Qualcomm AI Research Explores 3D Perception using Advanced ML

The trick of AI at the edge is reducing complexity while maintaining or even improving inference accuracy. And that takes a lot of primary research.

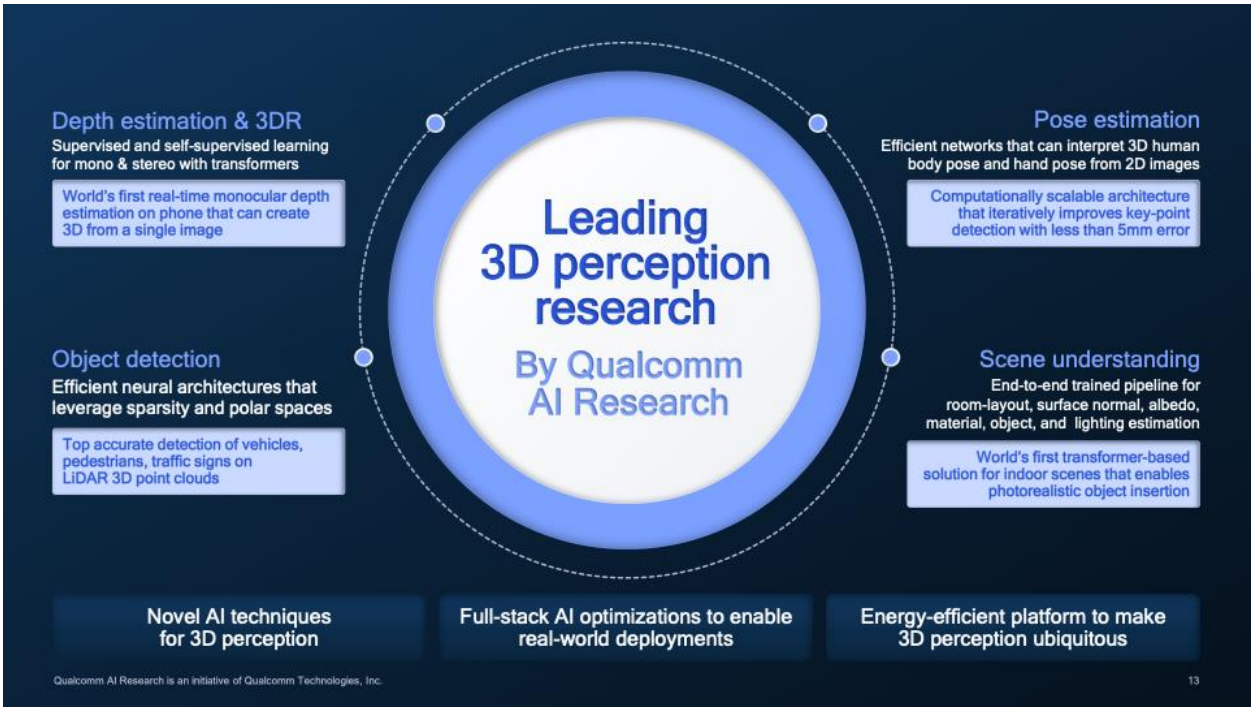
When we explored the eight "AI Firsts" from Qualcomm AI Research earlier this year, it was clear to us the company's full-stack approach to AI ensured that many of the innovations under investigation would translate to superior platforms for 5G, the edge, and the data center. Now the organization is tackling the next big challenge: 3D Perception.

One of the thorniest problems mobile and edge platforms wrestle with is achieving 3D perception on 2D images or 3D point clouds using an energy-efficient platform like Snapdragon. Now the company is the first to demonstrate 3D perception proof-of-concepts on edge devices and is working to solve system and feasibility factors to move from research to commercialization. This article explores the problems and solutions addressed by Qualcomm AI Research to evolve 3D perception.

The Benefit of 3D over 2D perception

3D perception offers many benefits over 2D, as a 3D point cloud is a more reliable model of reality and provides confident cues for object detection and scene recognition, which is essential in robotic and autonomous vehicles. 3D can also provide accurate size, pose, and motion estimation as well as enable realistic rendering from light and radio frequency signals.

