HiMODE:
A Hybrid Monocular Omnidirectional Depth Estimation Model

Masum Shah Junayed, Arezoo Sadeghzadeh, Md Baharul Islam, Lai-Kuan Wong, Tarkan Aydin

20th June, 2022

Funded by: TÜBİTAK

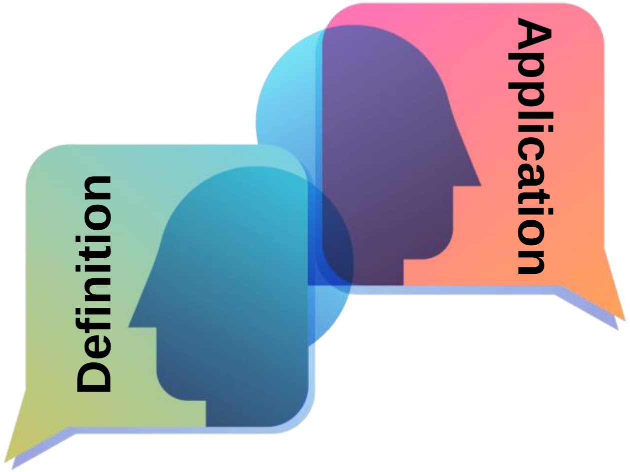
BAU BAHÇEŞEHİR ÜNİVERSİTESİ

AUM AMERICAN UNIVERSITY OF MALTA

MMU MULTIMEDIA UNIVERSITY

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

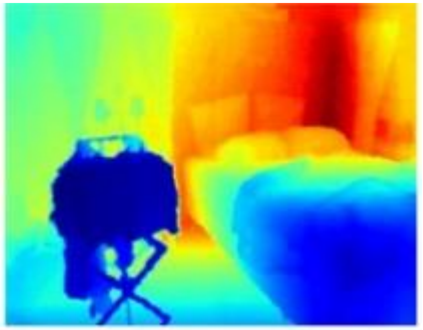
Introduction



Depth Estimation:

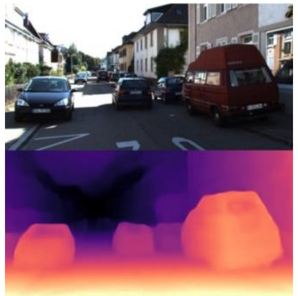
3D scene understanding from a single 2D image

Single RGB Image



Depth Map

Autonomous Driving



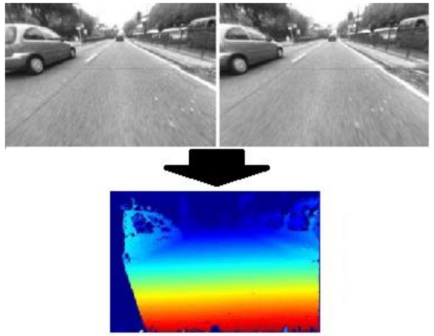
Virtual/Augmented Reality

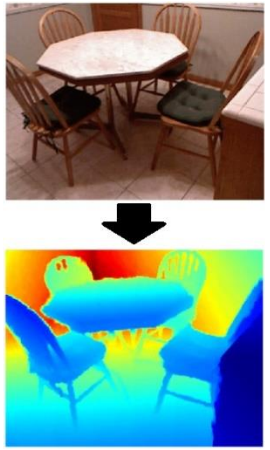
Robotics

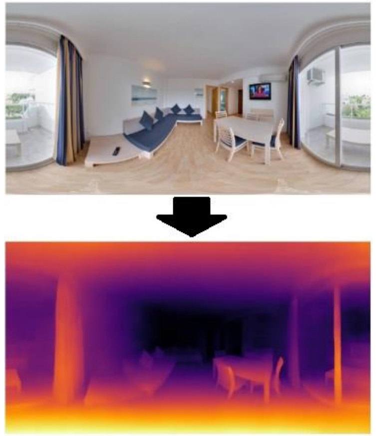
3D Reconstruction

Object Detection

Background







Depth Sensors

- Accurate depth measurement
- Inefficient in sunlight, nearby absorbing materials, and reflective surfaces
- Laborious and time-consuming
- Only available to a few high-end products

Stereo Images

- Lighter, robust, and compact
- Emitting no signal
- Challenging camera setting and alignment
- Unavailability of stereo datasets

Monocular Images

- Available to many phones
- Availability of large-scale datasets
- Limited field of view

