct RDMA on Clara Holoscan and run the same assisting AI tasks with low latency.


Display
4k
SDI/
HDMI
Camera
Head
CMOS
FPGA
Clara Holoscan (Media Manager)
AJA Vid
Cap
RDMA
GPU
PCIe
Orin
SSD
4k
SDI/
HDMI
Camera Control Unit
ISP | Image Proc |
Control
Light Source
Xenon Bulb or LED

AI reference application

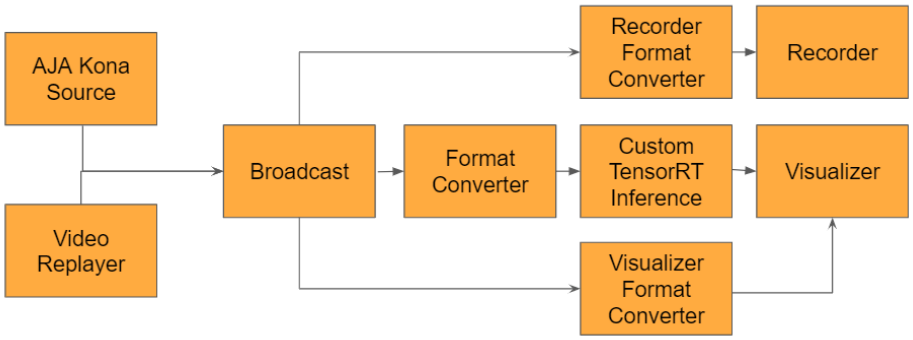
Setup

The AI reference applications in the Clara Holoscan SDK are based on GXF, which is a modular and extensible framework to build high-performance AI applications. This enables developers to reuse components and app graphs in existing Clara Holoscan reference applications as templates and build their own custom AI pipelines using common data formats and the foundational GXF Core.

Clara Holoscan SDK features an endoscopy tool tracking reference application in the v0.2 release. The main features in the GXF reference application are:

- GPUDirect RDMA with AJA video capture card over PCIe which eliminates the overhead of copying to and from system memory
- NVIDIA Performance Primitive Library (NPP) which enables CUDA-accelerated 2D image transformations
- TensorRT runtime for optimized AI Inference
- CUDA and OpenGL interoperability which provides efficient resource sharing on the GPU in the visualizer

The endoscopy tool tracking app is based on a LSTM (long-short term memory) model. The app can have three types of input streams: an external video stream via an AJA capture card, a mock video buffer to simulate an AJA capture card stream for testing purposes, and a pre-recorded video. The custom LSTM stateful inference module uses TensorRT for tool tracking, and developers can bring their own LSTM model into the inference module. The custom visualizer component handles compositing, blending, and visualization of tool labels, tips, and masks given the output tensors of the custom LSTM inference. The optional recorder serializes incoming messages, in this case RGBA888 frames, and writes them to a file.



Each of the blocks in the above diagram is a GXF extension, which is a shared library and/or header file containing one or more components. There are two types of extension: Standard Extensions that come with GXF and Custom Extensions. Developers and NVIDIA product teams such as Isaac, Jarvis Speech, Jarvis vision, Clara Holoscan, Metropolis, DeepStream can create, store and share custom extensions in the GXF Registry for all developers to reuse.

In the reference application shown in the block diagram, we have a mix of Clara Holoscan GXF extensions (AJASource, FormatConverter, custom_lstm_inference::TensorRtInference, visualizer_tool_tracking) and GXF standard extensions (Broadcast, EntityRecorder).

There are additional Clara Holoscan GXF extensions available in the Clara Holoscan v0.2 release not shown in the endoscopy tool tracking application:

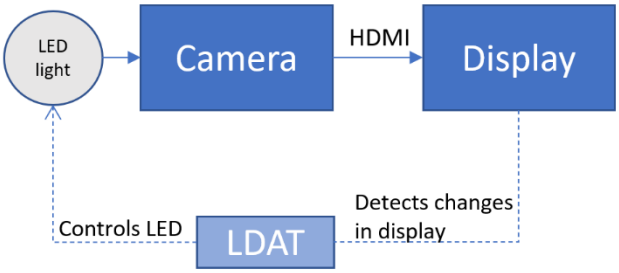
1. v4l2 video source for a Linux-source USB Camera, with output as a VideoBuffer object
2. Probe to print Tensor information during testing
3. Post-processor for segmentation models, converting the inference output to the highest-probability class index, including support for sigmoid, softmax, and other activations
4. Visualizer for segmentation, which is an OpenGL renderer that combines segmentation output overlayed on video input, using CUDA/OpenGL interoperability.

