Qualcomm

July 19, 2022

3D perception:

Cutting-edge AI research to understand the 3D world

Qualcomm Technologies, Inc.

@QCOMResearch

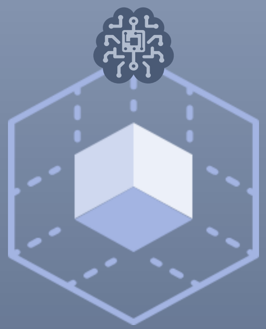
Today's agenda

- Advantages of 3D perception over 2D
- The need for 3D perception across applications
- Advancements in 3D perception by Qualcomm AI Research
- Future 3D perception research directions
- Questions?

We perceive the world in 3D

3D perception offers many benefits and new capabilities over 2D

- 3D structure is more reliable than a 2D image for perception
- 3D provides confident cues for object and scene recognition
- 3D allows accurate size, pose, and motion estimation
- 3D is needed for rendering by light and RF signals

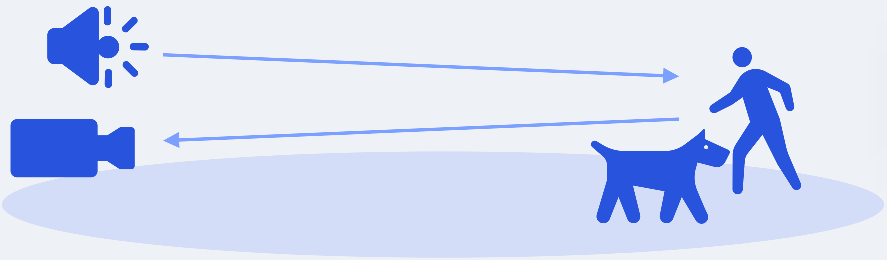


3D data acquisition

Multiple methods offering different benefits

Active sensing

Illuminating and reconstructing 3D from reflected back signals



Acquisition Methods

Light: Time-of-flight, structured light, LiDAR

RF: Radar, SAR, THz imaging

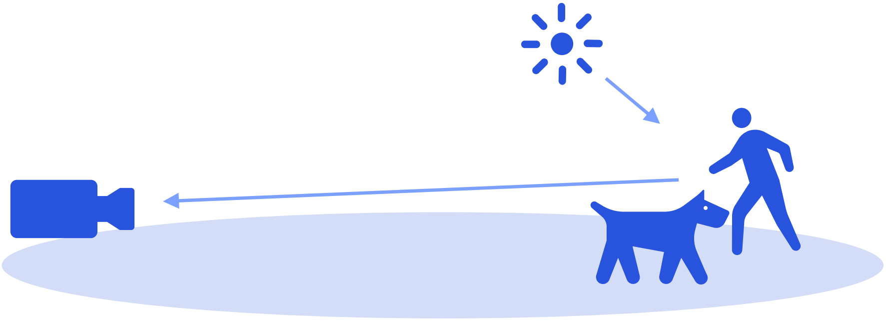
Other: CT scan, MRI, ultrasound, …

Benefits

Works in dark and various conditions

Passive sensing

Determining 3D from signals emitted by an uncontrolled illuminator



Computational Methods

Geometry: Traditional CV using multiple cameras (DFS, MVS)

Inference: Via shadow, blur, and motion parallax cues

Regression: From appearance to 3D (AI driven)

Benefits

Lower cost and lower power

DFS: Depth From Stereo; MVS: Multi-View Stereo; SAR: Synthetic-Aperture Radar

3D perception enables diverse applications that make our lives better

XR

Autonomous vehicles

IOT

Camera

Mobile

3D perception greatly facilitates immersive XR

6-DoF: 6 degrees-of-freedom

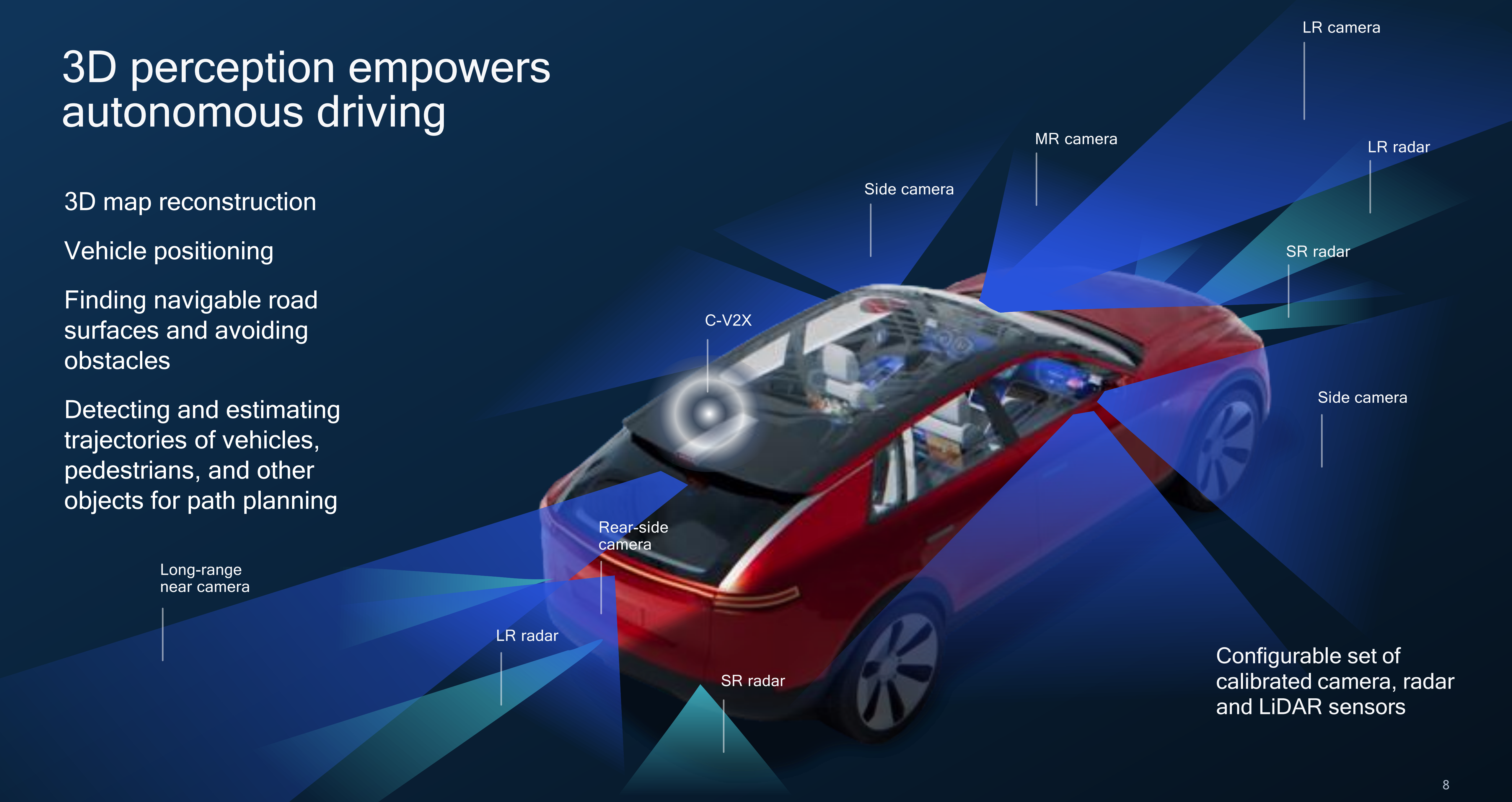
3D perception empowers autonomous driving