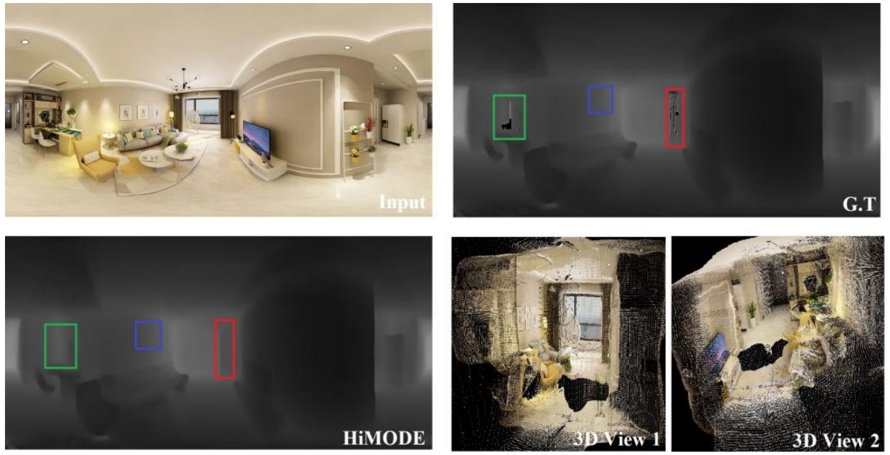
360º Depth

Omnidirectional monocular depth estimation

Providing full perception of the surroundings for a safe navigation

Motivation

HiMODE: CNN+Transformer



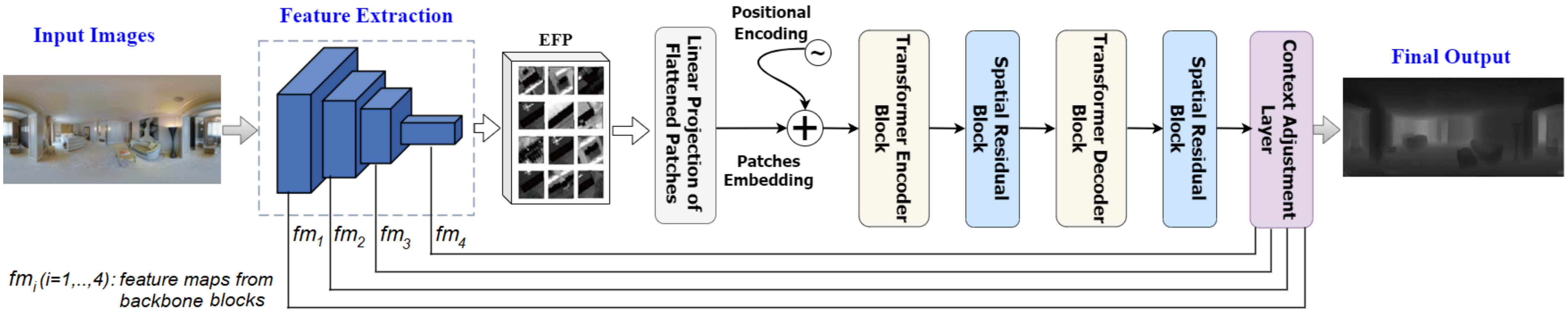
Solution

CNN-based Methods: successful estimation around the equator but significant distortions in the poles due to limited receptive field

Transformer-based Methods: inferior performance with small-scale datasets, still cannot deal with the data loss of the ground-truth, recovering the small objects details is challenging

