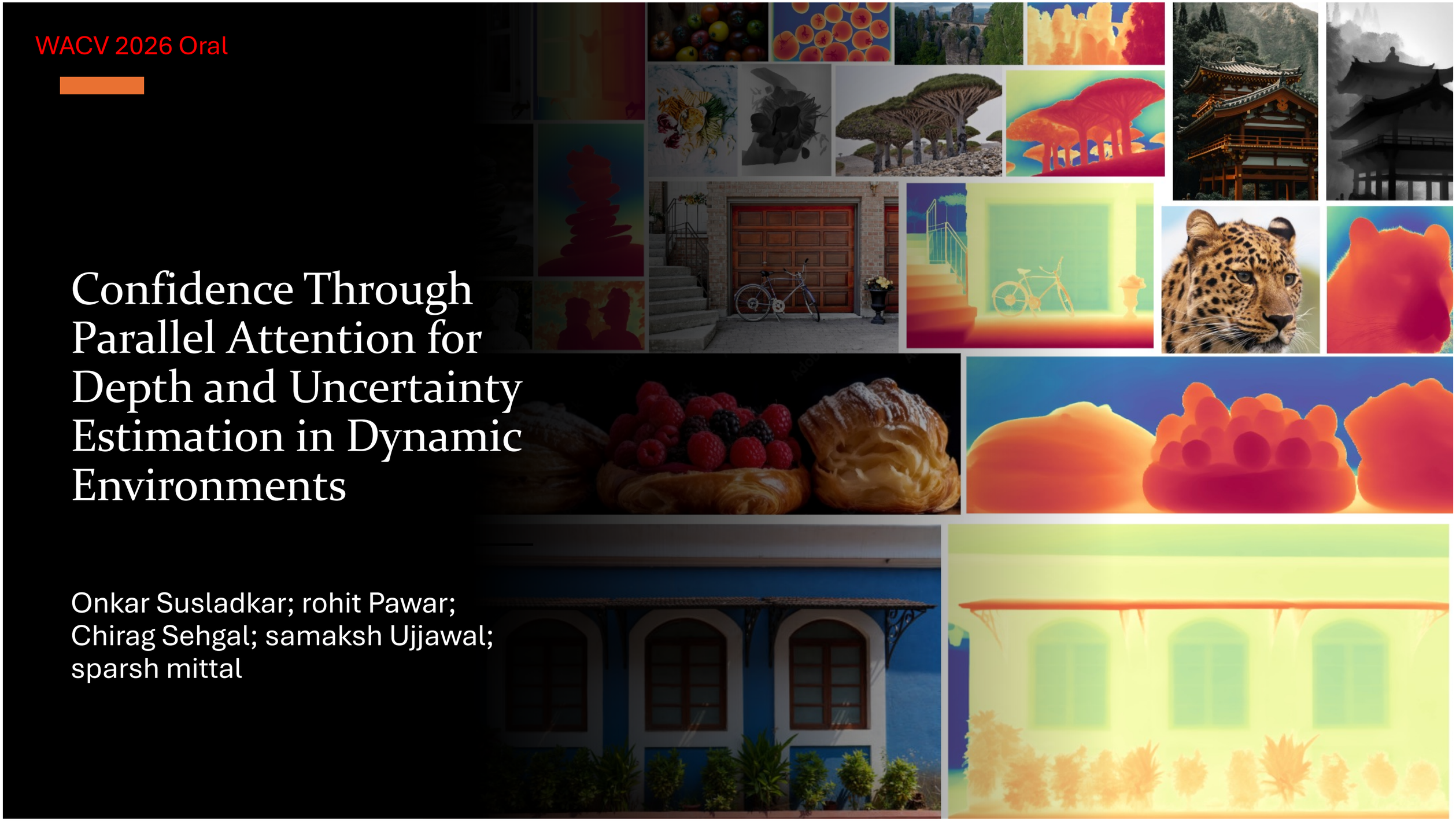
WACV 2026 Oral

Confidence Through Parallel Attention for Depth and Uncertainty Estimation in Dynamic Environments