3D map reconstruction

Vehicle positioning

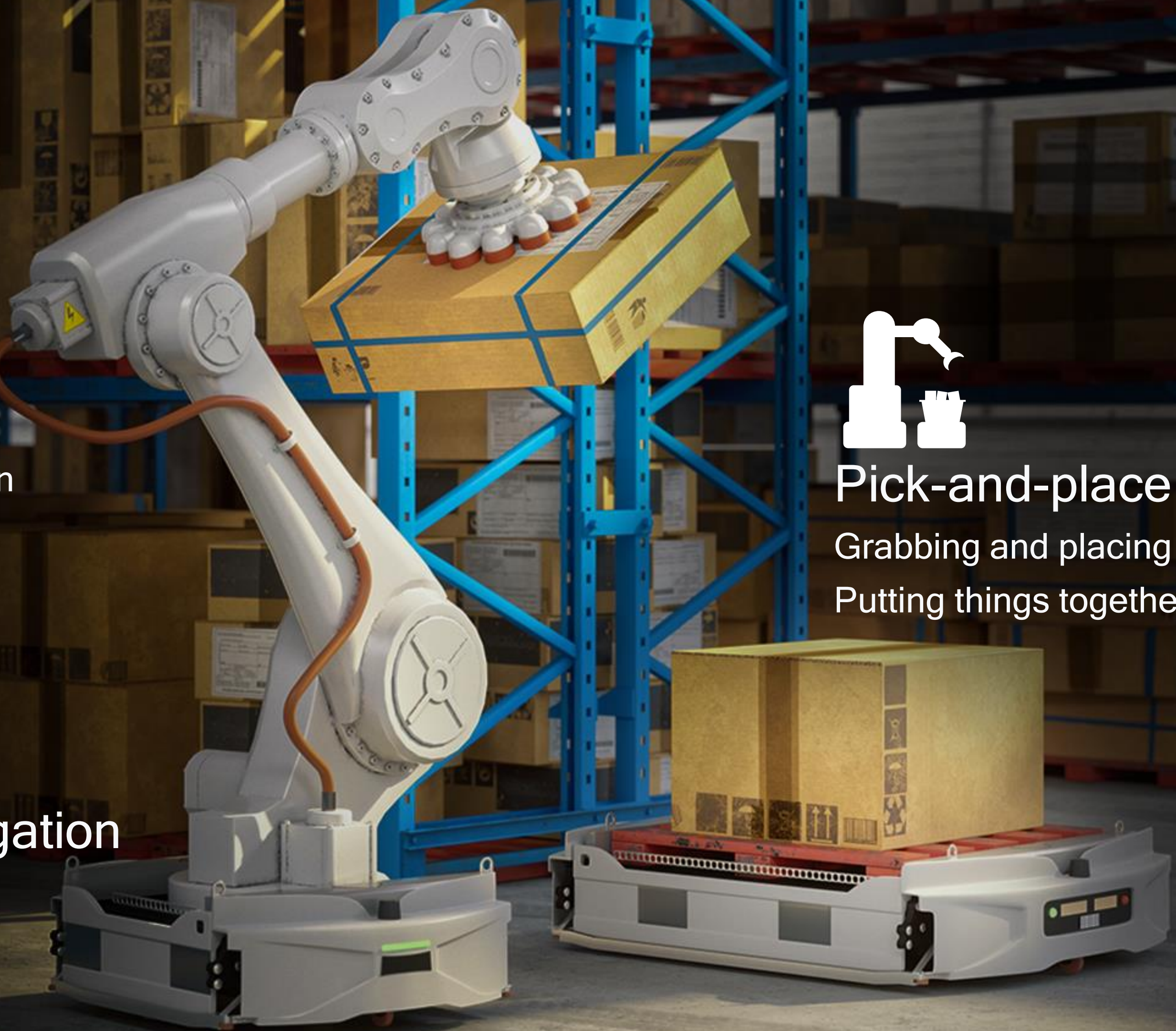
Finding navigable road surfaces and avoiding obstacles

Detecting and estimating trajectories of vehicles, pedestrians, and other objects for path planning



Configurable set of calibrated camera, radar and LiDAR sensors

3D perception is important for robotics



Situational awareness

Scene and human interaction

Object, scene, and motion recognition

Pick-and-place and assembly

Grabbing and placing objects

Putting things together

Motion planning and navigation

Obstacle avoidance

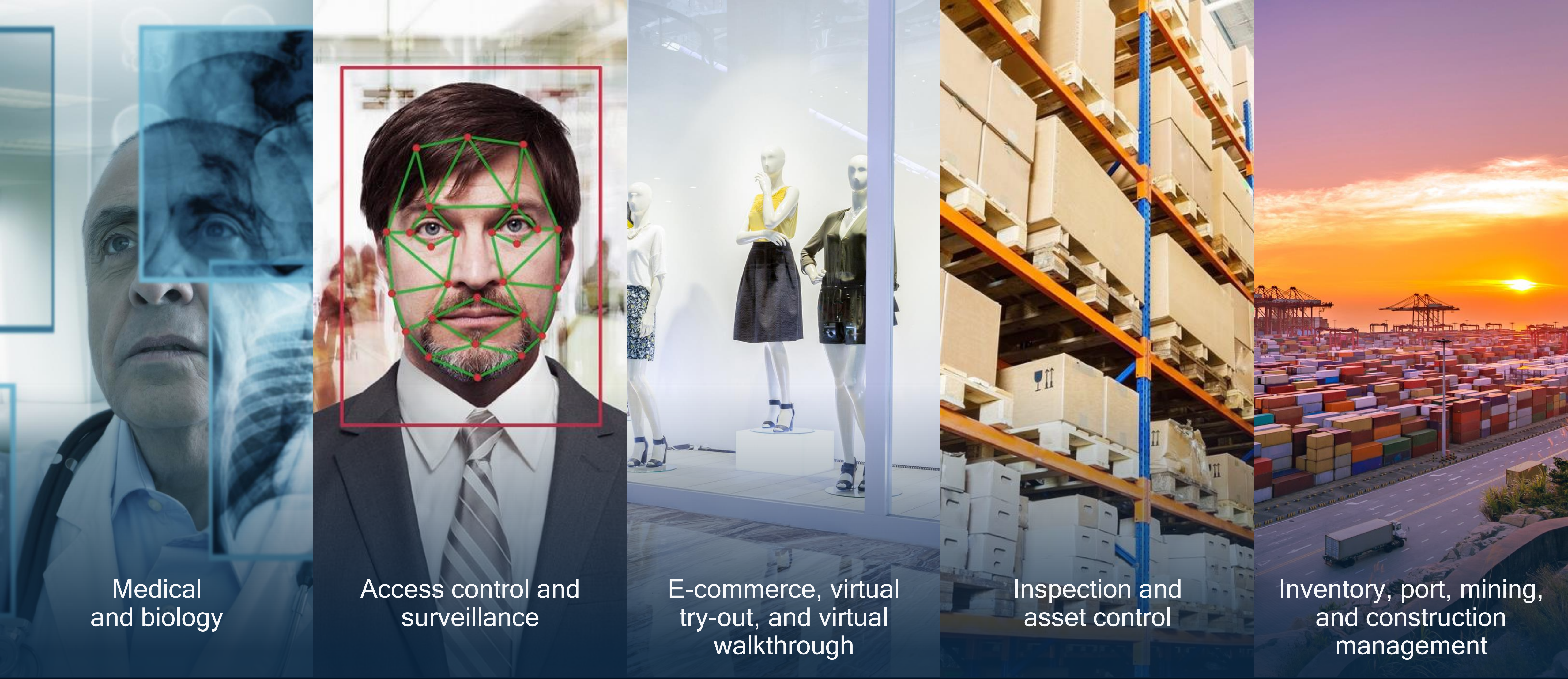
Delivery and factory automation

3D perception vastly improves computational photography and cinematography

Image quality
Denoising, deblurring, relighting, HDR, etc.

Filtering effects
Bokeh, depth-of-field, etc.

Content editing
Seamless object removal, beautification, etc.



3D perception is highly valuable in many more areas

3D perception

is crucial for understanding the 3D world

Depth estimation & 3DR
Creating 3D models of scenes and objects from 2D images

Pose estimation
Finding orientation and key-points of objects

Object detection
Finding positions and regions of individual objects

Scene understanding
Decomposing a scene into its 3D and physical components

Neural radiance fields
Networks to synthesize virtual views

Neuro SLAM
AI for simultaneous localization & mapping

3D imitation learning
Learning 3D robotics tasks from human videos

3D scene in RF
Scene understanding using RF signals

3DR: 3D Reconstruction

Leading 3D perception research

By Qualcomm AI Research

Depth estimation & 3DR

Supervised and self-supervised learning for mono & stereo with transformers

World's first real-time monocular depth estimation on phone that can create 3D from a single image

Object detection

Efficient neural architectures that leverage sparsity and polar spaces

Top accurate detection of vehicles, pedestrians, traffic signs on LiDAR 3D point clouds

Pose estimation

Efficient networks that can interpret 3D human body pose and hand pose from 2D images

Computationally scalable architecture that iteratively improves key-point detection with less than 5mm error

Scene understanding

End-to-end trained pipeline for room-layout, surface normal, albedo, material, object, and lighting estimation

World's first transformer-based solution for indoor scenes that enables photorealistic object insertion

Novel AI techniques for 3D perception

Full-stack AI optimizations to enable real-world deployments

Energy-efficient platform to make 3D perception ubiquitous

3D perception challenges

Data challenges

- Sparse vs volumetric nature of 3D point cloud
- Incompleteness in 3D acquisition
- Availability of high-quality 3D video datasets

Implementation challenges

- Computational load (training/inference)
- HW/SW platform (memory, SDKs, tools)
- Manipulation, viewpoint management

Image pixels are arranged on a uniform grid,
while 3D point cloud faces accessibility vs memory trade-off

Enabling AI-based self-supervised depth estimation from a single image

No need for annotated data

- Self-supervised learning from unlabeled monocular videos
- Utilizes geometric relationship across video frames

Builds on semantics[1]