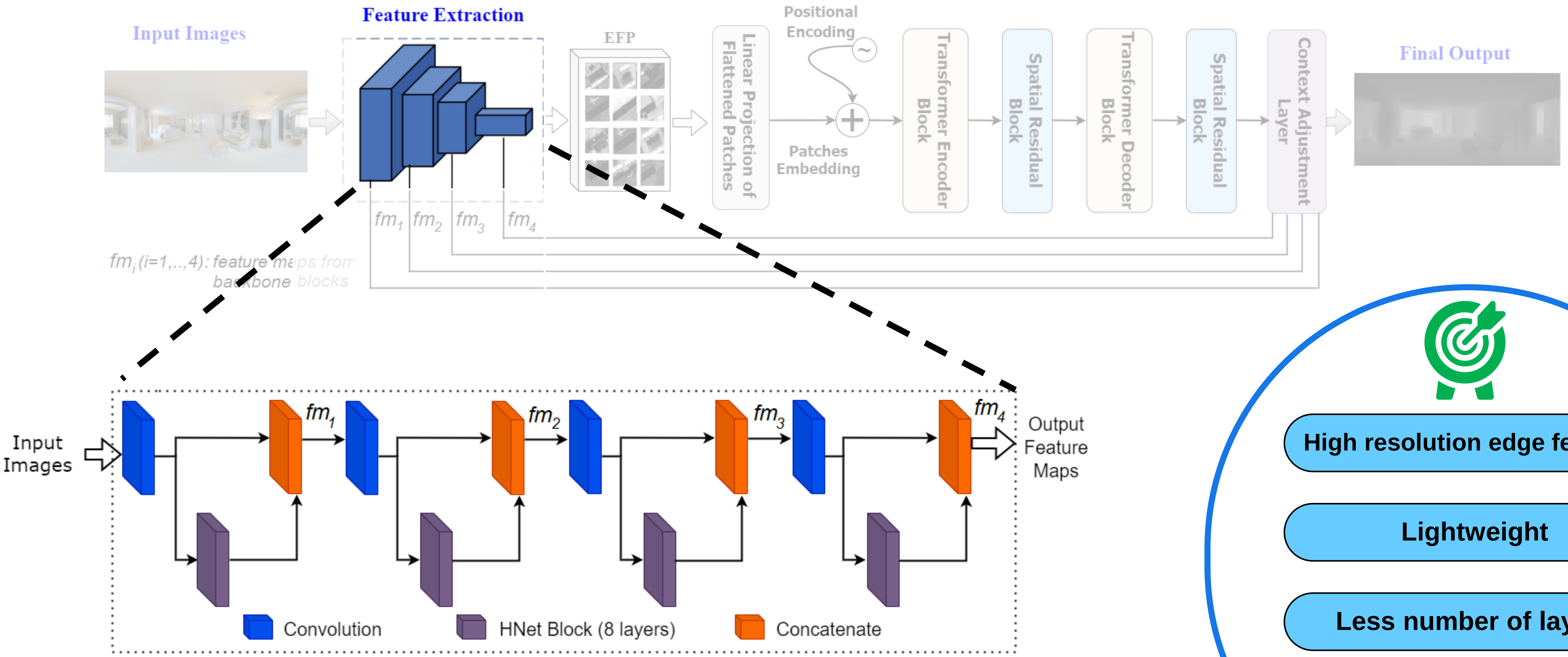
Challenges

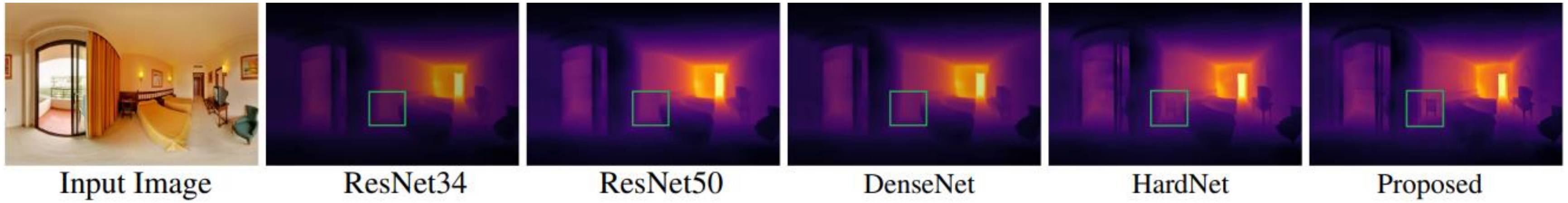
The Proposed HiMODE



HiMODE architecture overview.

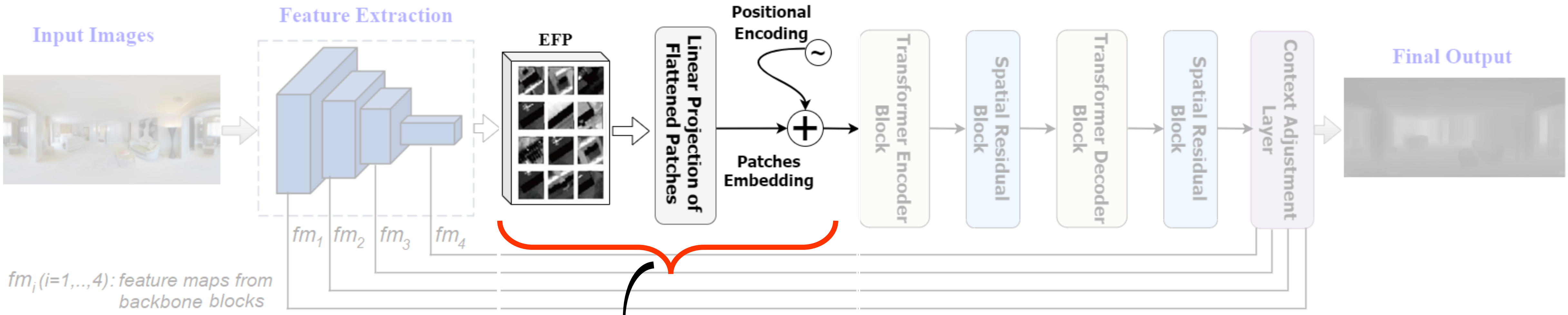
The Proposed HiMODE





- High resolution edge features
- Lightweight
- Less number of layers
- Superior Performance

