Solution Brief

Data Science/AI Algorithm Development
Video Inferencing

ENERZAi 3D Hand Pose Estimation Model Runs 3.25X Faster on Intel® Xeon® Platinum 8360Y Processor[1]

ENERZAi compressed and optimized OpenVINO™ model runs faster on Intel while providing flexibility to adapt to hardware constraints of edge inference platforms

As sensors and other acquisition devices at the far edge of the network expand, there is greater need to inference data remotely. Inference platforms can be built on different technologies with hardware constraints defined by the inference site, such as mobile devices, edge servers, or specialized appliances. One such use case for remote inference is with hand pose estimation to determine the physical view of a hand in real time, such as pointing, signaling, or sign language.

3D hand pose estimation using neural networks and artificial intelligence (AI) can open new use cases and interactive applications. These applications include VR/AR, video game control, gesture-based system control, driver monitoring systems, and other areas. In 2019, Google presented its hand pose estimation software to turn sign language into speech.[2] Over the years, other researchers have presented model approaches that recognize imagery from RGB cameras and RGB plus depth cameras (RGBD).[3]

Today, most 3D hand pose estimation models rely on dedicated depth-sensing cameras and/or specialized hardware support to handle both the high computation and memory requirements for real-time inference. Such requirements hinder the practical application of these models on mobile devices or in other embedded computing contexts.

ENERZAi (enerzai.com) designs tools and solutions to compress and optimize neural network models. The company's automated model compression and low-level optimization toolkit delivers best-in-class edge AI models. The models can be optimized for inference platforms based on different technologies: microcontroller unit, CPU, and application processer (AP).

ENERZAi engineers recently created a 3D hand pose estimation solution designed for Intel architecture that delivered up to 3.25X performance improvements for inference.[1] Inference on an Intel CPU allows clients to migrate from expensive GPU-based platforms to lower-cost edge AI or data center deployment. The solution also enabled flexibility to choose between speed, memory, and accuracy based on the inference platform resources available and desired outcomes.

3D Hand Pose Estimation

Hand pose estimation requires an AI algorithm to locate each joint of a hand from imagery captured by one or more cameras. Hand pose models have defined 21 different points:

- One wrist
- Three joints per digit
- Five fingernails

Hand pose estimation has been inferenced from RGB frames, but the improvement of depth imaging sensors, such as OAK-D, make it possible to use a depth map. Additionally, depth images can capture the hand pose even in low-light environments, since depth imaging sensors do not use visible light. With depth values, accurate 3D locations of each joint can be estimated through neural network modeling.

Most existing 3D hand pose estimation models are too large to handle live video streams with high frame rates (FPS) or manage multiple requests in a short time. Hardware with large computational capacity—such as expensive GPUs—could be adopted to meet the required throughput. But such an approach drives up the cost of edge inference, and it may not be possible to support high-power platforms in remote locations or mobile devices.

The model could be significantly downsized and trained in the same way as the larger model. However, accuracy degradation could occur due to the low capacity of the small model.

ENERZAi presented a compressed and optimized model solution using their expertise and tools.

ENERZAi Hand Pose Model Compression and Optimization

ENERZAi has experience in compressing and optimizing client models, with results of compressing an original model size to .05435 percent (1/1840) of its original size and accelerating inference time by 160x, while maintaining 99.3 percent of the model accuracy.[4] The model allowed the customer to migrate inference from a high-end GPU platform to an Intel Core™ i5 processor-based platform.

For 3D hand pose estimation, ENERZAi engineers used the company's proprietary tools to compress and optimize an original model (AWR: Adaptive Weighting Regression for 3D Hand Pose Estimation, AAAI, 2020[2]). It was CPU-optimized for 3rd Generation Intel Xeon Scalable processors. Besides the model architecture, ENERZAi designed new training methods to minimize the accuracy degradation, which could come with the compressed models.

ENERZAi compression and optimization tools include the following:

- **EdOptimizer—**performs a high level of code optimization with minimal time and cost to easily convert AI models into a C library for inference.
- **EdLite—**automatically compresses AI models while minimizing accuracy loss.

Optimized Performance on Intel Xeon Platinum 8360Y Processor

The ENERZAi 3D hand pose estimation model was implemented with PyTorch, compressed, and then converted to an ONNX model. ENERZAi compressed the model for three different outcomes: latency-oriented compression, balanced (or normal), and accuracy-oriented compression. The model was then optimized for running with a low memory footprint and faster execution time—without compromising accuracy—on a CPU.

The Intel oneAPI toolkit includes the Intel Distribution for OpenVINO toolkit, which provides tools to convert an ONNX model to OpenVINO. Advanced quantization technology was used to achieve low memory utilization and faster execution times using Intel Deep Learning Boost (Intel DL Boost). Results of the performance testing are shown in the following charts.

The compressed and optimized model delivered 3.25X faster performance (figure 1).[1] The ability to compress for different outcomes shows the flexibility of ENERZAi tools, while achieving significant improvements on 3rd Gen Intel Xeon Scalable processors (figure 2).