Qualcomm AI Research has initially focused on 4 areas of 3D perception: depth estimation, pose estimation, object detection, and scene understanding. In all four areas, described in Figure 1, Qualcomm AI Research has applied novel AI techniques, such as using transformer models, with an eye toward power-efficient real-world deployment by considering the full stack optimizations, not just the hardware to accelerate the computation. We would note that this is consistent with the recently announced Qualcomm AI Stack and expect these innovations to be added to that framework.



Implementing these concepts is technically difficult mathematically, but especially so in the thermal- and power-constrained environments Qualcomm creates for mobile and edge devices. Consequently, the research team must constantly find ways to do computation more efficiently while maintaining accuracy. This is especially true when applying transformers, which are incredibly powerful but are also computationally hungry models built on massive networks of parameters. Useful techniques here include exploitation of sparsity, dealing with incomplete data sets, quantization, and self-supervised training of AI models.

Depth Estimation

Self-supervised training is used in estimating depth from unlabeled monocular videos utilizing geometric relationships across video frames. Qualcomm AI Research has developed a novel transformer architecture that leverages “spatial self-attention” for depth estimation, with a smaller model that runs in real time and impacts only the training process, requiring no additional inference computations. In fact, this technique has resulted in a model that is 26 times smaller and runs in real-time on the Snapdragon’s Hexagon Processor.

Beyond monocular videos, Qualcomm AI Research has developed stereo depth estimation models for increased accuracy, similar to human visual perception. These stereo techniques enable real-time estimation on a phone on today’s Hexagon

Processor with greater generalizability, increased precision, and over 20-times faster than the current state of the art (SOTA).

Object Detection

Enabling efficient and accurate 3D object detection is critical for understanding the physical world. Qualcomm AI Research has developed a transformer-based architecture that reduces latency and memory usage without sacrificing accuracy. And the model can function without requiring a complete and expensive 360 LiDAR scan. The remarkable results show a model that is more accurate, faster, and uses a fraction of the parameters of SOTA models in use today, all while consuming far less computation and energy.

Model	Parameters (M)	GFLOPs	Inference time (ms)	Accuracy (AP)
PointRCNN	2.20	25	620	90.34
PV-RCNN	13.10	69	80	92.24
ComplexYolo	65.50	31	19	75.32
PointPillars	1.43	32	16	88.36
Ours	0.59	6	14	94.70

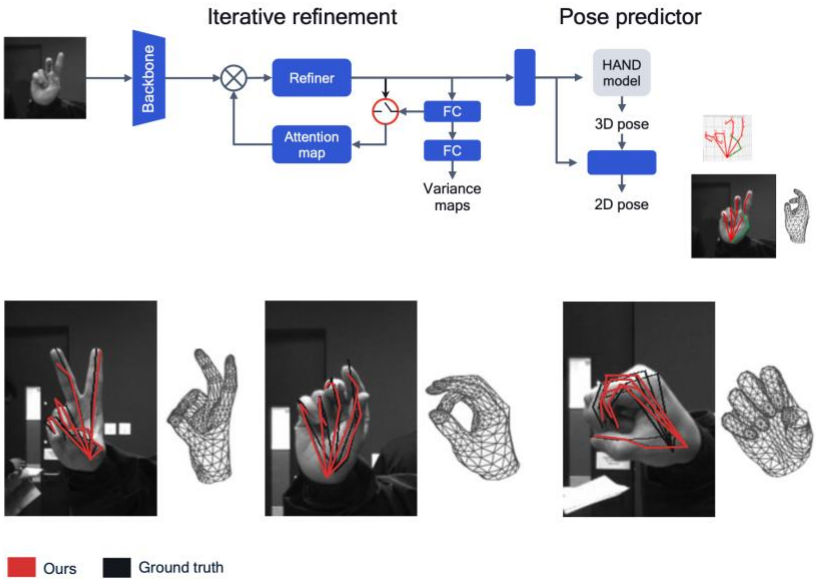
Pose Estimation

Inferring human pose, whether it be head, body, or hands, is important for many 3D applications. Interestingly, Qualcomm AI Research has developed a lightweight architecture that recursively refines estimation of hand poses in real time without the need for precise hand detection or massive computation. The technique incorporates attention and gating for dynamic refinement, again delivering a power and memory-efficient algorithm. This approach could be used in augmented or virtual reality experiences as well to someday facilitate communications between the deaf, translating American Sign Language to text, or audio for those who can hear. And one can imagine all this running on a Snapdragon-based edge device.