Onkar Susladkar; rohit Pawar; Chirag Sehgal; samaksh Ujjawal; sparsh mittal



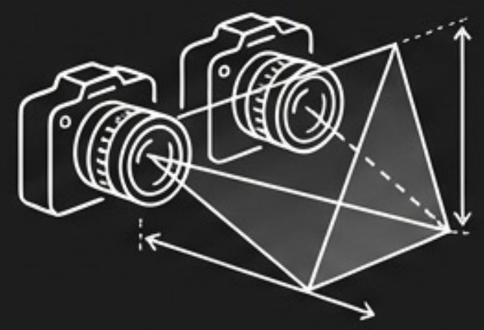

Why is need for ConFiDeNet?

The Quest for Scalable 3D Understanding

LiDAR

Precise but Costly

Accurate sparse depth, but prohibitive hardware costs limit mass deployment.



Stereo Vision

Dense but Rigid

Dual-camera setups struggle with textureless regions and require strict, fragile calibration.

Monocular

Scalable but Ambiguous

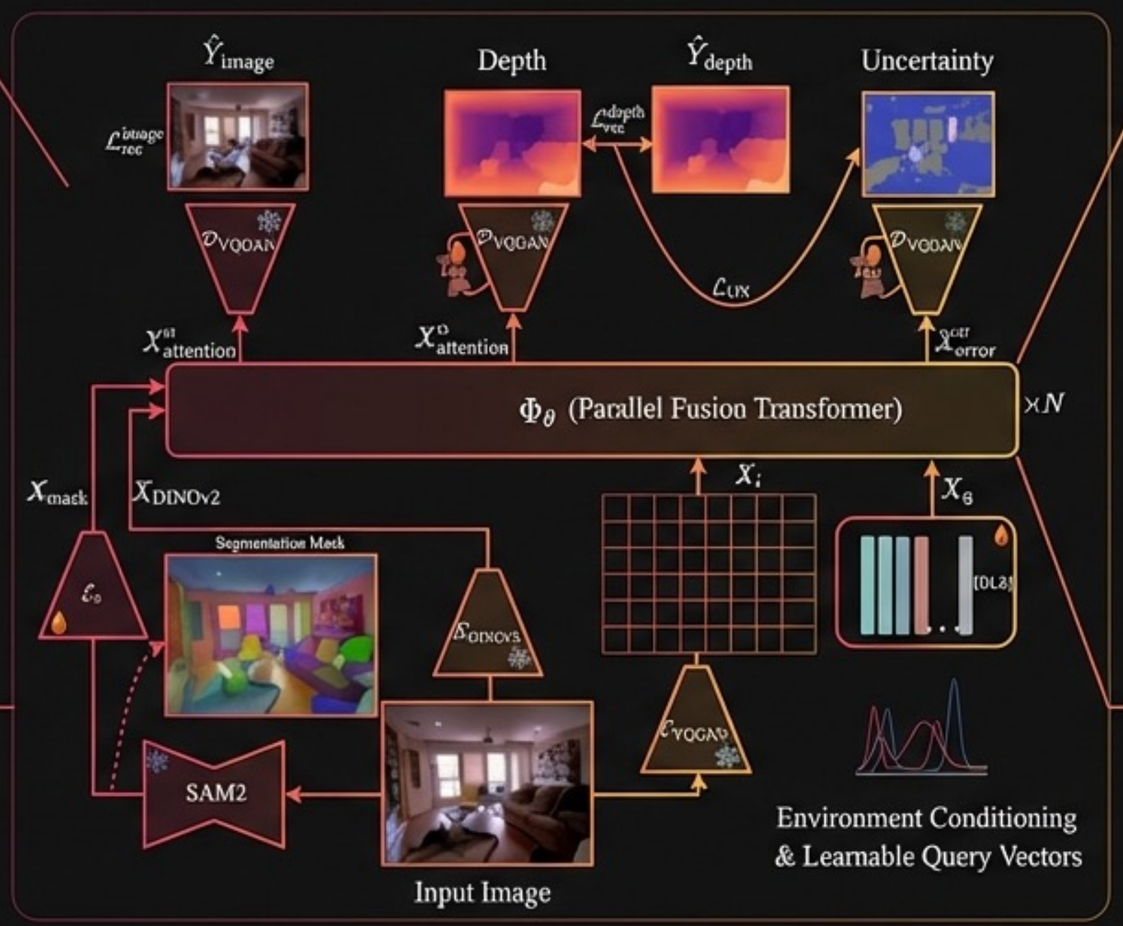
The ultimate scalable sensor. However, it struggles with 'The Real World'—occlusions, transparency, and unknown scale.

