M2H: Multi-Task Learning with Efficient Window-Based Cross-Task Attention for Monocular Spatial Perception

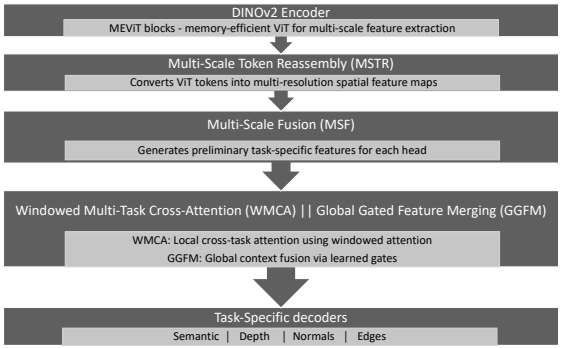
PRESENTER:
Bavantha Udugama
b.udugama@utwente.nl

INTRODUCTION

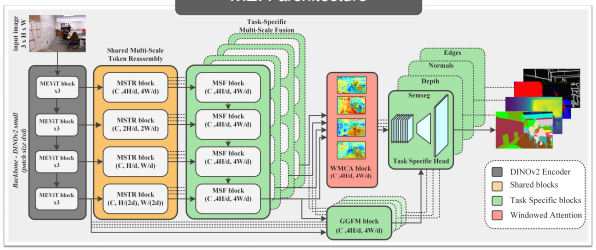
Compact and Efficient Spatial Perception for Robotics

- Autonomous systems require real-time scene understanding on edge hardware.
- M2H processes a single RGB image to predict depth, semantics, surface normals, and edges, enabling Mono-Hydra to build spatial relationships through a 5-layer hierarchical scene graph.

METHOD

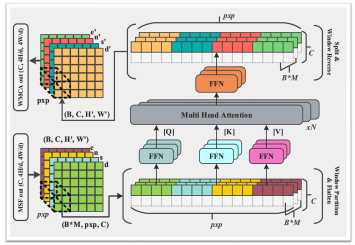


M2H architecture



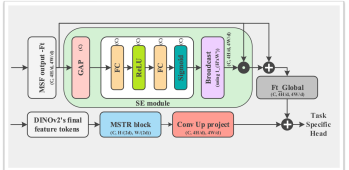
WMCA Block

- Shares features across tasks using localised attention windows
- Enables task interaction without high global attention cost
- Inspired by Swin Transformers for efficiency and locality

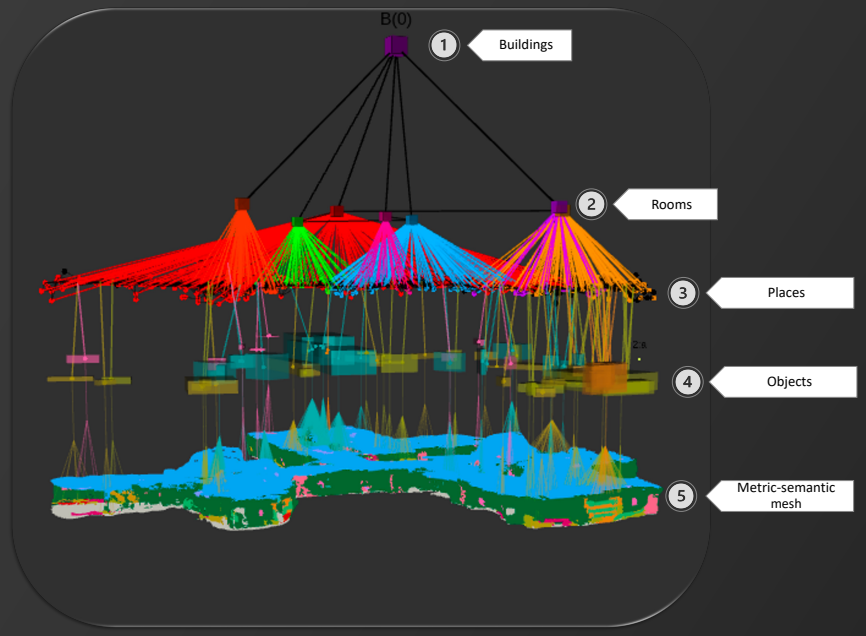


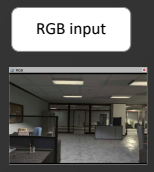
GGFM Block

- Merges global features across all tasks using gated attention
- Utilises learned gates to adaptively fuse complementary information