Dynamic refinements to reduce size and latency for hand pose estimation

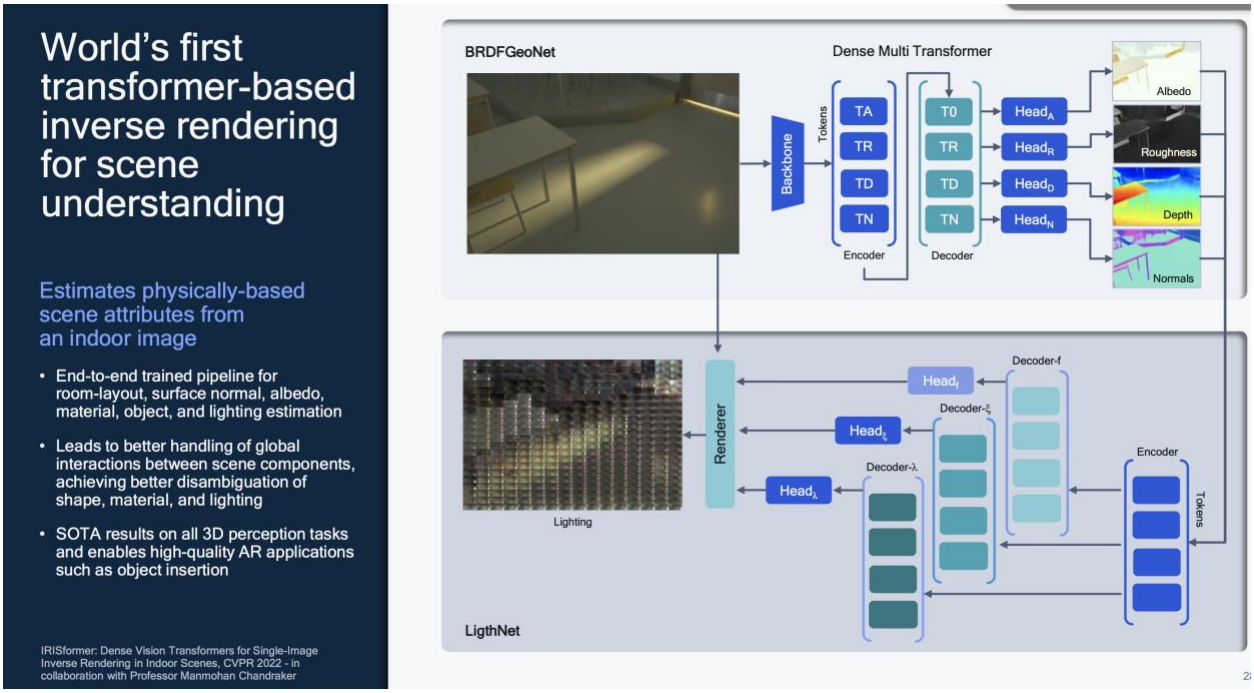
Lightweight architecture: applies recursively while incorporating attention and gating for dynamic refinement