The Proposed HiMODE



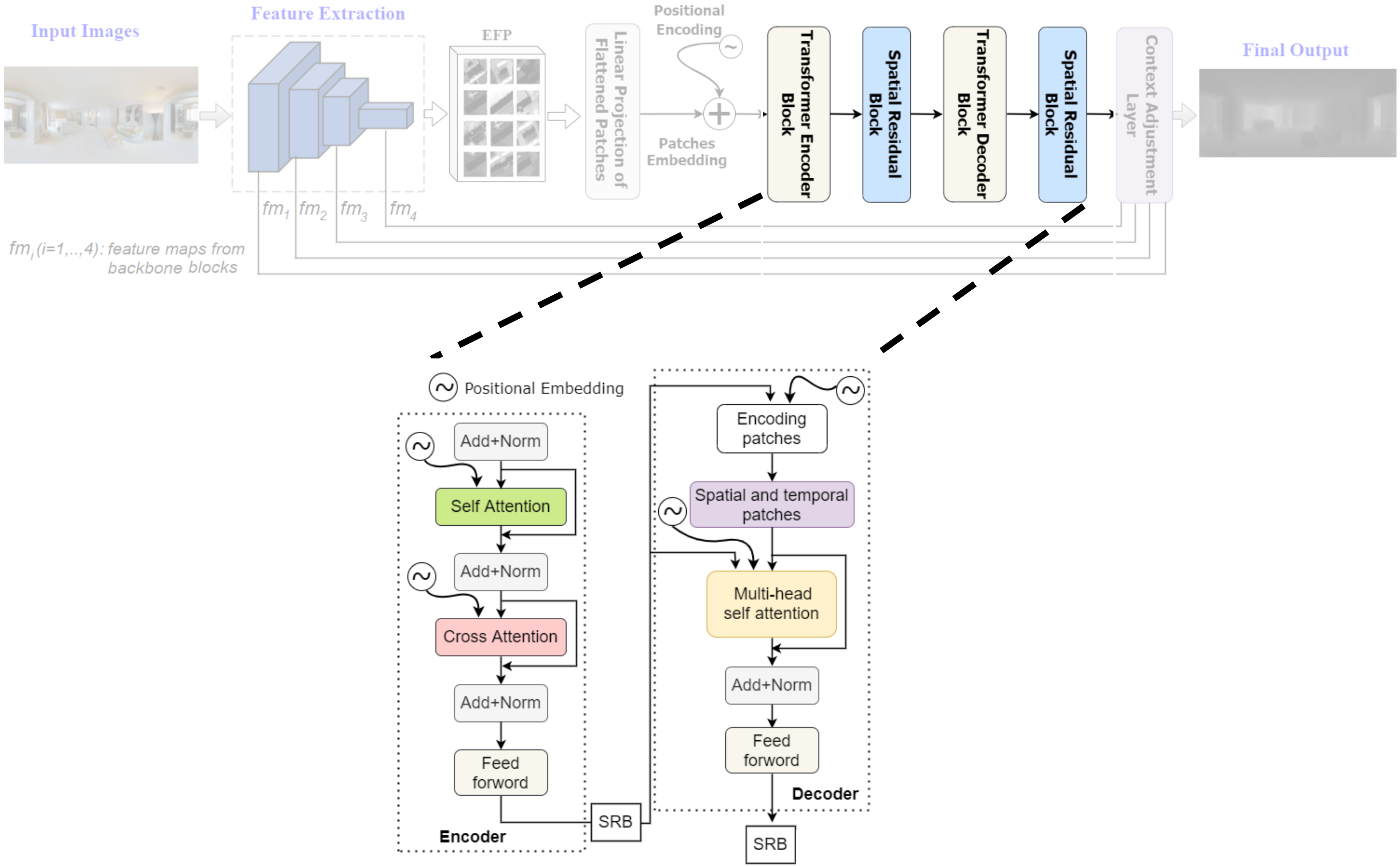
$$I = Concat(PE, PE')$$

Patch Embeddings

Positional Embeddings

$$PE'_{(pos,2i)} = sin(pos/10000^{2i/D})$$

The Proposed HiMODE



The Proposed HiMODE

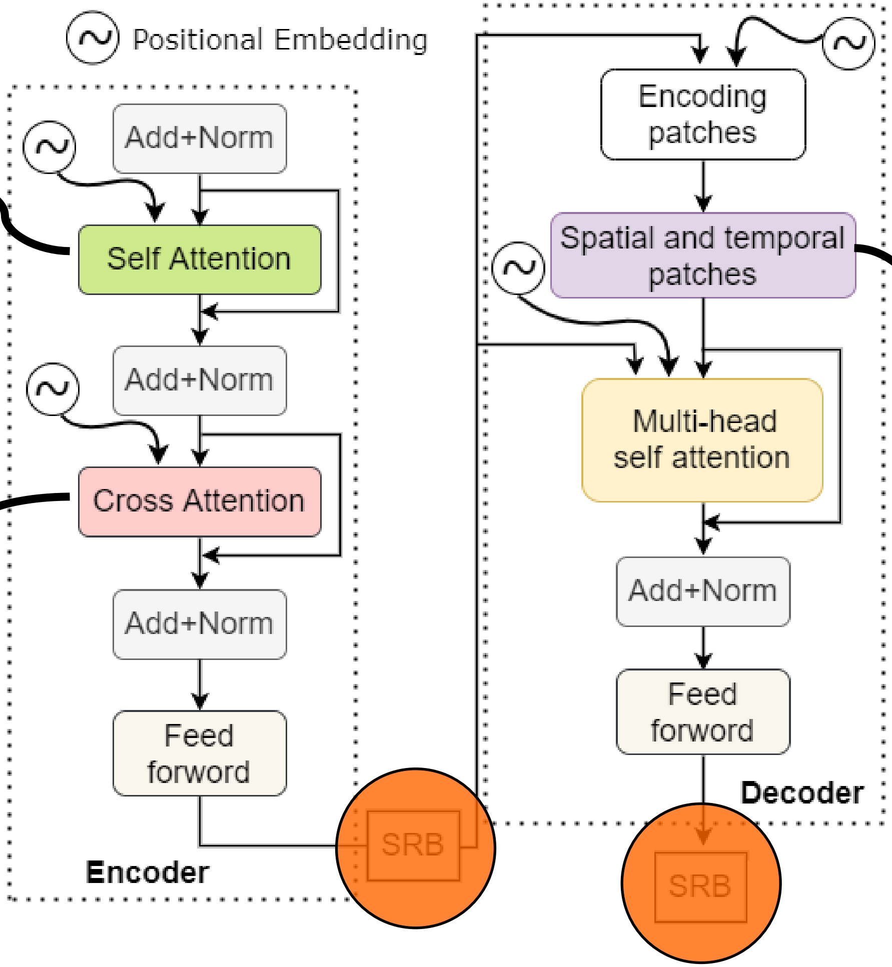
$$[Q, K, V] = I \times U_{QKV}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{D_k}})V = AV$$

Irrelevant and noisy data are filtered out

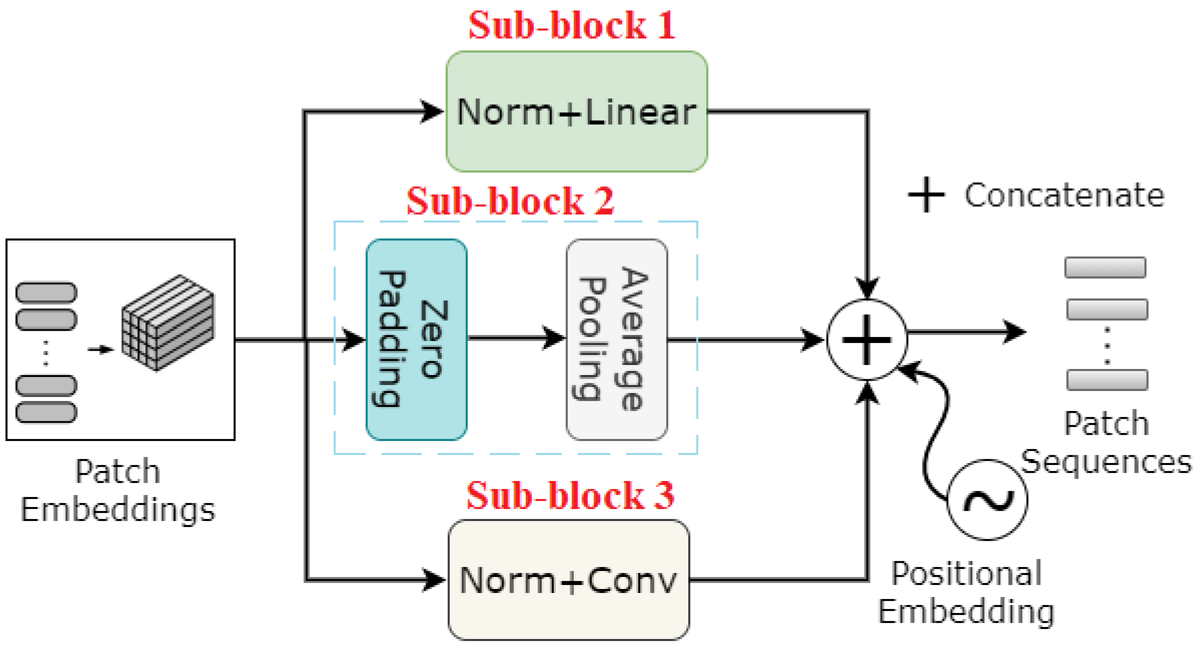
Computing attention between the pixels of different patches