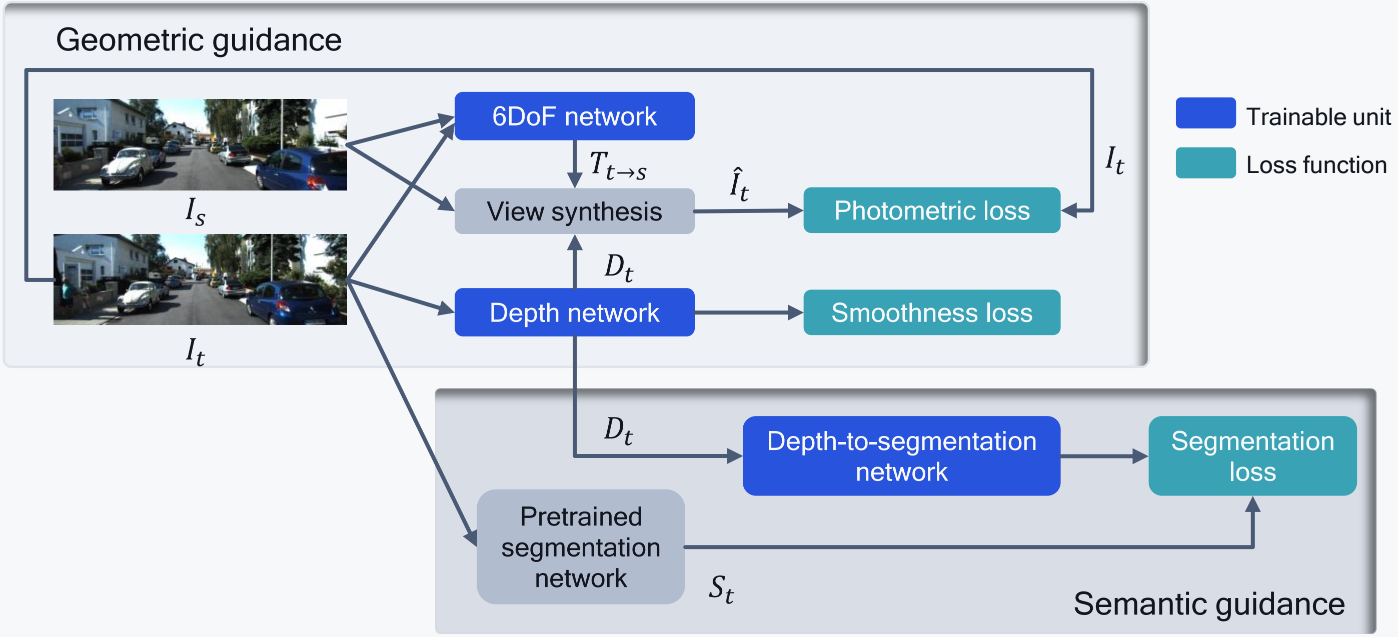
- Significantly improves accuracy by using semantic segmentation[2]

Works on any neural network architecture

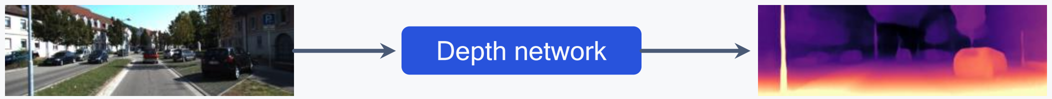
- Modifies only training and requires no additional inference computation

Enables automatic domain adaptation

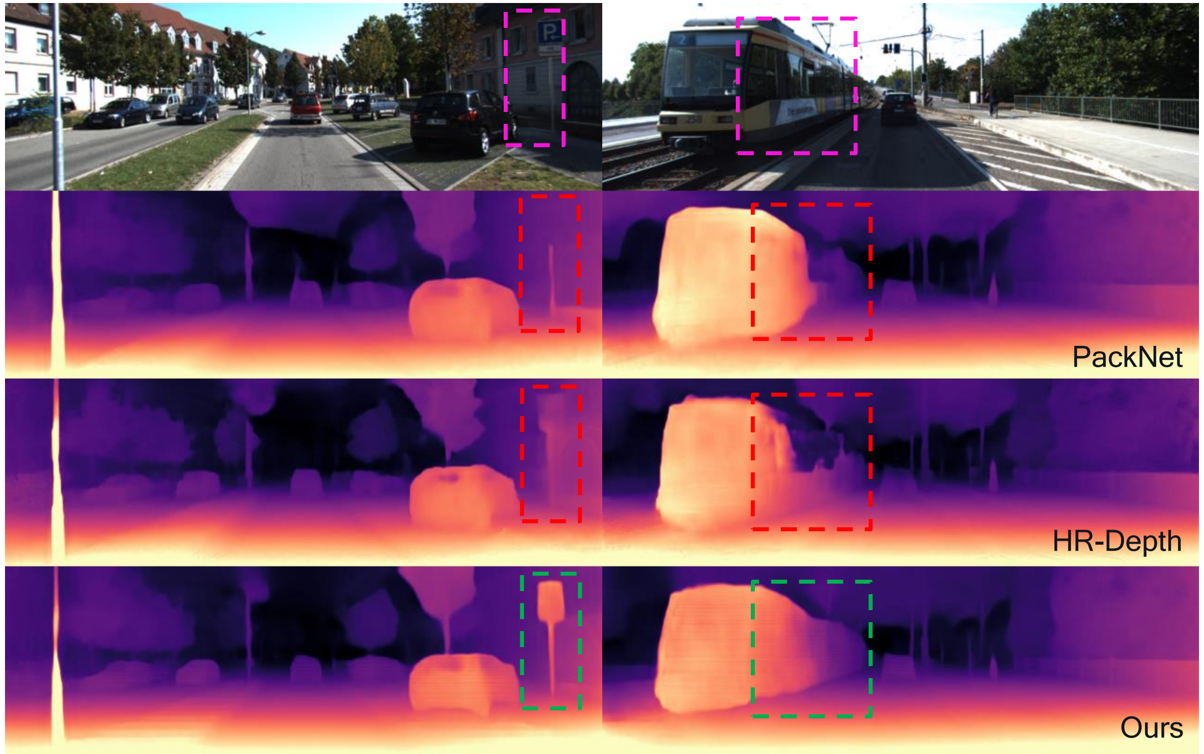
Self-supervised training pipeline of X-Distill

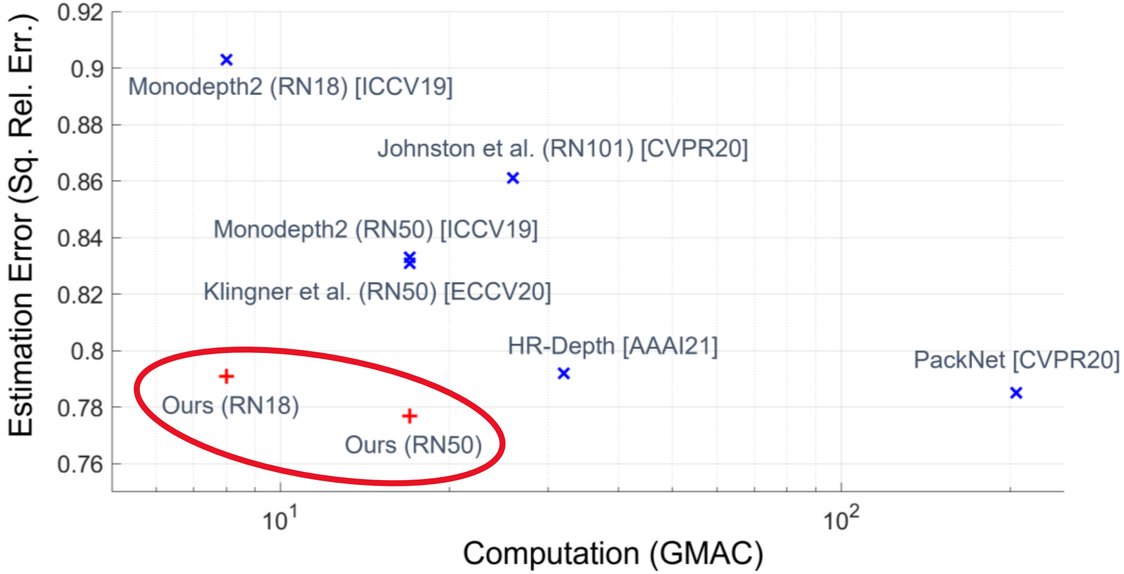


Test pipeline



1: X-Distill: Improving Self-Supervised Monocular Depth via Cross-Task Distillation, BMVC 2021
2: InverseForm: A Loss Function for Structured Boundary-Aware Segmentation, CVPR 2021
Raw video obtained from Cityscapes Benchmark: https://www.cityscapes-dataset.com/





Achieving SOTA accuracy with 90% less computation

Raw video obtained from Cityscapes Benchmark: https://www.cityscapes-dataset.com/