Eliminates the need for precise hand detection



SCENE UNDERSTANDING

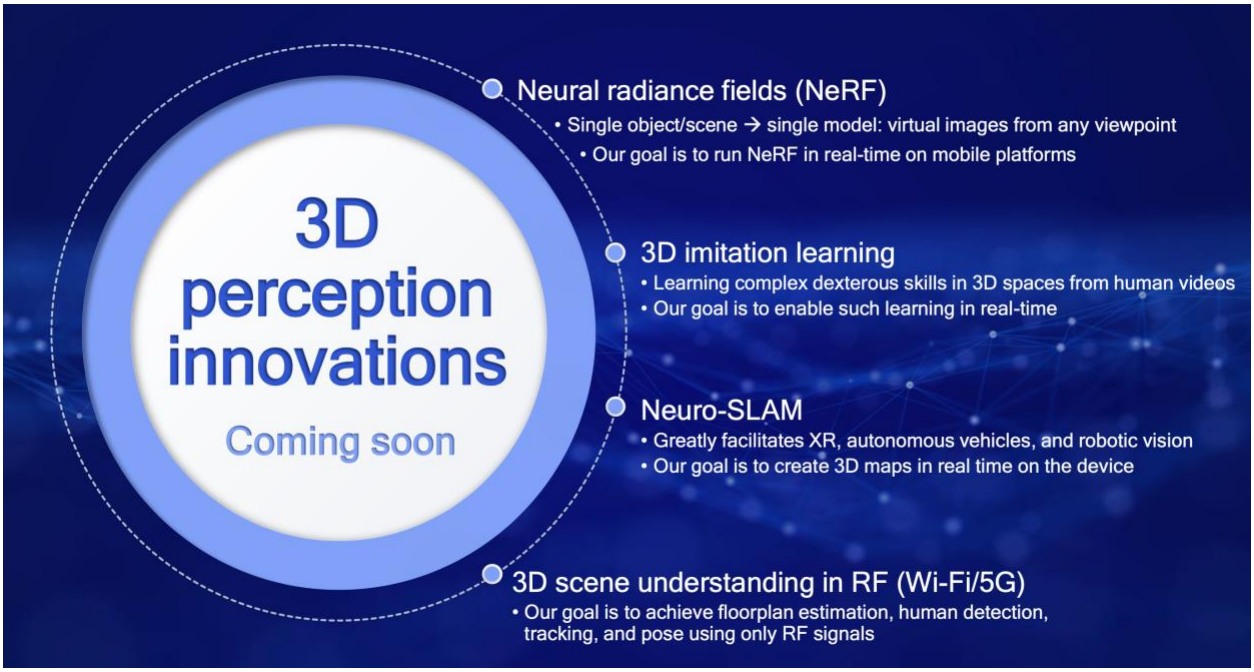
Scene understanding decomposes a scene into its 3D and physical components. In yet another industry first, Qualcomm AI Research has developed another transformer-based model that can use “inverse rendering” to estimate scene attributes from a 2D image. This model could be used to create meta data about a scene, such as room layout, surfaces, albedo, materials, objects, and lighting estimation. This could lead to better interactions between scene components, disambiguating shapes, materials, and lighting. This SOTA 3D scene understanding could enable high-quality AR applications, such as 3D object insertion into a real-world scene without spatial conflicts.



WHAT COULD BE NEXT?

Qualcomm AI Research is engaged in a menu of interesting projects and looks forward to sharing performance advancements in the future. One of the more intriguing and valuable of these is called Neural Radiance Fields (NeRFs). NeRFs can be used to create virtual images from any viewpoint, an astonishing scene at which we have all marveled when watching TV coverage of a football game. But instead of programming an FPGA to accomplish this visually amazing feat, Qualcomm believes their approach can run in real time on a Snapdragon-based mobile device.

In other endeavors, Qualcomm is researching "3D imitation learning" to enable robots to someday be able to mimic a dexterous motion using a video of a human maneuver. Neuro-SLAMs could be another innovation, enabling 3D mapping of interior spaces to facilitate XR, autonomous vehicles, and robotic vision with real-time 3D maps. And finally, Qualcomm AI Research is working on 3D scene understanding using RF signals, which could enable floorplan estimation, human detection, tracking, and pose using only RF signals.



CONCLUSIONS

These new 3D perception techniques go far beyond simply satisfying an engineer's penchant for creativity and curiosity. 3D perception lies at the very heart of solving critical real-world problems from autonomous vehicles and more efficient factories to saving lives on the operating table. These approaches are excellent examples of useful creativity being applied at Qualcomm AI Research to be at the heart of the connected intelligent edge enabled by smart devices that we can all afford and use in intuitive ways. We know of no other company who is tackling these problems as quickly and as effectively as Qualcomm AI Research.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR: Karl Freund, Founder Cambrian-AI Research

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be mentioned in the context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This document was developed with Qualcomm Technologies, Inc. (QTI) funding and support. Although the paper may utilize publicly available material from various vendors, including QTI, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The views expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and

forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.