Efficient process



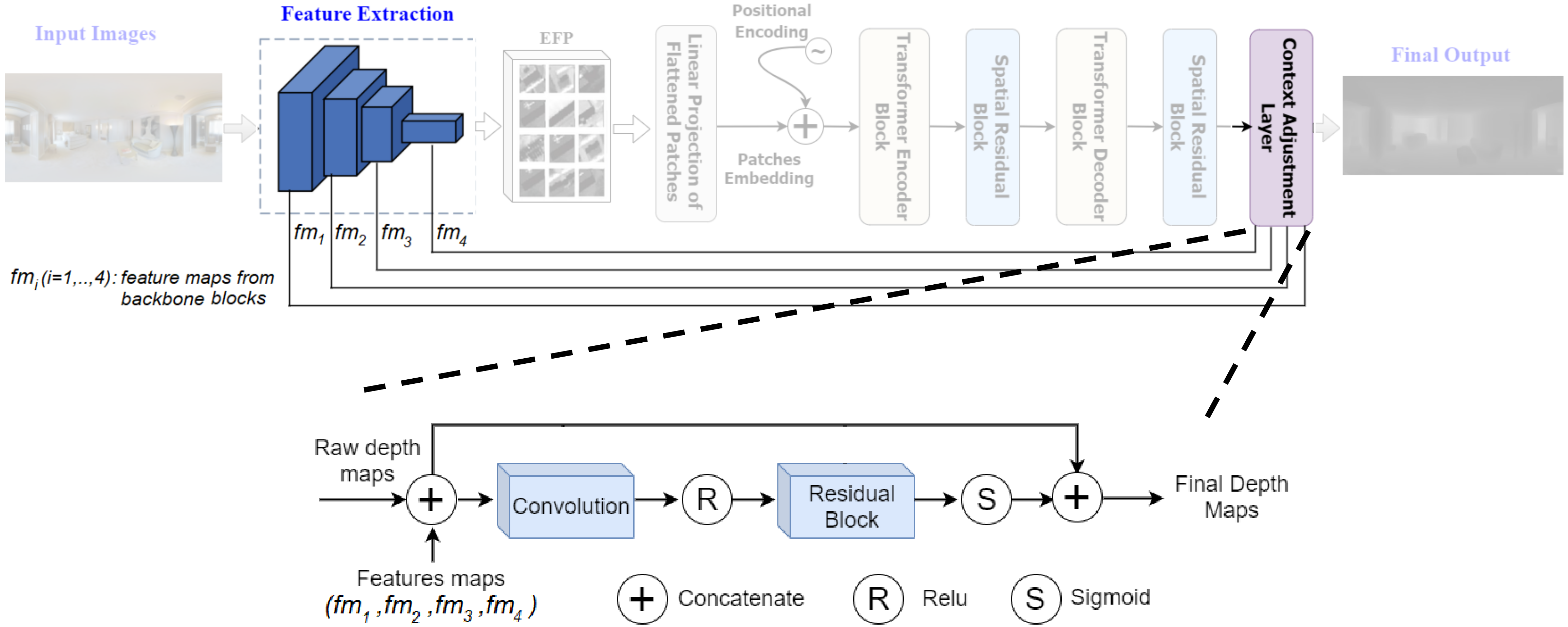
1) A temporal mechanism for finding the similarities of the patches from a smaller spatial area along with the temporal dimensions
2) A spatial mechanism for searching similarities of the patches

The Proposed HiMODE



- Improving the efficiency
- Decreasing computation cost
- Stabilizing training

The Proposed HiMODE



Experimental Setup and Datasets

1413 Images

Stanford3D Dataset

Training Details

- PyTorch
- Intel Core i9-10850K CPU with a 3.60GHz processor, 64GB RAM, and NVIDIA GeForce RTX 2070 GPU.
- Two T-blocks, 128 hidden nodes, one self-attention, one cross-attention, and one MHSA
- Adam optimizer with a batch size of 4 and 55 epochs
- Learning rates of 0.00001 and 0.0003 for the real-world and synthetic data.