For more information on GXF extensions, refer to the Clara Holoscan SDK User Guide.

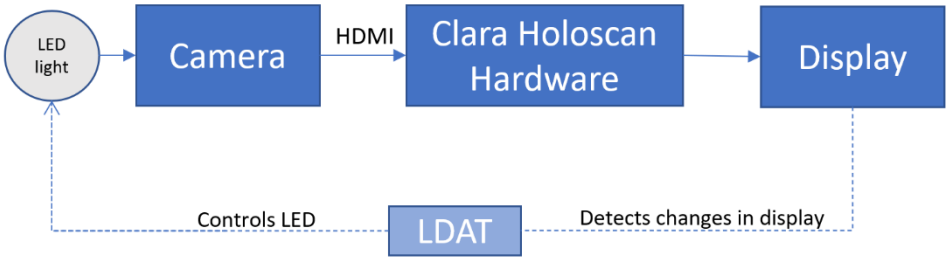
Latency Comparison of Sample AI Applications

The NVIDIA Latency Display Analysis Tool (LDAT) measures the photon-to-display latency, from the camera capturing light, to the camera transmitting data to Clara Holoscan, to Clara Holoscan running processing and compute workloads, and finally to rendering to the display.

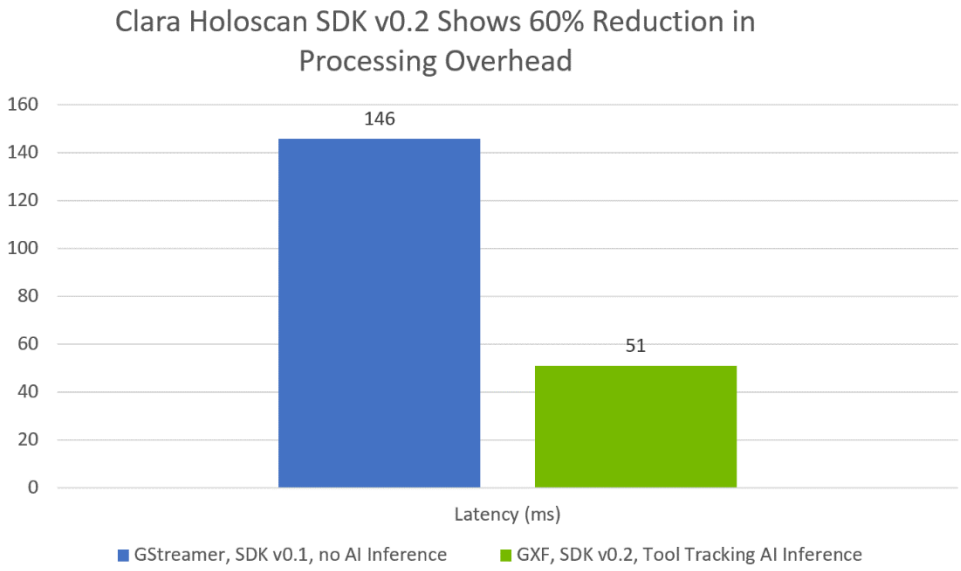
Setup 1: Baseline Setup



Setup 2: Clara Holoscan Hardware Setup



GXF reduces the overhead in an AI Inferencing application by more than 60% compared to a similar GStreamer pipeline in Holoscan v0.1. The measured latency for different configurations can be seen in the table below. On the same hardware (Clara AGX Developer Kit), with the same connection to an AJA KONA video capture card, the latency for the GXF pipeline with AI inference for surgical tool tracking (51 ms) is significantly smaller than the latency for a GStreamer pipeline with no AI inference (146 ms). Note that all latency measurements are the result of the original average latency subtracted by the average baseline latency of 74 ms.