Improved on-device efficiency with neural architecture search

Enabling real-time on-device use cases

Distilling Optimal Neural Architectures (DONNA)[1] → Hardware aware

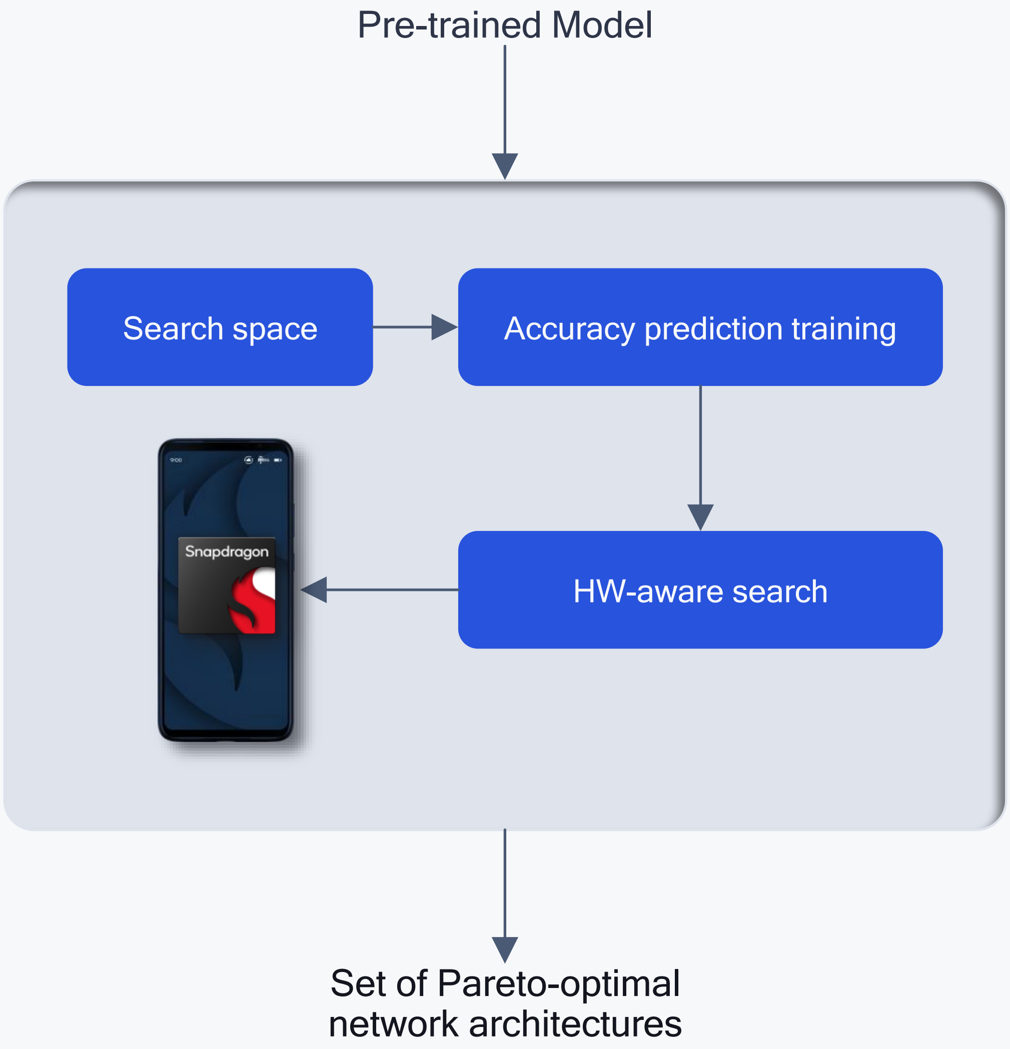
20-40% faster models

Diverse search
Supports diverse spaces with different neural operations to find the best models

Low cost
Scales to many HW devices at minimal cost

Scalable
Low start-up cost of 1000-4000 epochs, equivalent to training 2-10 networks from scratch

Reliable
Uses direct HW measurements instead of a potentially inaccurate HW model



1: Distilling Optimal Neural Networks: Rapid Search in Diverse Spaces, ICCV 2021

Model	Parameters (M)	FPS	Error
Monodepth2 (RN50) Quantized	32.0	23	0.83
Our X-Distill (RN50) Quantized	32.0	23	0.71
Our X-Distill DONNA Quantized[1]	**3.7**	**35**	**0.75**

~10X reduction in parameters

More efficient model

More accurate model

- Quantization through AI Model Efficiency Toolkit (AIMET)
- 30+ FPS by using more efficient backbone optimized via DONNA neural architecture search

Running monocular depth estimation in real time on Snapdragon mobile platform

1: X-Distill/DONNA demo at NeurIPS 2021

AIMET is a product of Qualcomm Innovation Center, Inc. Raw video obtained from Cityscapes Benchmark: https://www.cityscapes-dataset.com/

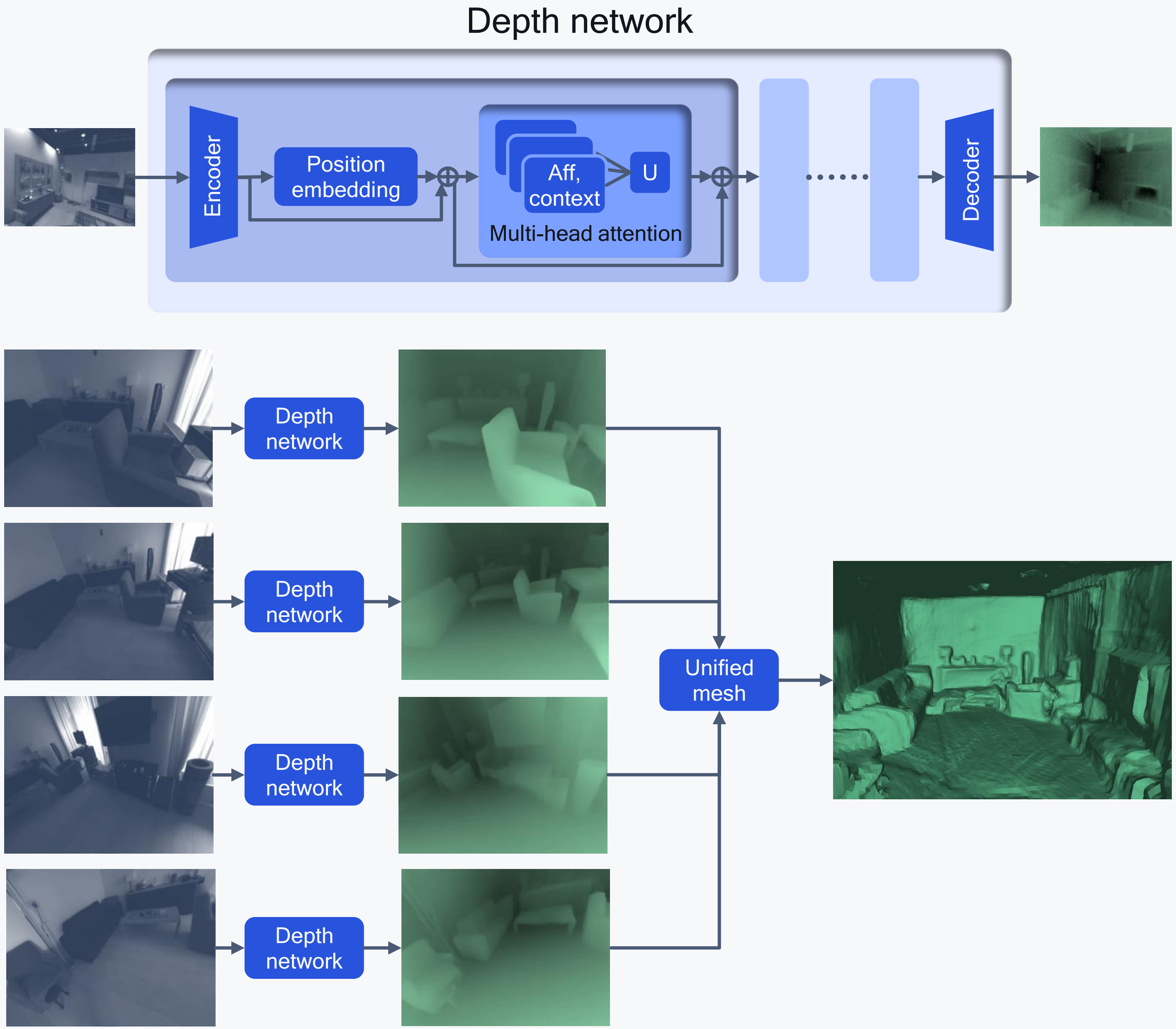
Enhancing the model with a new efficient transformer architecture

Novel transformer architecture leverages spatial self-attention for depth network

- Smaller model that runs in real time
- Improved 3D recall in reconstructed meshes

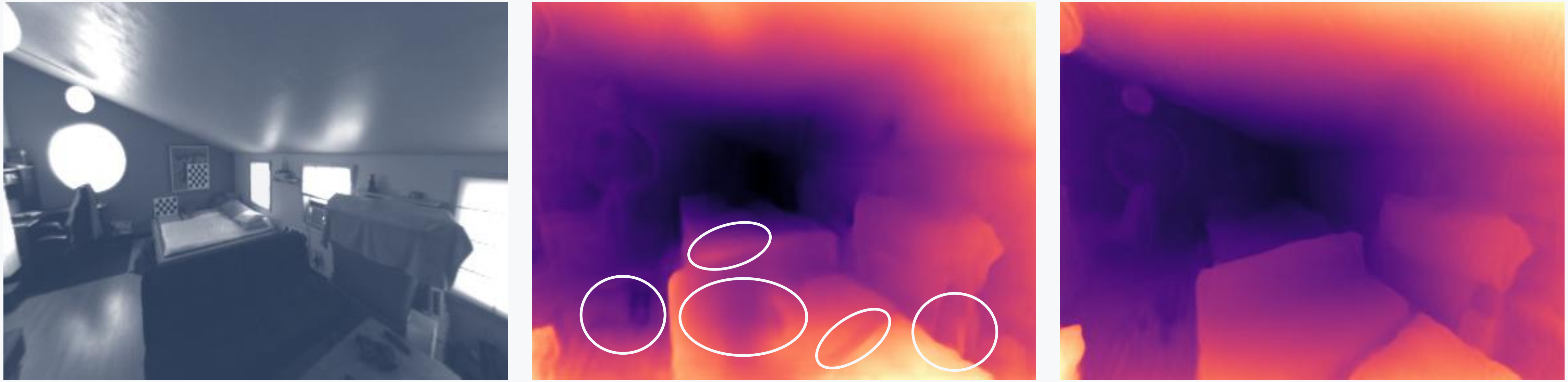
During training:

- Self-supervised learning for fine-tuning transformers to improve temporal consistency and eliminate texture copy
- Sparse depth from 6DoF to resolve global scale issues



6DoF: 6 degrees-of-freedom

Image from XR camera | Monodepth2 (RN34) model | Our transformer-based model

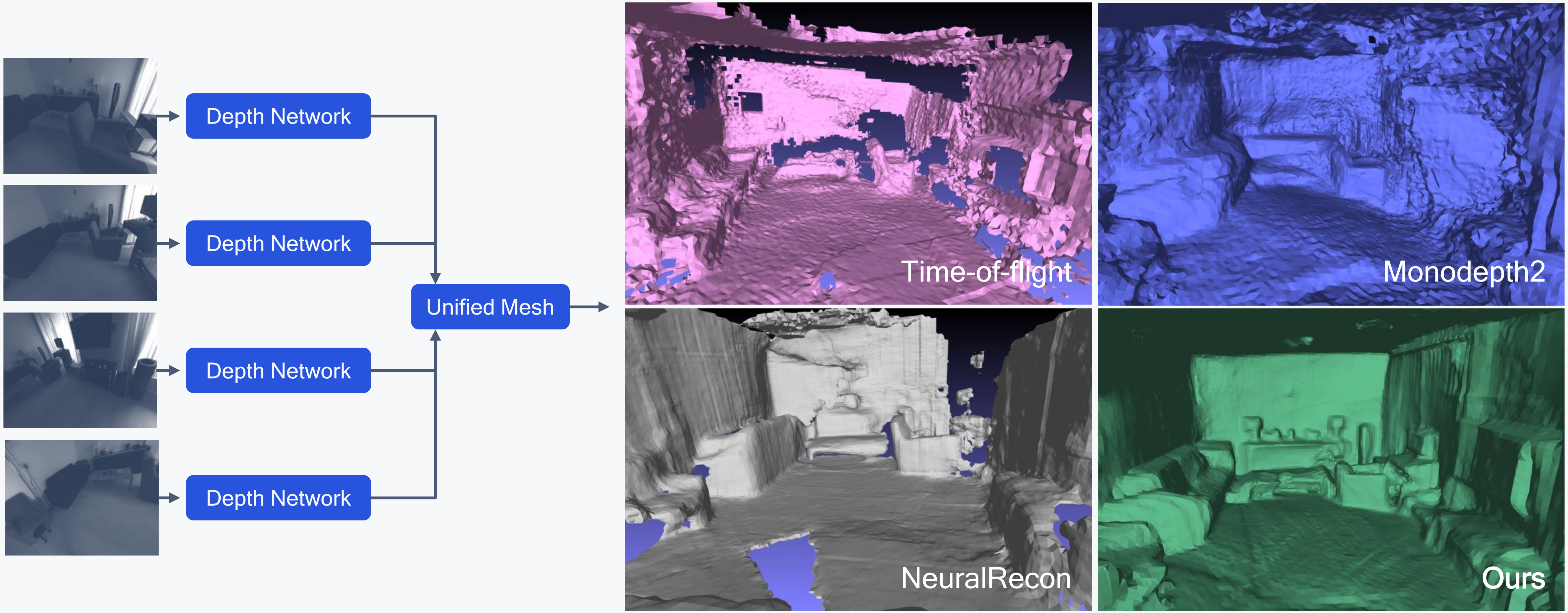


Model	Parameters (M)	3DR recall
Monodepth2 (RN34)	15.0	0.653
DPT (MIDAS, transformer based)	123.0	0.926
Ours (transformer based)	**4.7**	**0.930**

Our transformer-based model is both smaller and more accurate

26x smaller model fits on a phone processor and runs real-time on Qualcomm® Hexagon™ Processor

Qualcomm Hexagon is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.



Our transformer-based model achieves better visual quality than current SOTA

From single image to stereo depth estimation for increased accuracy

Similar to human visual perception

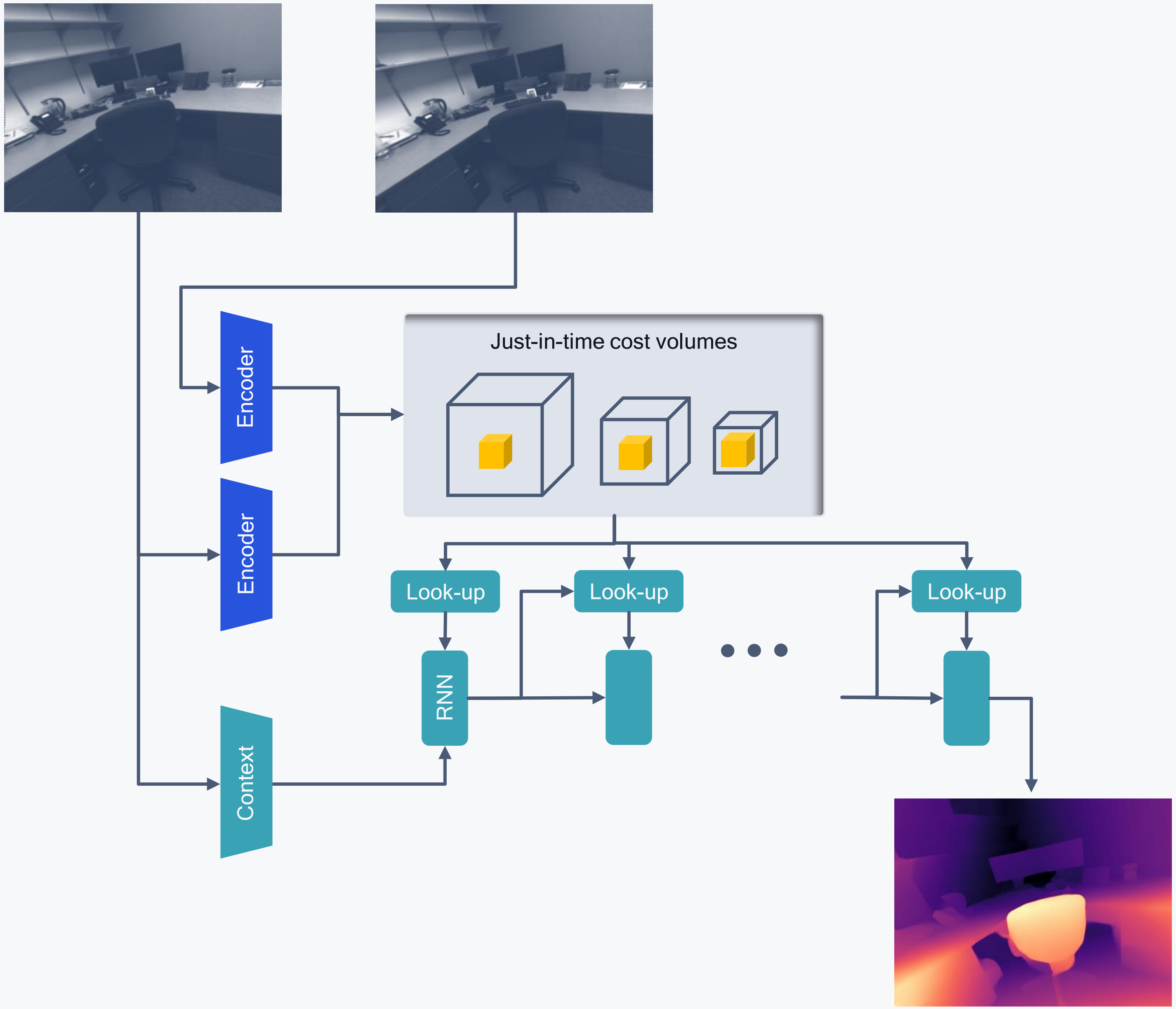
Highly accurate disparity estimation model using two images (left/right)

- Fits into memory by just-in-time computation
- Geometry is much sharper and more detailed
- Greater generalizability

Models leverage Snapdragon's heterogenous computing

- Real-time performance
- Subpixel disparity accuracy

Extends to motion parallax



Left image

Right image

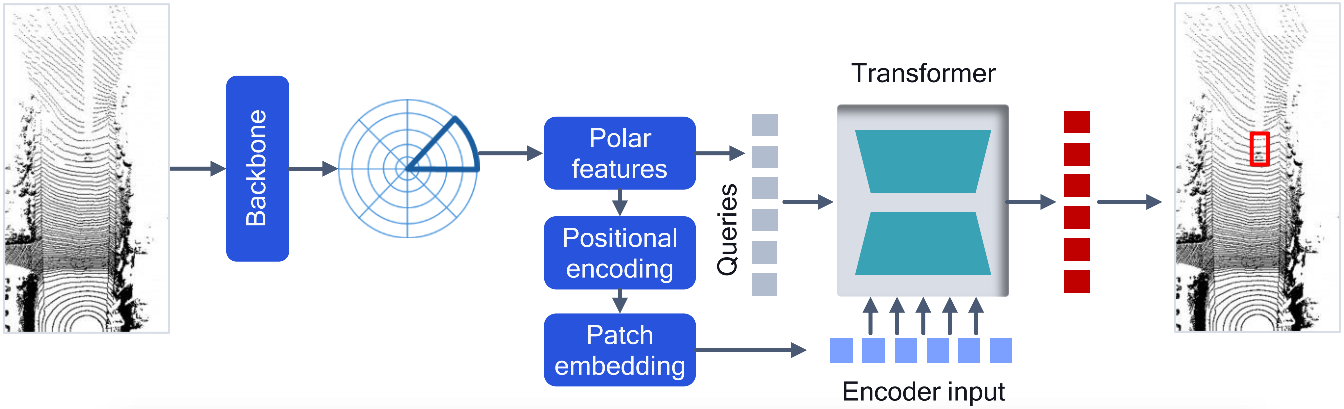
Our JiT stereo depth estimation model runs in real-time
20x faster than SOTA with similar accuracy on Hexagon Processor