Matterport3D Dataset

10800 Images

PanoSUNCG Dataset

25000 Images

Quantitative Results

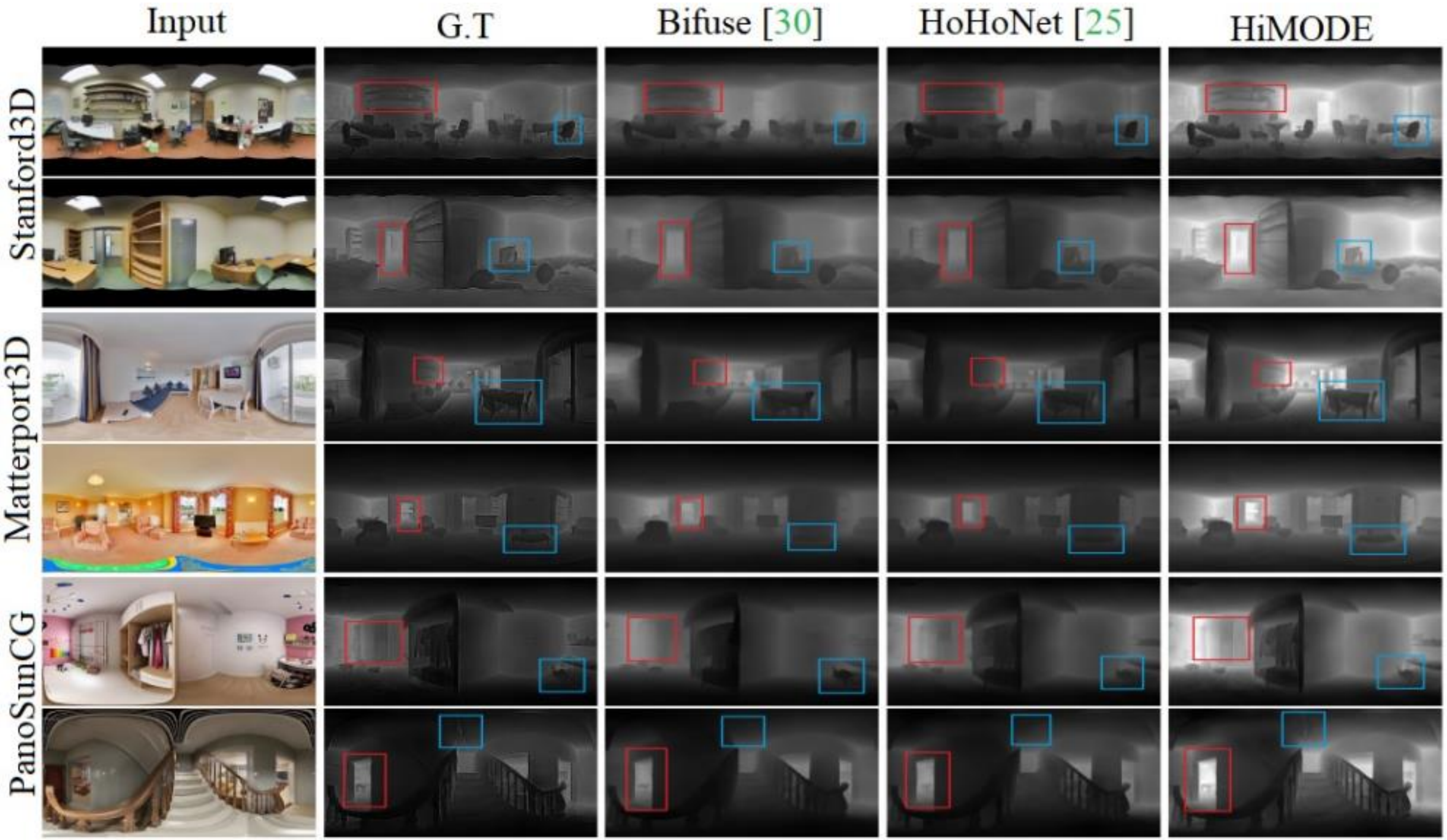
Datasets	Approaches	Abs-Rel	Sq-Rel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Stanford3D	Omnidepth [39]	0.1009	0.0522	0.3835	0.1434	0.9114	0.9855	0.9958
	SvSyn [38]	0.1003	0.0492	0.3614	0.1478	0.9296	0.9822	0.9949
	Bifuse [30]	0.1214	0.1019	0.5396	0.1862	0.8568	0.9599	0.9880
	HoHoNet [25]	0.0901	0.0593	0.4132	0.1511	0.9047	0.9762	0.9933
	NLDPT [36]	0.0649	0.0240	0.2776	0.993	0.9665	0.9948	0.9983
	HiMODE	**0.0532**	**0.0207**	**0.2619**	**0.0821**	**0.9711**	**0.9965**	**0.9989**
Matterport3D	Omnidepth [39]	0.1136	0.0691	0.4438	0.1591	0.8795	0.9795	0.9950
	SvSyn [38]	0.1063	0.0599	0.4062	0.1569	0.8984	0.9773	0.9974
	Bifuse [30]	0.139	0.1359	0.6277	0.2079	0.8381	0.9444	0.9815
	HoHoNet [25]	0.0671	0.0417	0.3416	0.1270	0.9415	0.9838	0.9942
	NLDPT [36]	0.0700	0.0287	**0.3032**	0.1051	0.9599	0.9938	0.9982
	HiMODE	**0.0658**	**0.0245**	0.3067	**0.0959**	**0.9608**	**0.9940**	**0.9985**
PanoSunCG	Omnidepth [39]	0.1450	0.1052	0.5684	0.1884	0.8105	0.9761	0.9941
	SvSyn [38]	0.1867	0.1715	0.6965	0.2380	0.7222	0.9427	0.9840
	Bifuse [30]	0.2203	0.2693	0.8869	0.2864	0.6719	0.8846	0.9660
	HoHoNet [25]	0.0827	0.0633	0.3863	0.1508	0.9266	0.9765	0.9908
	NLDPT [36]	0.0715	0.0361	0.3421	**0.1042**	0.9625	0.9950	0.9989
	HiMODE	**0.0682**	**0.0356**	**0.3378**	0.1048	**0.9688**	**0.9951**	**0.9992**

Quantitative performance comparison of the proposed HiMODE with the state-of-the-art methods