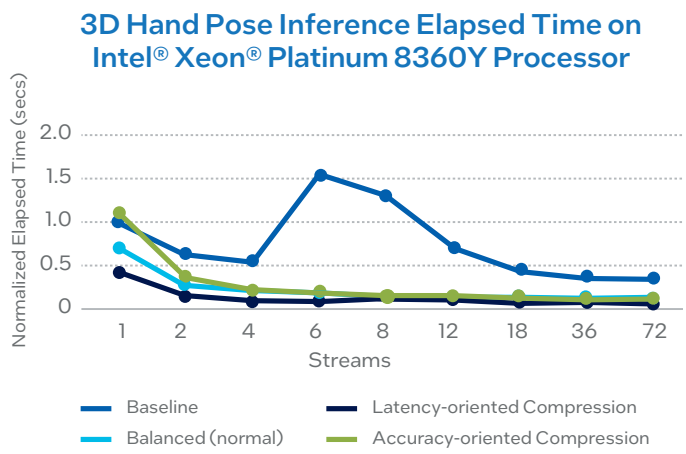

Figure 1. Running on Intel Xeon Platinum 8360Y processor the 3D hand pose estimation model achieves 3.25X higher throughput inference using 72 streams.[1] The CPU utilization is approximately 90 percent, with the vectorization reaching 99 percent.

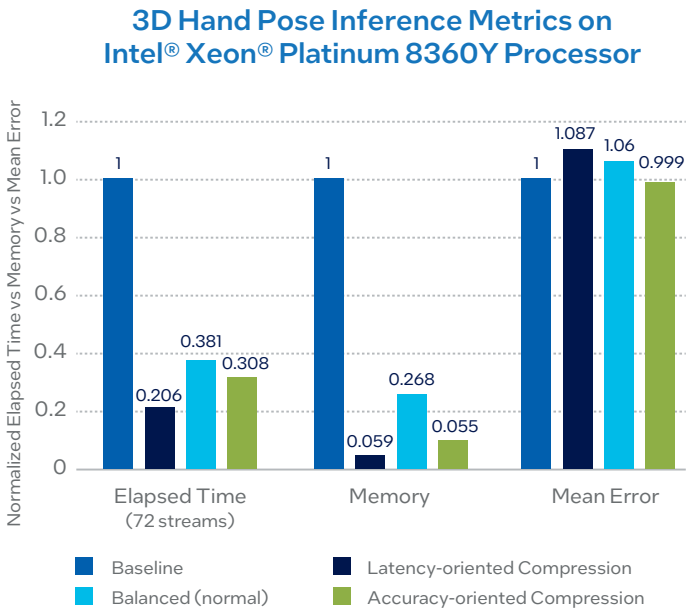

Figure 2. The flexibility to choose between speed, memory, and accuracy allows deployment to wide variety of hardware. Migration from GPU to CPU at edge is a possibility with the performance observed.

Conclusion

ENERZAi provides compression and optimization expertise and tools for AI models. ENERZAi tools have achieved significant model size reduction without impacting accuracy, enabling inference speedup.

The ENERZAi compressed 3D hand pose estimation model runs best on OpenVINO-supported hardware. Thus, for on-premises data center or cloud inference, the model runs well on 3rd Gen Intel Xeon Scalable processors as shown above. The compressed model ran 3.25X faster on Intel Xeon Platinum 8360Y processor with OpenVINO acceleration.[1] This acceleration can help reduce hardware costs by allowing more parallelized model deployments.

The model can also be deployed on OAK-D camera platforms with stereo depth cameras for depth image sensing. OAK-D contains the Intel Movidius™ Myriad™ X Visual Processing Unit (VPU) with OpenVINO support.

ENERZAi technology provides flexibility to choose between speed, memory and accuracy, which enables AI on edge devices that have strong hardware constraints. Furthermore, since the technology does not compromise accuracy while minimizing inference time and memory, it supports clients who want to migrate from GPU to cost-efficient hardware platforms, including Intel CPUs.

For more information about ENERZAi, visit **enerzai.com**

Learn more about the Intel AI Builders program at **builders.intel.com/ai**

intel AI Builders

ENERZAi provides best-in-class Edge AI technology to overcome limitations of cloud-based artificial intelligence such as latency, network connection, operating costs, and privacy. ENERZAi have developed its own compression recipe and optimization resources to achieve improved inference on various technologies.

[1] **CONFIG:** Test by Enerzai as of Apr/2022. 1-node, 2x Intel® Xeon® Platinum 8360Y CPU @ 2.40GHz Processor, 36 cores HT On Turbo ON Total Memory 256 GB (16 slots/16GB/3200 MT/s DDR4), BIOS: SE5C6200.86B.0022.D64.2105220049 (ucode: 0xd0002b1), Ubuntu 20.04.3 LTS, 5.4.0-91-generic, openvino (2022.1), torch (1.11.0), Inference Framework: OpenVINO (2022.1), 3D Hand Pose Estimation: Adaptive Weighting Regression (AWR), customer data, 1 instance/2 socket, Datatype: INT8, Images: 30000, Size: 1x1x128x128, Streams: 72, Batch Size: 1.

[2] https://www.bbc.com/news/technology-49410945.

[3] https://arxiv.org/pdf/2007.09590.pdf.

[4] See enerzai.com: Cloud-To-Edge Migration Project With Global Top Medical AI Venture.

0822/AU/HBD/PDF ♻Please Recycle 352608-001US

intel