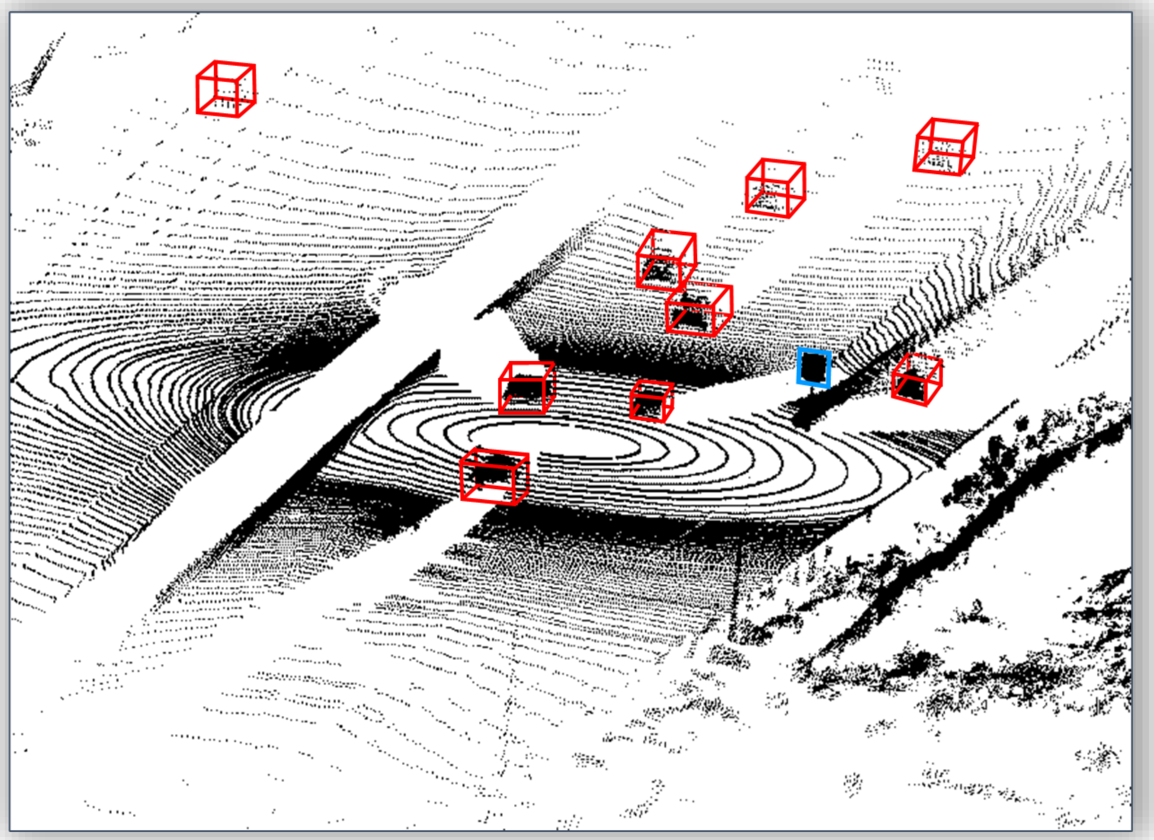
JiT: Just-in-time
SOTA: RAFT-Stereo

Enabling efficient object detection in 3D point clouds

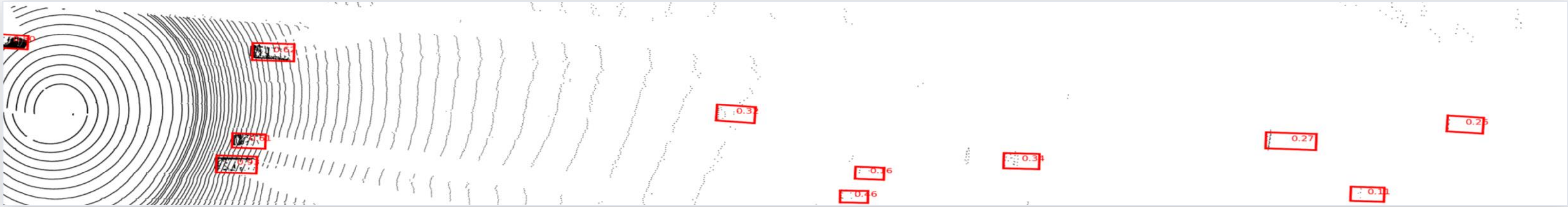
A transformer-based architecture

- Leverages 2D pseudo-image features extracted in the polar space
- Reduces latency and memory usage without sacrificing detection accuracy
- Sectors data for a stream-wise processing to make predictions without requiring a complete 360 LiDAR scan





Fast Polar Attentive 3D Object Detection on LiDAR Point Clouds, NeurIPS MLAD



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

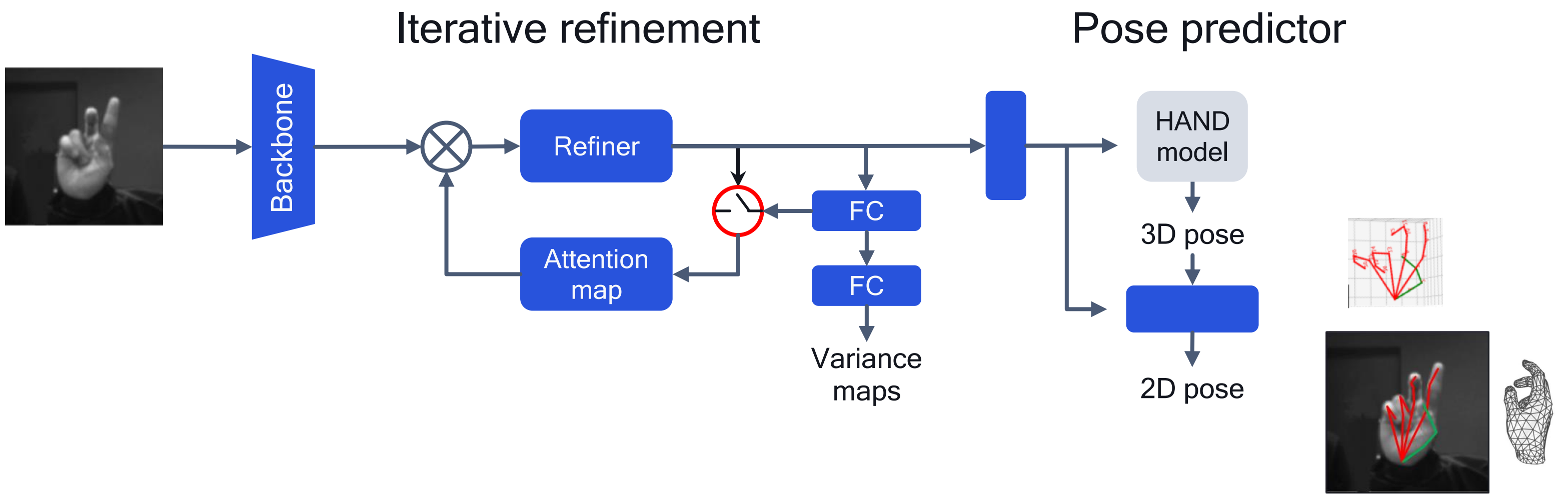
Smaller, faster, lower power model that achieves the top precision