The Challenge: We need the scalability of monocular vision with the reliability of LiDAR.

A Unified Framework for Robust Perception

1. Metric Depth

Accurate distance measurement, not just relative scaling.



2. Uncertainty Estimation

Joint learning of depth and pixel-wise confidence (X_{error}).

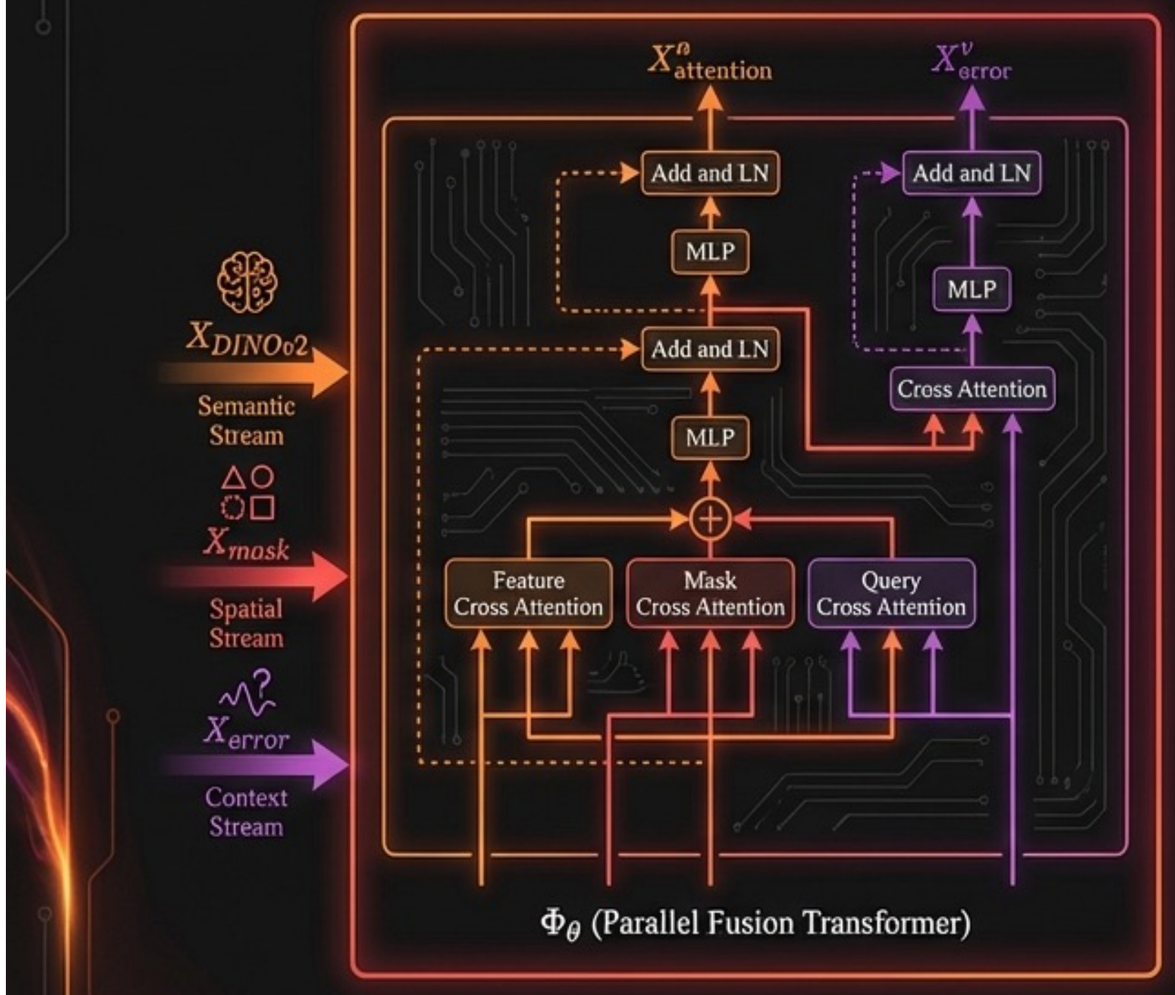
3. Zero-Shot Generalization

Deploys instantly on unseen datasets without retraining.

4. Real-Time Efficiency

672ms latency for 4K resolution images—ready for deployment.