Approaches	Threshold	Recall	Precision	F1-Score
Laina et al. [16]	0.25	0.435	0.489	0.454
	0.50	0.422	0.536	0.463
	1.00	0.479	0.670	0.548
Xu et al. [16]	0.25	0.400	0.516	0.436
	0.50	0.363	0.600	0.439
	1.00	0.407	0.794	0.525
Fu et al. [33]	0.25	0.583	0.320	0.402
	0.50	0.473	0.316	0.412
	1.00	0.512	0.483	0.485
Hu et al. [10]	0.25	0.508	0.644	0.562
	0.50	0.505	0.668	0.568
	1.00	0.540	0.759	0.623
Yang et al. [34]	0.25	0.518	0.652	0.570
	0.50	0.510	0.685	0.576
	1.00	0.544	0.774	0.631
HiMODE	0.25	**0.598**	**0.703**	**0.634**
	0.50	**0.569**	**0.720**	**0.605**
	1.00	**0.641**	**0.815**	**0.656**

Performance comparison on edge pixels recovery for MDE on NYU Depth V2 dataset (non-panoramic images)

Qualitative Results



Ablation Study

		Errors				Accuracy		
Datasets	Backbones	Abs-Rel	Sq-Rel	RMSE	RMSElog	δ	δ^2	δ^3
Stanford3D	ResNet34 [12]	0.1128	0.0635	0.3665	0.1873	0.9149	0.9884	0.9880
	ResNet50 [12]	**0.0509**	0.0682	0.3177	0.1185	0.9349	0.9906	0.9923
	DenseNet [14]	0.1045	0.0624	0.3358	0.1621	0.9076	0.9839	0.9889
	HardNet [5]	0.0789	0.0352	0.3041	0.1215	0.9234	0.9947	**0.9992**
	Proposed	0.0532	**0.0207**	**0.2619**	**0.0821**	**0.9711**	**0.9965**	0.9989
Matterport3D	ResNet34 [12]	0.1078	0.1139	0.4587	0.1786	0.8946	0.9792	0.9800
	ResNet50 [12]	0.1014	0.0856	0.4189	0.1251	0.9257	0.9755	0.9945
	DenseNet [14]	0.0935	0.0472	0.3548	0.1547	0.9138	0.9668	0.9829
	HardNet [5]	0.0769	**0.0244**	0.3628	0.1174	0.9415	0.9831	0.9902
	Proposed	**0.0658**	0.0245	**0.3067**	**0.0959**	**0.9608**	**0.9940**	**0.9985**
PanoSunCG	ResNet34 [12]	0.1353	0.1471	0.4823	0.2379	0.9183	0.9947	0.9926
	ResNet50 [12]	0.1094	0.1043	0.3847	0.2149	0.9524	0.9918	0.9989
	DenseNet [14]	0.0949	0.0987	0.4283	0.1958	0.9245	0.9909	0.9895
	HardNet [5]	0.0726	0.0557	0.3985	0.1305	**0.9693**	0.9897	0.9877
	Proposed	**0.0682**	**0.0356**	**0.3378**	**0.1048**	0.9688	**0.9951**	**0.9992**

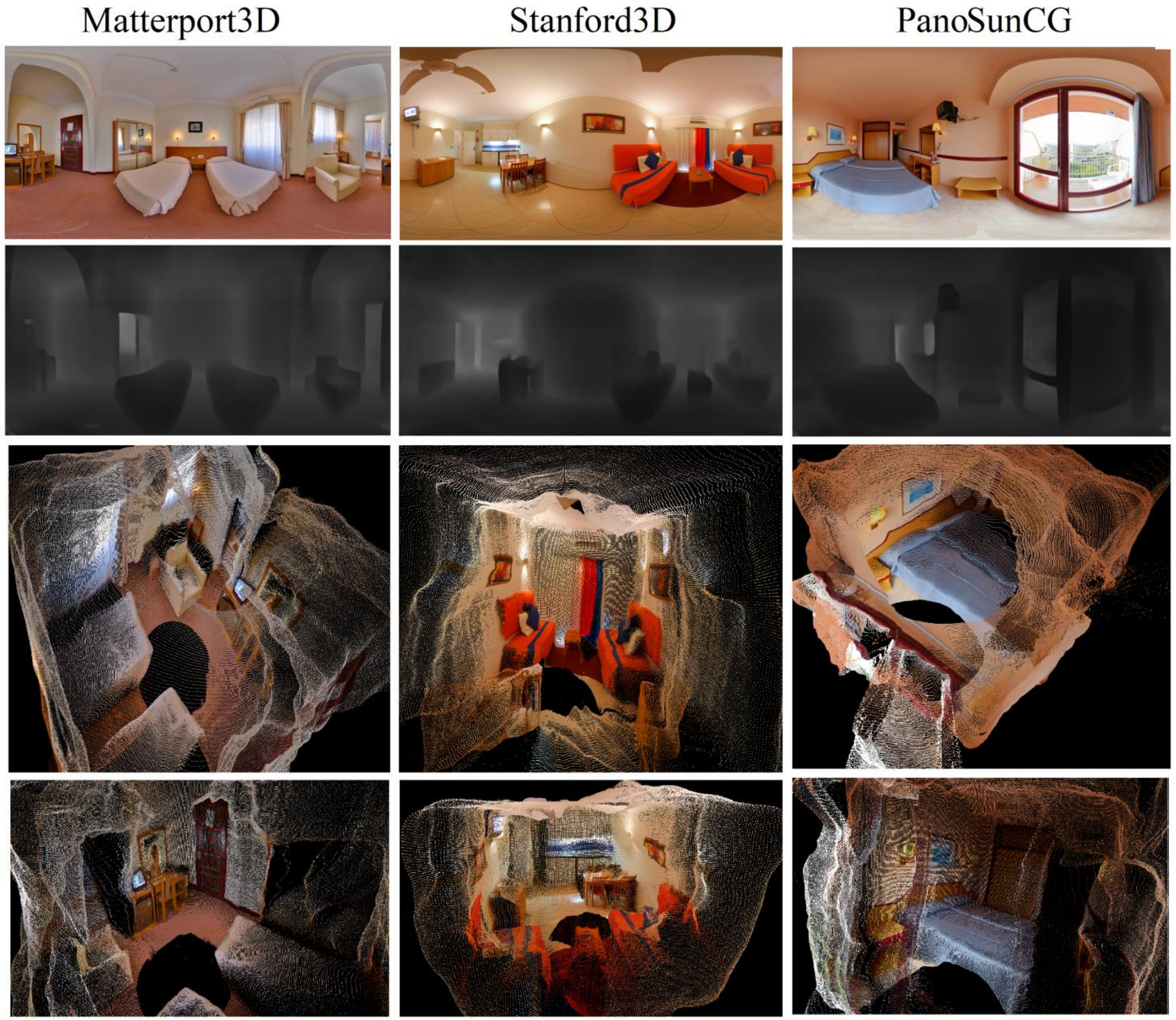
Datasets	SRB	Attention	Abs-Rel	Sq-Rel	RMSE	RMSElog	δ	δ^2	δ^3
Stanford3D	✓	SCA	**0.0532**	**0.0207**	**0.2619**	**0.0821**	**0.9711**	**0.9965**	**0.9989**
	×	SCA	0.0698	0.0395	0.2846	0.1028	0.9574	0.9898	0.9787
	✓	MHSA	0.0746	0.0590	0.3548	0.1529	0.9358	0.9748	0.9695
Matterport3D	✓	SCA	0.0658	**0.0245**	**0.3067**	**0.0959**	**0.9608**	**0.9940**	**0.9985**
	×	SCA	**0.0514**	0.0358	0.3108	0.1073	0.9480	0.9799	0.9891
	✓	MHSA	0.0629	0.0854	0.4098	0.1889	0.9466	0.9709	0.9770
PanoSunCG	✓	SCA	0.0682	**0.0356**	**0.3378**	0.1048	**0.9688**	**0.9951**	**0.9992**
	×	SCA	**0.0540**	0.0541	0.3586	**0.1038**	0.9555	0.9869	0.9902
	✓	MHSA	0.0640	0.0849	0.3928	0.1044	0.9497	0.9672	0.9816