Clara Holoscan SDK Version	Software Pipeline	Hardware Connection	AI Inference	Latency
N/A	N/A	Camera to Monitor directly via HDMI cable	No Inference	Avg latency: **74** ms (This is the baseline to be subtracted)
V0.1	GStreamer	Camera to AJA Kona HDMI	No Inference	Avg latency: **146** ms
V0.1	GStreamer	Camera to AJA Kona HDMI	Resnet18 based Object Detection model	Avg latency: **170** ms
V0.2	GXF	Camera to AJA Kona HDMI	Tool Tracking model	Avg latency: **51** ms

The 146 ms latency for the v0.1 GStreamer pipeline was obtained using Jetpack 4.5 with Holoscan SDK v0.1:

```
$ gst-launch-1.0 ajavideosrc mode=1080p60-rgba input-mode=hdmi nvmm=true ! m.sink_0
nvstreammux name=m width=1280 height=720 batch-size=1 ! nvdsosd ! nv3dsink
sync=false
```

The 170 ms latency for the v0.1 GStreamer pipeline was obtained using Jetpack 4.5 with Holoscan SDK v0.1:

```
$ gst-launch-1.0 ajavideosrc mode=1080p60-rgba input-mode=hdmi nvmm=true ! m.sink_0
nvstreammux name=m width=1280 height=720 batch-size=1 ! nvinfer config-file-
path=/opt/nvidia/clara-holoscan-sdk/clara-holoscan-deepstream-
sample/build/DeepstreamJetsonAGXSample/dsjas_nvinfer_config.txt ! nvdsosd ! nv3dsink
sync=false0.
```

The 51 ms latency for the v0.2 Tool Tracking model was obtained by running within the runtime Docker container with Jetpack 5.0 - HP and Holoscan SDK v0.2:

```
cd /opt/holoscan_sdk/tracking_replayer &&
./apps/endoscopy_tool_tracking/run_tracking_replayer
```

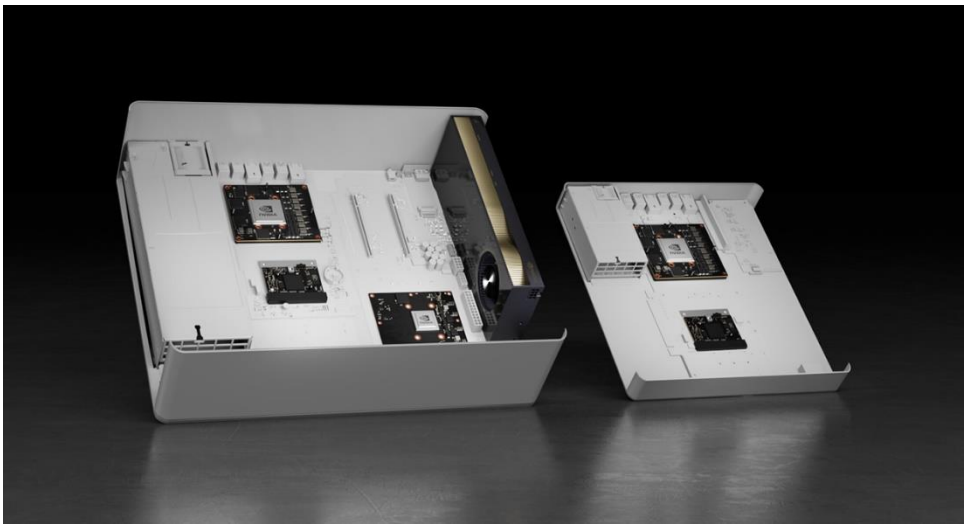
The following table is a detailed look at each of the components in the GXF application on the Clara AGX Developer Kit with Clara Holoscan SDK v0.2. These values are generated using the same reference application

(`apps/endoscopy_tool_tracking/run_tracking_replayer`) within the runtime docker container.