The Power of Parallel Attention

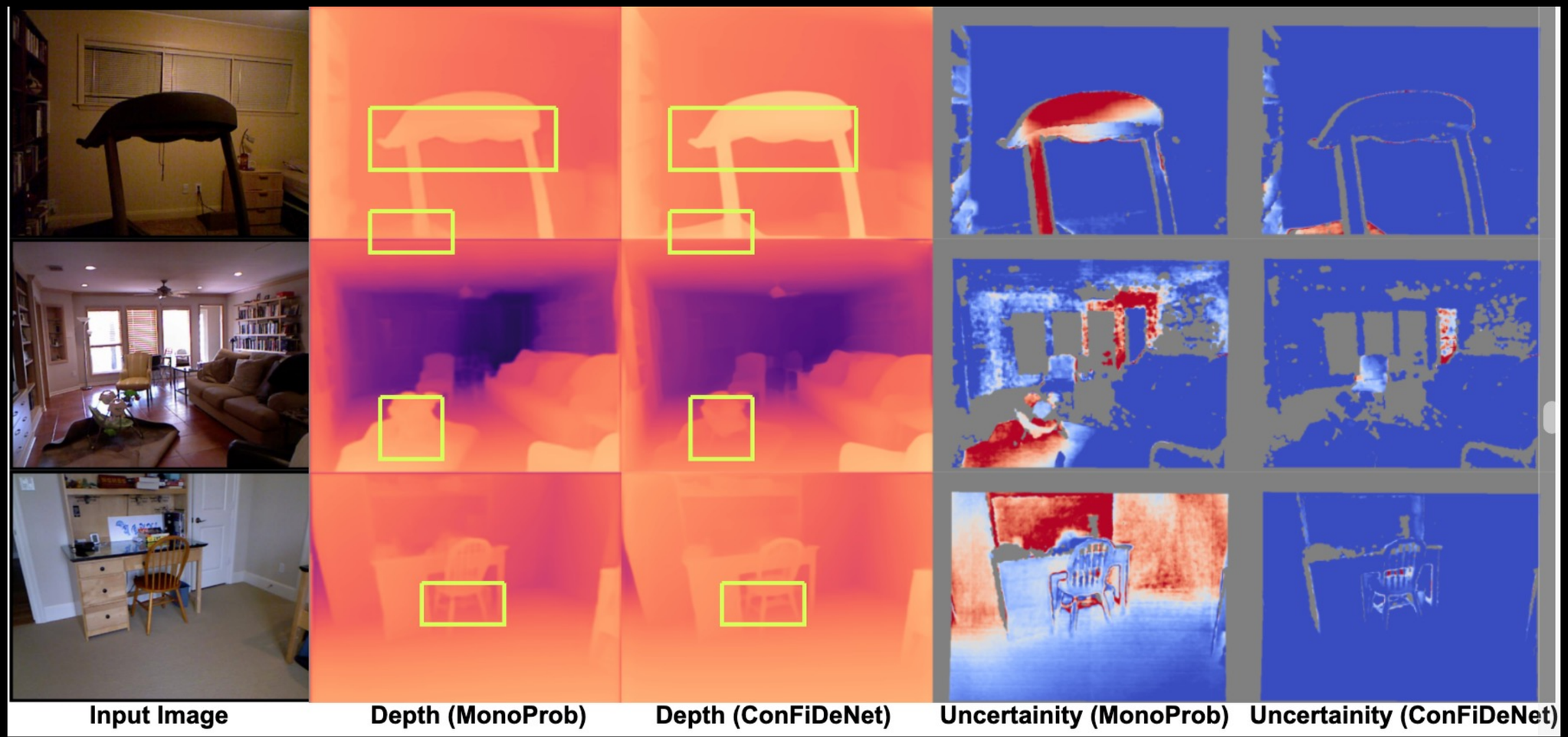


$$X_{attention} = \text{MHA}(K_{DINOv2}, Q_I, V_{DINOv2}) + \text{MHA}(K_{mask}, Q_I, V_{mask}) + \text{MHA}(K_{error}, Q_I, V_{error})$$

Unlike sequential processing, ConFiDeNet uses a lightweight module that simultaneously attends to three distinct streams:

- **Semantic Stream (X_{DINOv2}):** High-level object understanding.
- **Spatial Stream (X_{mask}):** Geometric boundaries and segmentation.
- **Context Stream (X_{error}):** Environment and uncertainty tokens.