Computation Cost

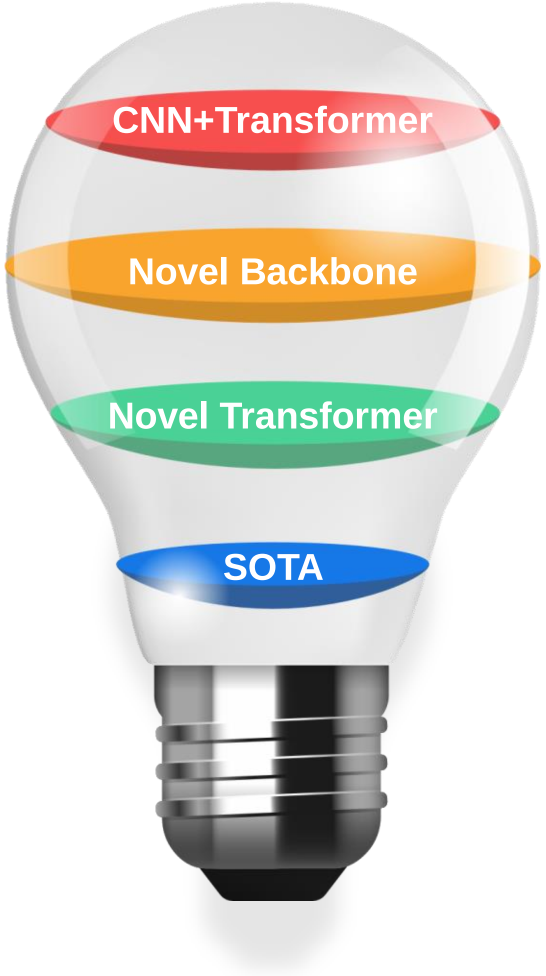
	SRB	TEB		TDB	Computation Cost	Accuracy		
		SCA	MHSA	STP	#Parm	δ	δ^2	δ^3
1	✓	✓	×	✓	**79.67M**	**0.9711**	**0.9965**	**0.9989**
2	✓	×	✓	✓	84.59M	0.9358	0.9748	0.9695
3	×	✓	×	✓	88.47M	0.9574	<u>0.9898</u>	0.9787
4	✓	✓	×	×	<u>81.37M</u>	<u>0.9623</u>	0.9746	<u>0.9877</u>
5	×	×	✓	✓	93.59M	0.9398	0.9655	0.9629
6	×	✓	×	×	95.36M	0.9238	0.9481	0.9642

Results of the ablation study on different modules in terms of computation cost and accuracy (on Stanford3D dataset). Bold and underlined numbers indicate the first and s econd best results.

3D structure Reconstruction



Conclusion



- To capitalize on the strengths of CNN-based feature extractors and the power of Transformers for monocular omnidirectional depth estimation
- The high-level features near the edges were extracted by using a pyramid-based CNN as the backbone, with the HNet block inside.
- Further improvement was achieved by applying self and cross attention along with the spatial-temporal patches and the spatial residual block.
- It not only achieved the state-of-the art performance on three datasets, but also was capable to recover the lost data in the ground-truth depth map.

Thank you for your attention!

PyTorch code and supplementary material available:
https://github.com/ himode5008/HiMODE