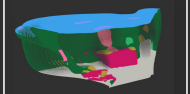
Advanced Mental Model for Robots; Human-like **spatial understanding** is possible with a single **camera in real-time.**

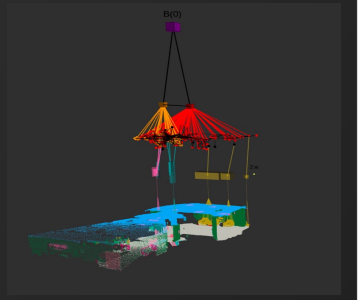



RGB input


Predicted Semantics


Predicted Depth


Semantic Point Cloud



3D Scene Graph

Scan QR for more details

RESULTS

Multi-Task Benchmarks:

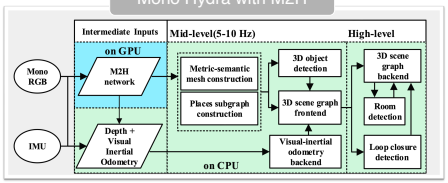
- M2H consistently outperforms task-specific and multi-task baselines across NYUDv2, Hypersim and Cityscapes.
- Cityscapes results confirm that M2H generalizes well to outdoor domains.

Dataset	Metric		SOTA (Method)	M2H (Ours)	Improvement
NYUDv2	Semseg	(mIoU ↑)	58.14 (SwinMTL)	61.54	+3.4
	Depth	(RMSE ↓)	0.4818 (MTMamba++)	0.4196	-13%
Hypersim	Semseg	(mIoU ↑)	46.66 (EMSANet)	52.31	+5.65
	Depth	(RMSE ↓)	4.825 (ScaleDepth-NK)	3.19	-33.9%
Cityscapes	Semseg	(mIoU ↑)	76.41 (SwinMTL)	77.60	+1.19
	Depth	(RMSE ↓)	6.32 (SwinMTL)	6.10	-3.5%

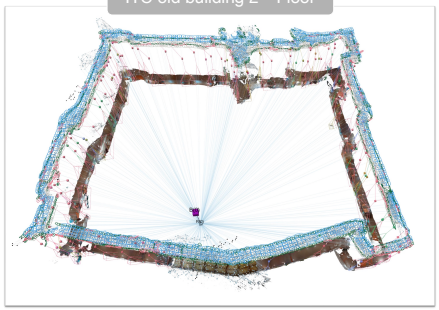
Mono Hydra Real-World Test (ITC Dataset):

- 34.2% lower reconstruction error vs MTMamba++
- 21.8% improvement over DistDepth + HRNetv2 (single-task combo)
- Runs at 30 FPS on RTX 3080 in a real-world monocular + IMU setup

Mono Hydra with M2H



ITC old building 2nd Floor



Ablation:

- Removing WMCA leads to a 6.1% drop in mIoU and a 19.2% increase in RMSE.
- Removing GGFM results in a 4.6% drop in mIoU.
- WMCA outperforms generic Multi Head Attention with ~16% lower FLOPs and better accuracy.

CONCLUSION

- M2H outperforms SOTA in multi-task dense perception from a single RGB image.
- Lightweight and real-time, suitable for monocular 3D scene understanding in real-world systems.

DISCUSSION

- How to best leverage M2H for temporal multi-view consistency?

Prof. Francesco Nex, Prof. George Vosselman

UNIVERSITY OF TWENTE. | ITC