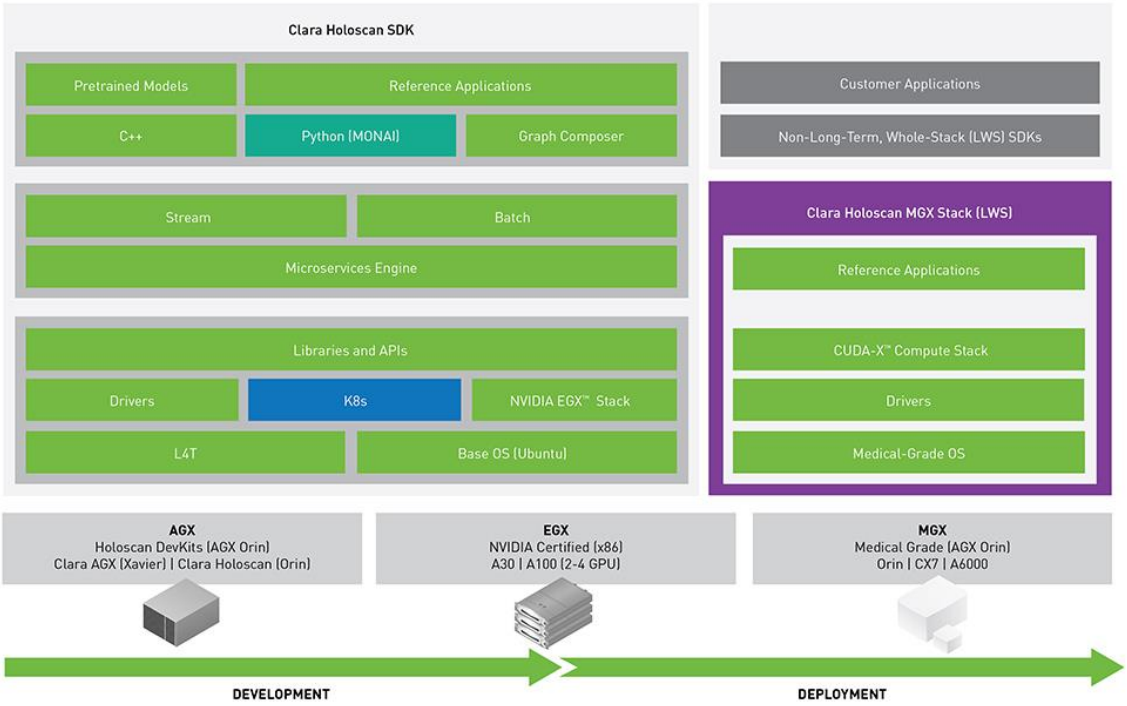
	Time (Median) [ms]	Load (%)
visualizer	1.72	18.8 %
lstm_inference	4.29	24.8 %
format_converter	0.40	2.5 %
broadcast	0.17	1.1 %
visualizer_format_converter	0.40	2.5 %
source	8.92	50.3 %
Total	15.90	

Path to Production

If you develop with the Clara Holoscan SDK and Clara Developer Kits, you can deploy to the medical-grade MGX, a production hardware and software platform optimized to deliver AI to the medical edge.



The NVIDIA Clara Holoscan MGX platform is built with a high-performance NVIDIA AGX Orin™ module, a safety, security and management module, extensive input/output (IO), and expansion slots available for PCIe. This high-performance platform is optimized for low-latency, real-time applications, making it the ideal solution for deploying the next generation of software-defined medical devices. MGX will come with long-life NVIDIA hardware components and documentation to support IEC 60601 certification.



The MGX software will have 10-year long-term support provided by NVIDIA, as well as documentation to support IEC 62304 certification. The core of the MGX stack includes a Linux support package for the MGX hardware, drivers for supported IO devices, NVIDIA AI inference and acceleration libraries, and reference applications. Customer-specific software components and applications can be added at the top of the MGX core stack.

MGX will also include a board management controller (BMC) and software stack for remote system management and software updates, a secure boot controller that provides external root of trust, and a dedicated safety monitor that supports a real-time operating system (RTOS) and built-in safety monitoring software. Custom safety monitoring applications that are tailored to specific needs can be added.

Get Started with Holoscan Today

Visit the Clara Holoscan product page to learn more about the platform. See the Clara Holoscan developer page to get started on your development journey with the Clara Holoscan and Clara AGX Developer Kits, and learn more about the MGX platform for production hardware and software options.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021. https://doi.org/10.3322/caac.21660

[2] Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 2014;370:1298-306. 10.1056/NEJMoa1309086

[3] Stulberg JJ, Huang R, Kreutzer L, Ban K, Champagne BJ, Steele SR, Johnson JK, Holl JL, Greenberg CC, Bilimoria KY. Association Between Surgeon Technical Skills and Patient Outcomes. JAMA Surg. 2020 Oct 1;155(10):960-968. doi: 10.1001/jamasurg.2020.3007. Erratum in: JAMA Surg. 2020 Oct 1;155(10):1002. Erratum in: JAMA Surg. 2021 Jul 1;156(7):694. PMID: 32838425; PMCID: PMC7439214.

[4] Pham, TC., Luong, CM., Hoang, VD. et al. AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. Sci Rep 11, 17485 (2021). https://doi.org/10.1038/s41598-021-96707-8

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Trademarks

NVIDIA, the NVIDIA logo, are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com