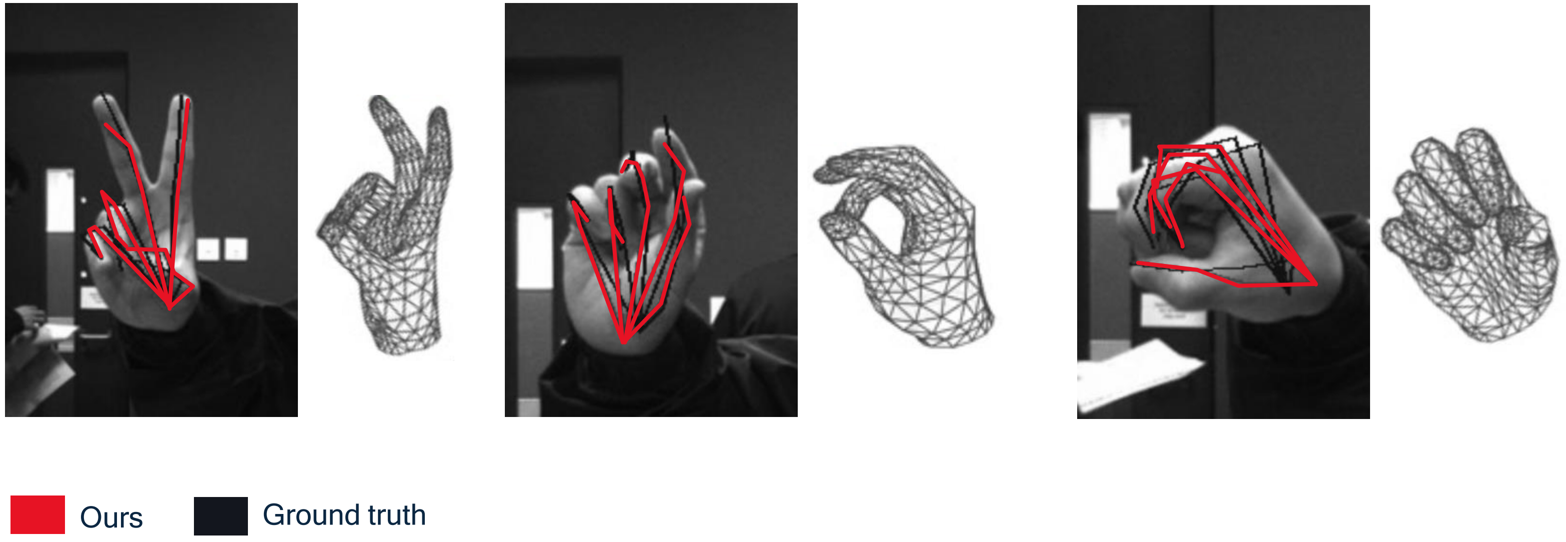
Fast Polar Attentive 3D Object Detection on LiDAR Point Clouds, NeurIPS MLAD workshop, 2021

Dynamic refinements to reduce size and latency for hand pose estimation

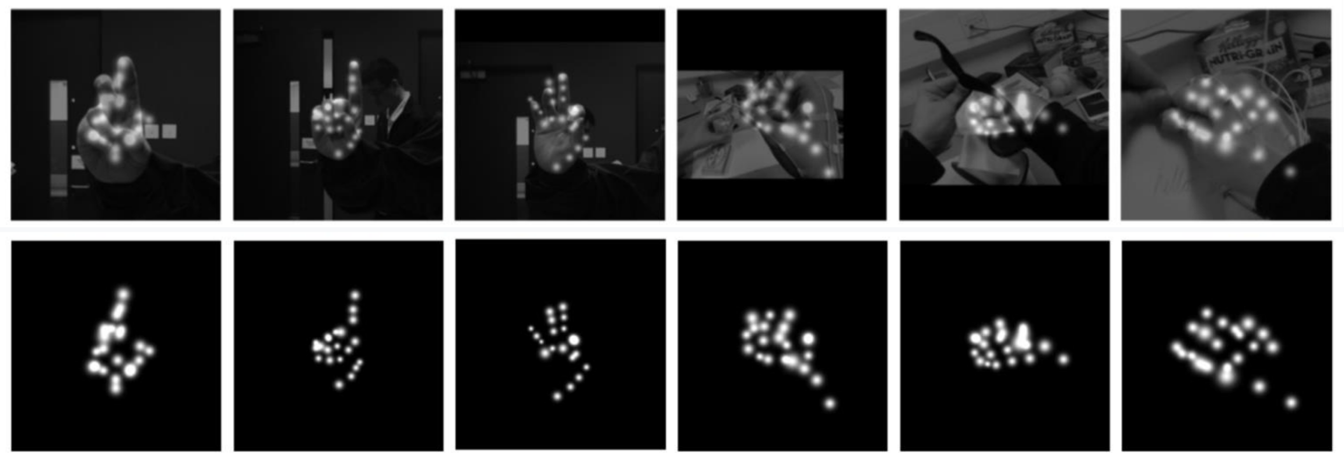
Lightweight architecture: applies recursively while incorporating attention and gating for dynamic refinement

Eliminates the need for precise hand detection





Dynamic iterative refinement for efficient 3D hand pose estimation, WACV 2022



Heatmaps for hand landmark points from our model

Methods	AUC (20-50)	GFLOPs	#Params
Z&B	0.948	78.2	-
Liu et al.	0.964	16.0	-
HAMR	0.982	8.0	-
Cai et al.	0.995	6.2	4.53M
Fan et al.	0.996	1.6	4.76M
Ours	**0.997**	1.3	1.68M

Methods	AUC (0-50)	GFLOPs	#Params
Tekin et al.	0.653	13.62	14.31M
Fan et al.	0.731	1.37	5.76M
Ours	**0.768**	0.28	0.46M

Significant reduction in GFLOPs while achieving better accuracy

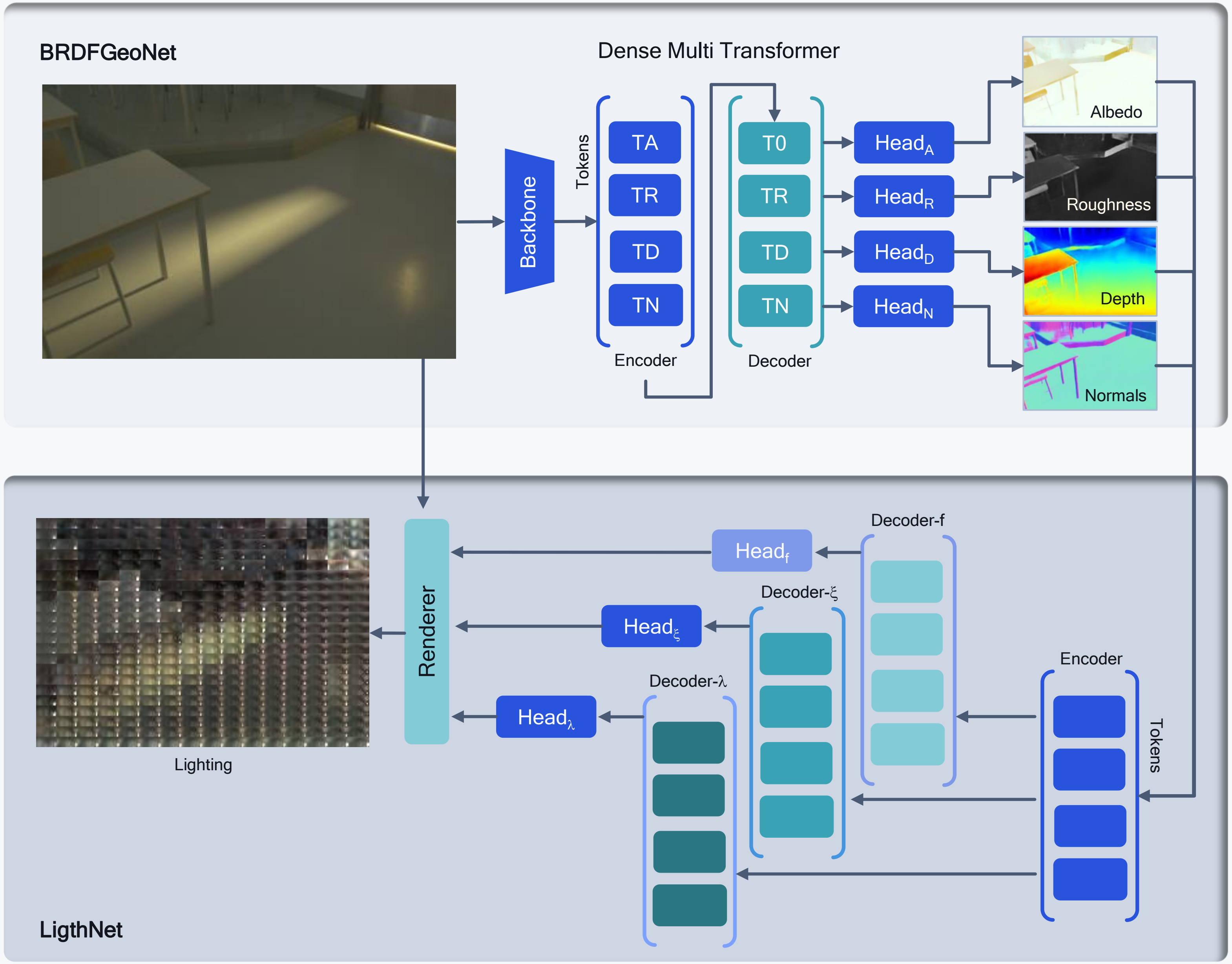
Our method also achieves the best average 3D error
9.76mm (SOTA) → 7.24mm (Ours)

AUC: Area under curve
Dynamic iterative refinement for efficient 3D hand pose estimation, WACV 2022