Result: The model reasons about structure, meaning, and ambiguity simultaneously.



Uncertainty in ConFiDeNet

State-of-the-Art Performance

ConFiDeNet demonstrates superior accuracy and calibrated uncertainty across challenging benchmarks.

Method	KITTI AbsRel	KITTI RMSE	NYUv2 AbsRel	Uncertainty (AUSE)
Marigold	0.089	22.8	0.085	0.032
PatchFusion	0.075	19.4	0.071	0.028
ZoeDepth	0.091	24.5	0.105	0.035
ConFiDeNet	**0.062** ⏫	**15.3** ⏫	**0.066** ⏫	**0.016** ⏫

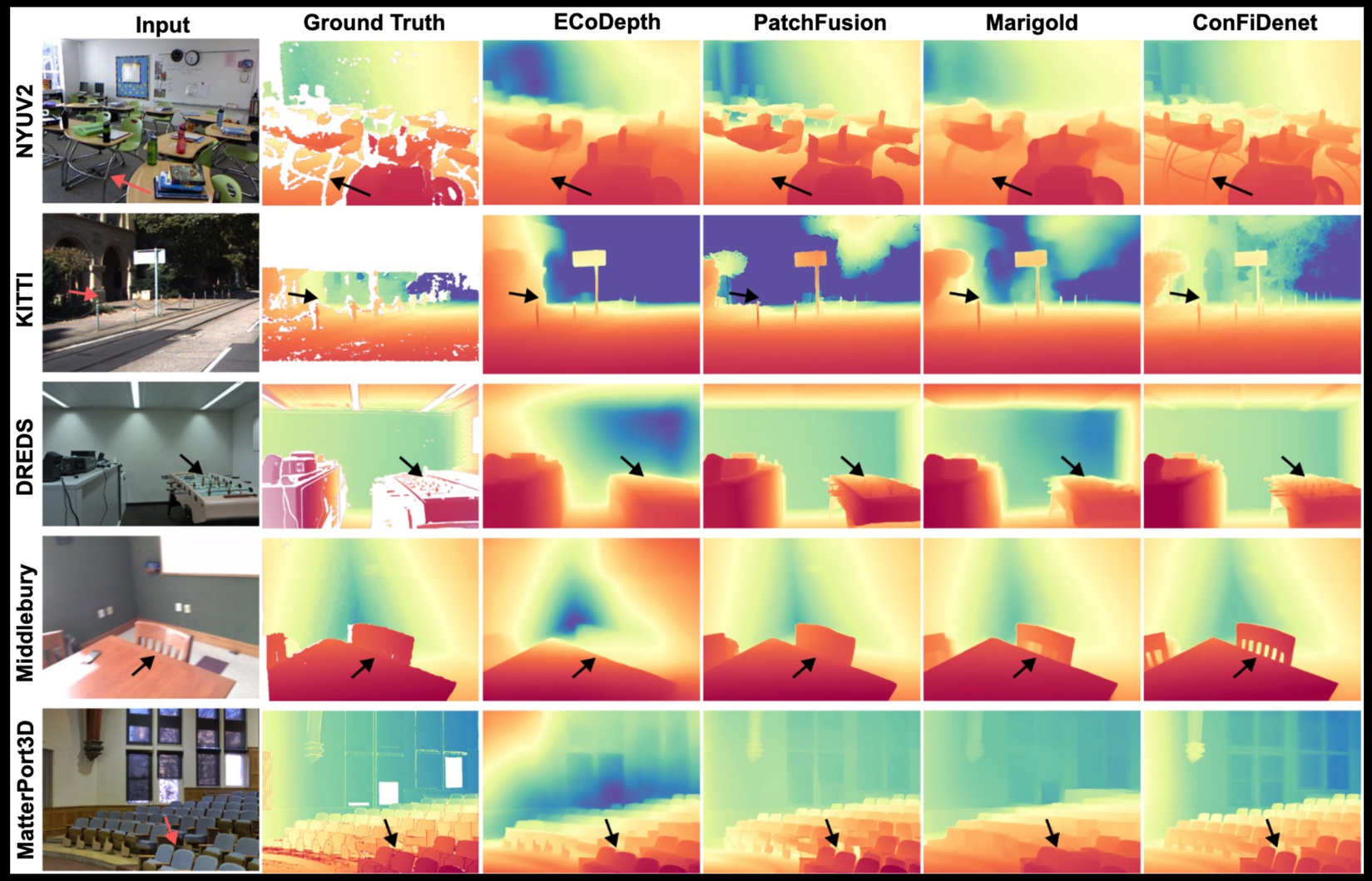
Dominance in Accuracy:

- Lowest AbsRel **(0.062)** & RMSE **(15.3)** on KITTI.
- Outperforms Marigold by **1.5x**.

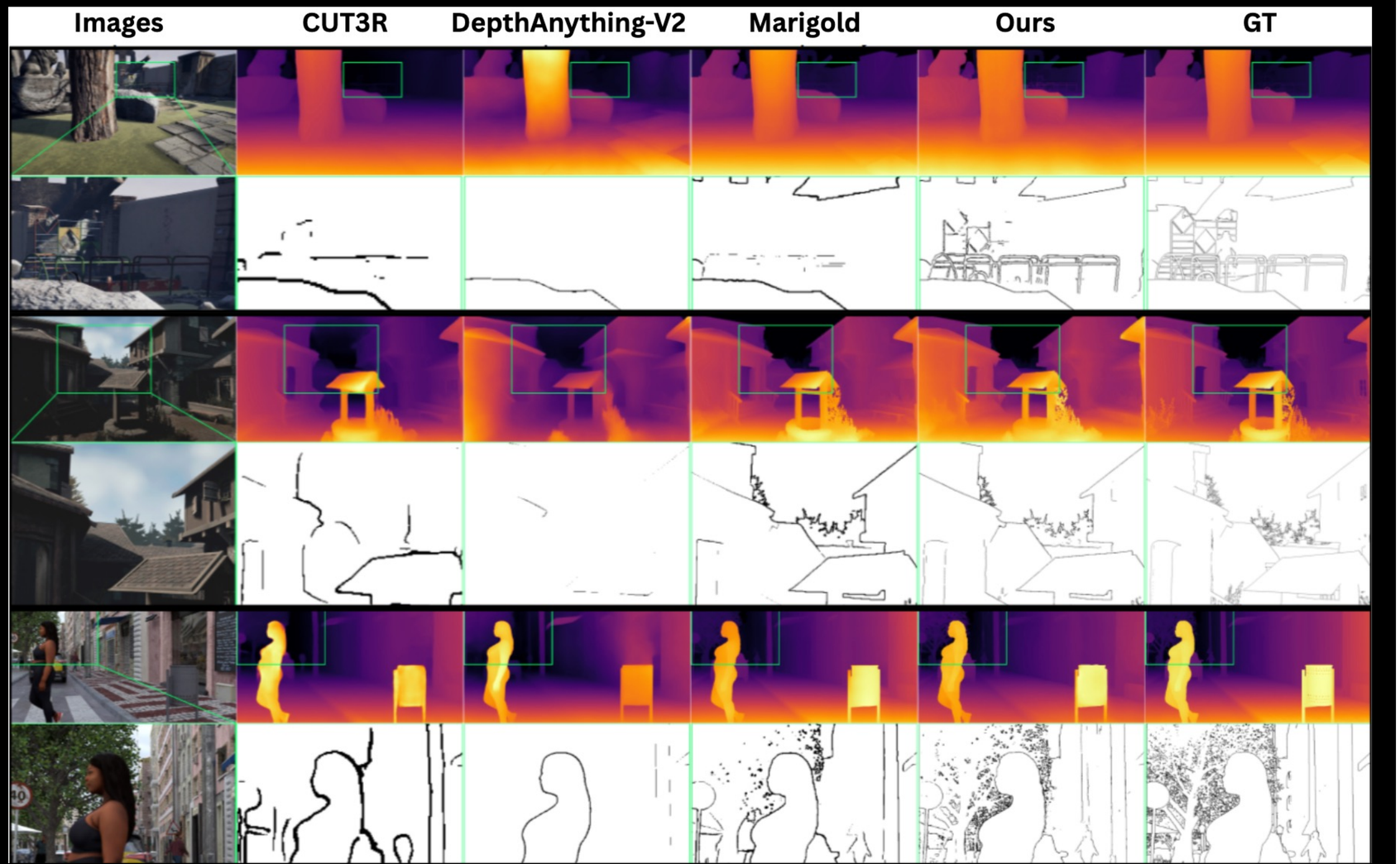
Calibrated Confidence:

- Best-in-class AUSE **(0.016)** scores prove the confidence maps match reality.

Visual Comparison



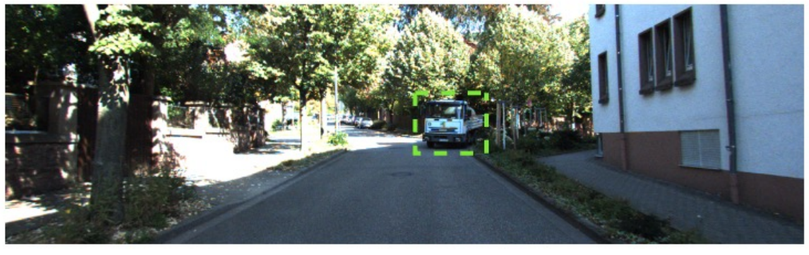
Edge Fidelity



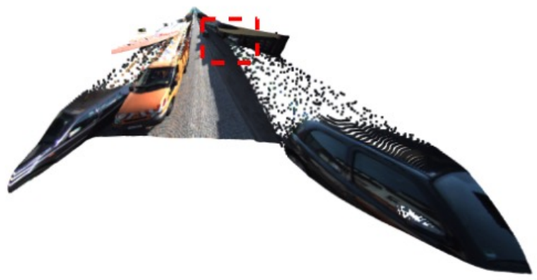
Normal Estimation with Pretrained ConFiDeNet

Point cloud from 16Bit Depth

Monocular RGB Image



Point cloud from Depthmap (Marigold)



Point Cloud from Depthmap (ConFiDeNet)

Future Work

•**Extension to Video Modality**
Enable temporally consistent predictions for dynamic scenes through video-based modeling.

•**Unified Multi-Property Geometric Model**
Extend the framework to jointly predict additional intrinsic properties such as **metallic surfaces, reflectivity, and albedo maps**.

•**Aerial Depth & Uncertainty Estimation for Drones**
Generalize the model to estimate **aerial depth and aerial uncertainty**, enabling safer and more reliable drone navigation.

Thank You