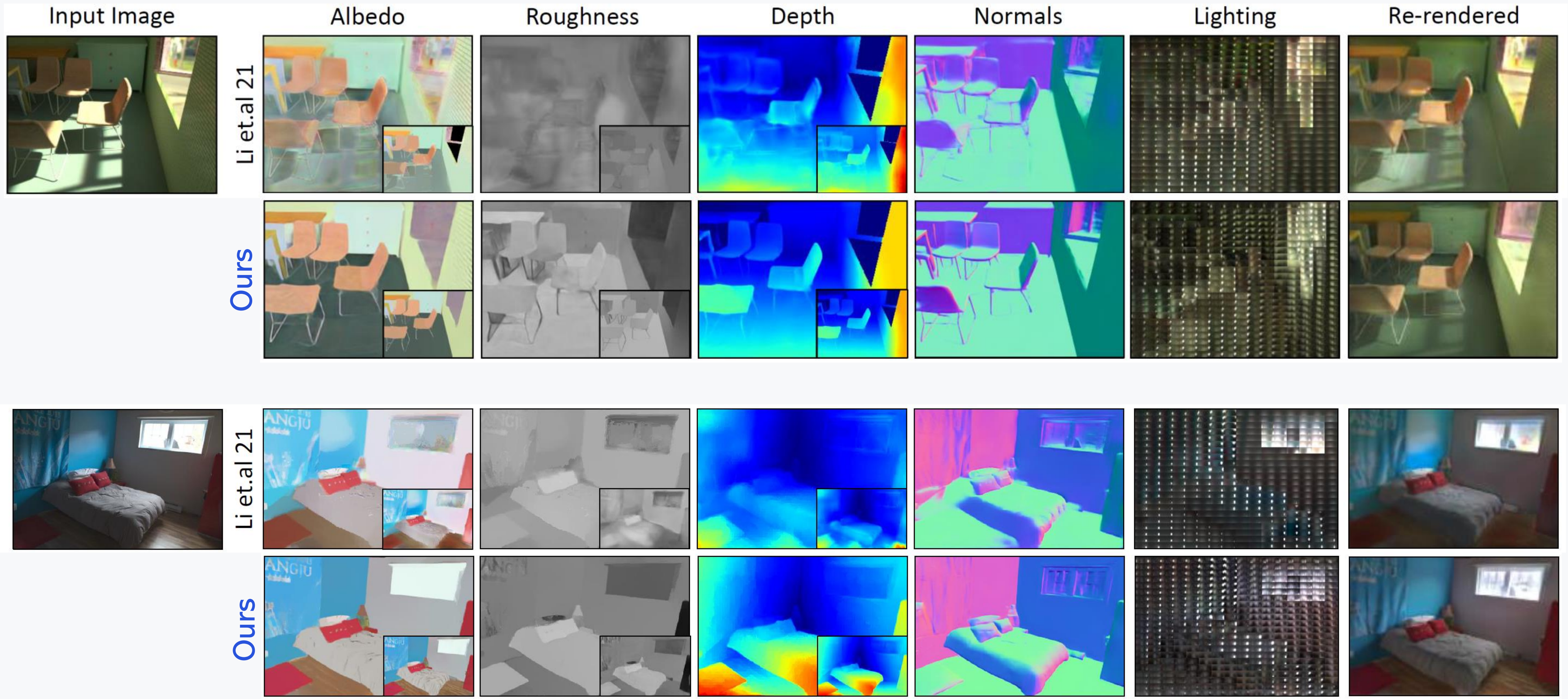
World's first transformer-based inverse rendering for scene understanding

Estimates physically-based scene attributes from an indoor image

- End-to-end trained pipeline for room-layout, surface normal, albedo, material, object, and lighting estimation
- Leads to better handling of global interactions between scene components, achieving better disambiguation of shape, material, and lighting
- SOTA results on all 3D perception tasks and enables high-quality AR applications such as object insertion



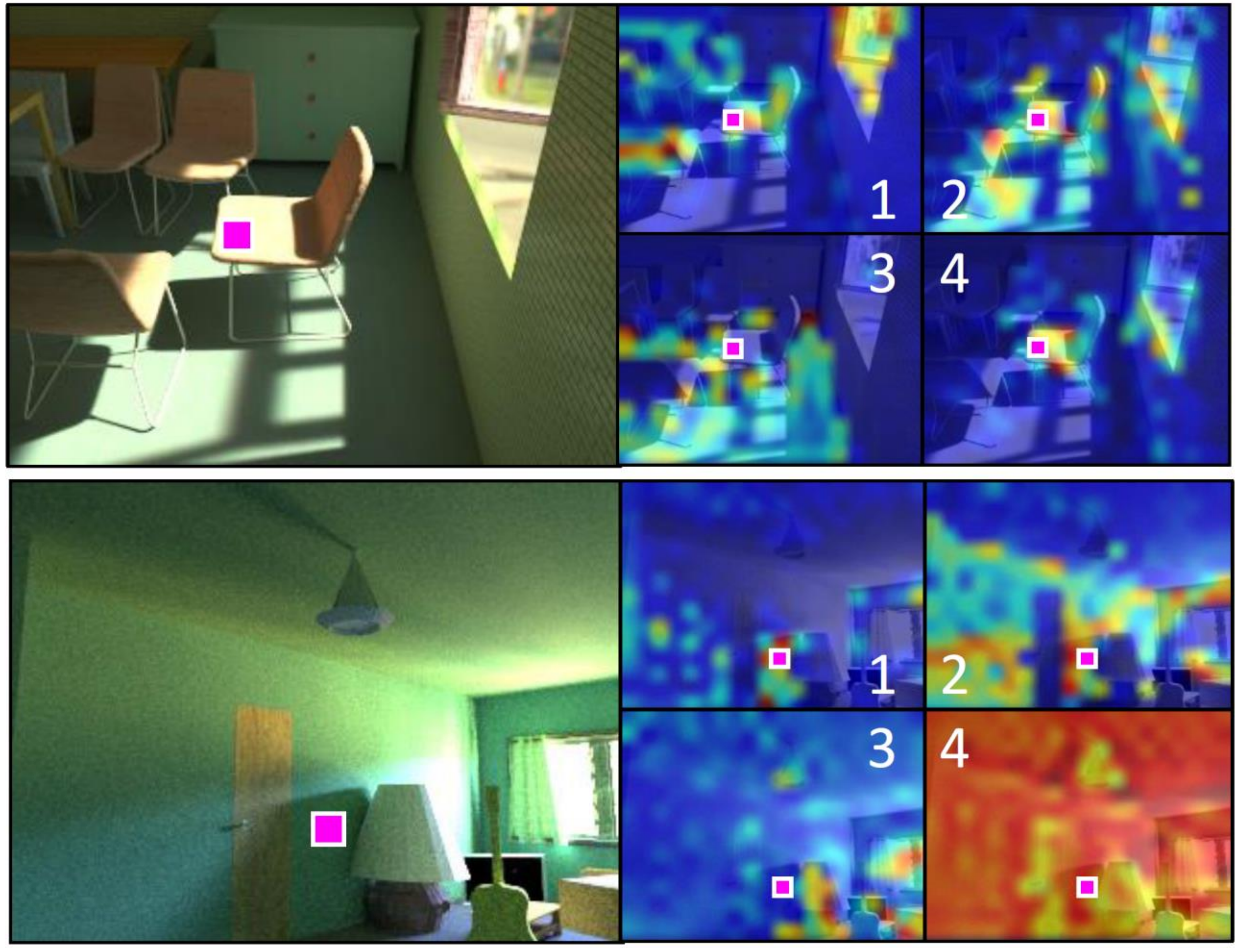
IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes, CVPR 2022 - in collaboration with Professor Manmohan Chandraker



Small insets are refined estimations obtained by further post-processing with bilateral solvers (BS).

Our model finds more accurate scene attributes compared to SOTA

Input images | Attention Maps (Albedo)



Focuses attention on:

1. Chairs and window
2. Highlighted regions over the image
3. Entire floor
4. The chair itself

Focuses attention on:

1. Neighboring shadowed areas of the wall
2. The entire wall
3. Potential light sources and occluders
4. Ambient environment

Transformer algorithm automatically learns attention maps to determine the important areas of the image

Barron et al. 2013	Gardner et al. 2017	Li et al. 2021	Ours

Our method correctly estimates lighting to realistically insert objects

Qualcomm AI Stack

Tools:

AIMET

AIMET Model Zoo

NAS

Model analyzers

Infrastructure:

Prometheus

kubernetes

docker

AI Frameworks

TensorFlow PyTorch ONNX

AI Runtimes

Qualcomm® Neural Processing SDK | ONNX RUNTIME | TF Lite Micro | Direct ML | TF Lite

Qualcomm® AI Engine Direct

Math Libraries | Compilers | Virtual platforms

Profilers & Debuggers | Programming Languages | Core Libraries

System Interface | SoC, accelerator drivers | Emulation Support

android | Zephyr | ubuntu® | CentOS | QNX

Platforms: Smartphones | XR | ACPC | IoT | Robotics | Auto | Cloud

3D perception innovations

Coming soon

Neural radiance fields (NeRF)

- Single object/scene → single model: virtual images from any viewpoint
  - Our goal is to run NeRF in real-time on mobile platforms

3D imitation learning

- Learning complex dexterous skills in 3D spaces from human videos
- Our goal is to enable such learning in real-time

Neuro-SLAM

- Greatly facilitates XR, autonomous vehicles, and robotic vision
- Our goal is to create 3D maps in real time on the device

3D scene understanding in RF (Wi-Fi/5G)

- Our goal is to achieve floorplan estimation, human detection, tracking, and pose using only RF signals

We are conducting leading research to enable 3D perception

Thanks to our full-stack AI research, we are first to demonstrate 3D perception proof-of-concepts on edge devices

We are solving system and feasibility challenges to move from research to commercialization

Questions

Connect with us

www.qualcomm.com/research/artificial-intelligence

www.qualcomm.com/news/onq

@QCOMResearch

www.youtube.com/c/QualcommResearch

www.slideshare.net/qualcommwirelessevolution

Thank you

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog