Ordinal Depth Classification using Region-based Self-attention

by Minh Hieu Phan, Lam Phung, Abdesselam Bouzerdoum

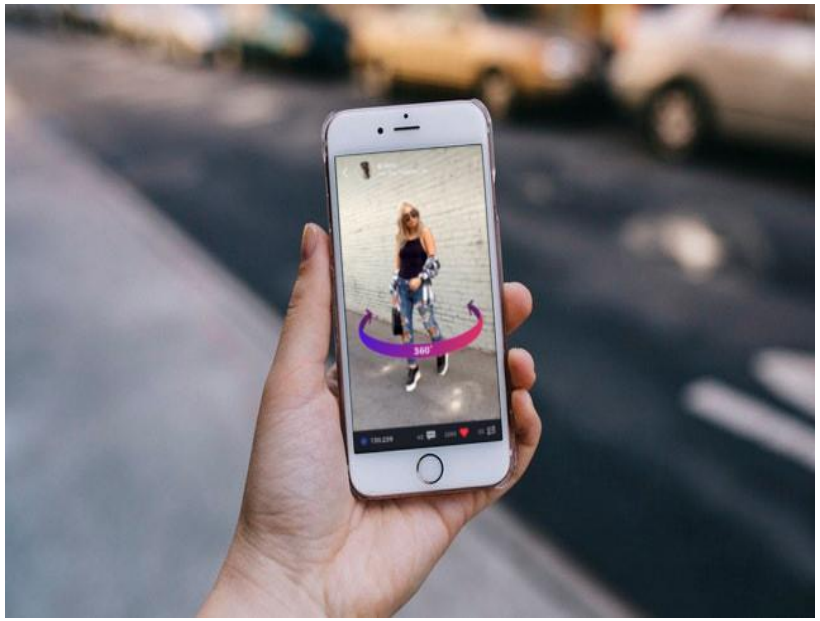
Monocular Depth Estimation

.

Monocular Depth Estimation

.

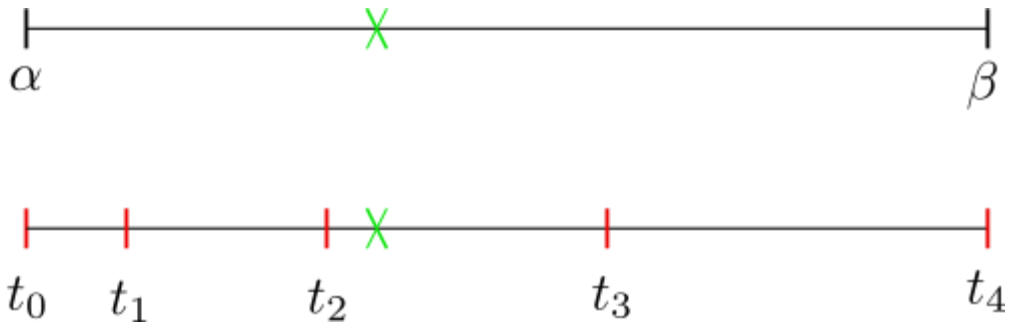
AR application



2D-3D reconstruction

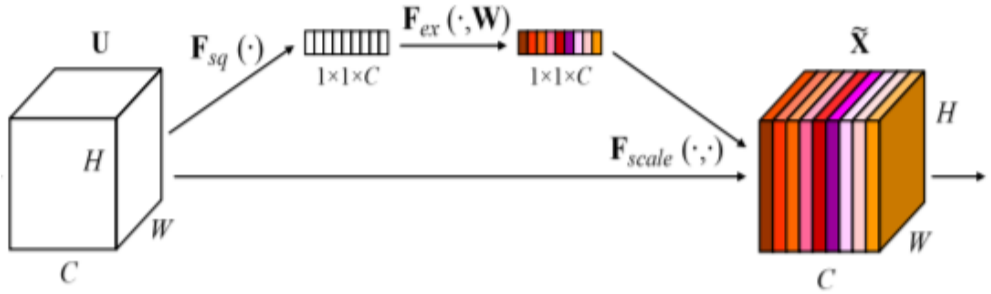
Assistive navigation

Ordinal Classification

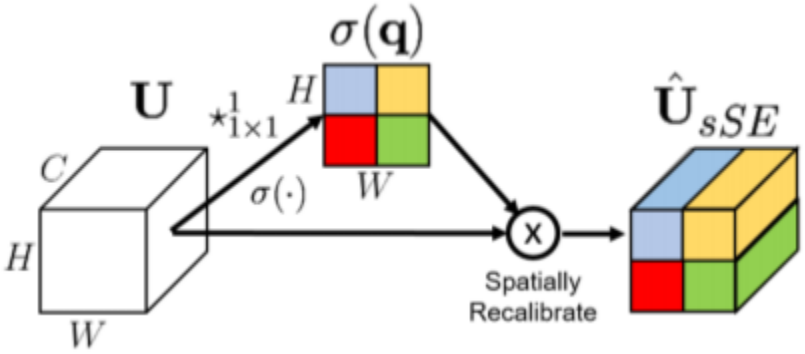


$$t_k = \exp\left(\log \alpha + \frac{\log \beta/\alpha * k}{K}\right)$$

Attention Mechanism



Channel-wise squeeze and excitation[1]



Spatial squeeze and excitation[2]

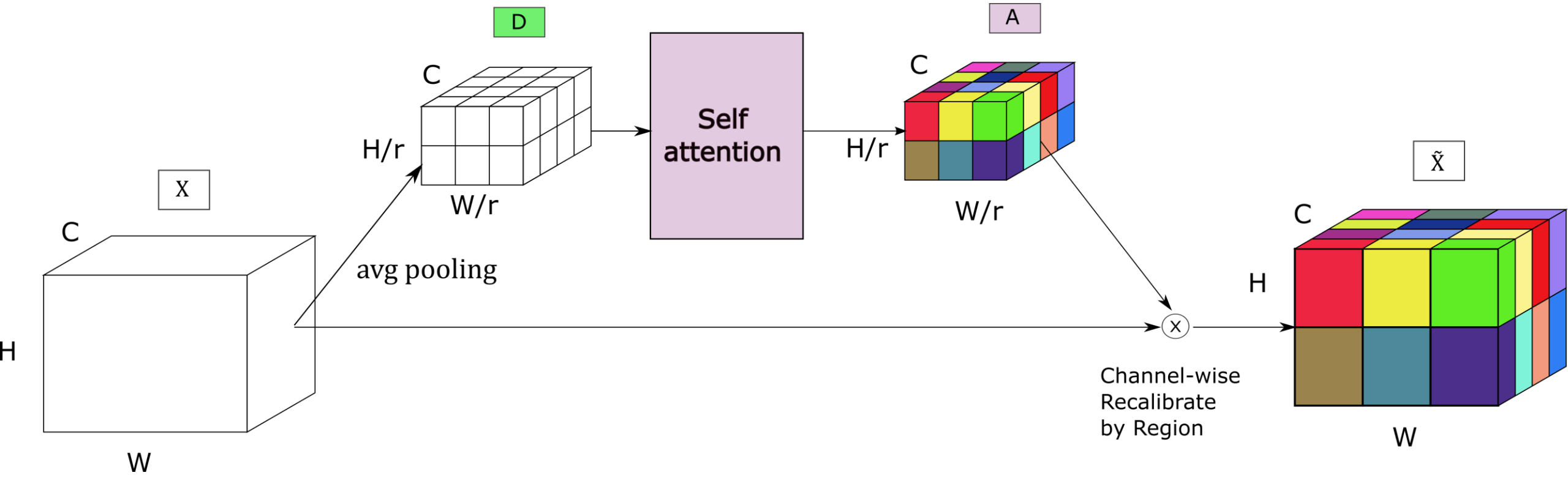
[1] Eigen and Fergus, CVPR, 2015.
[2] Hu et al., NIPS, 2018.

Monocular Depth Cues

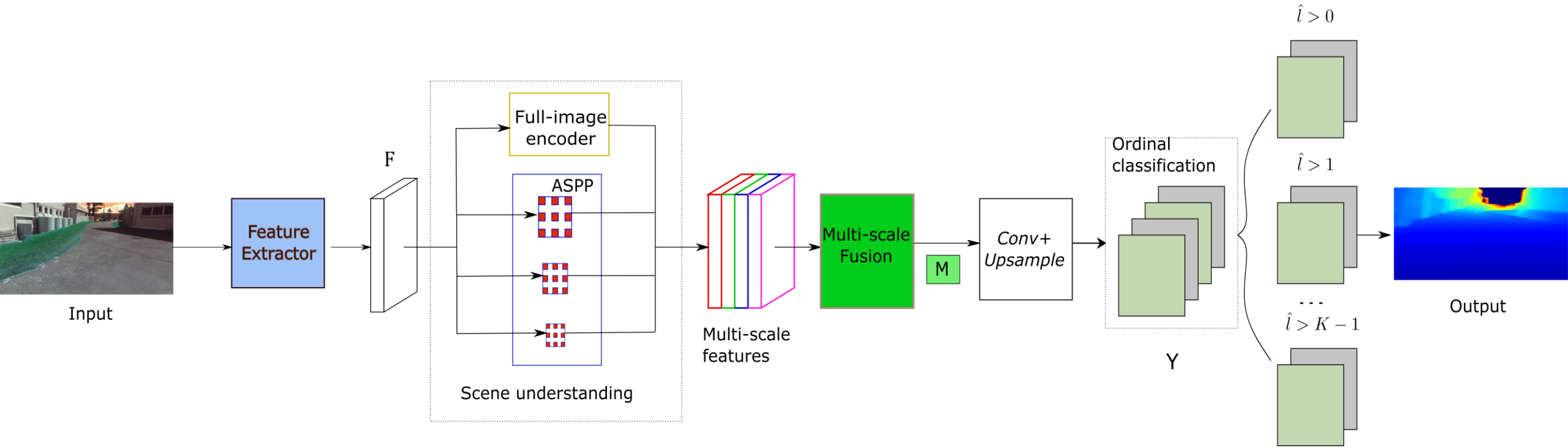
Human uses different depth cues when interpreting the scene, depending on the texture, and our prior knowledge[1]

[1] Knill, Vision Research, 2003

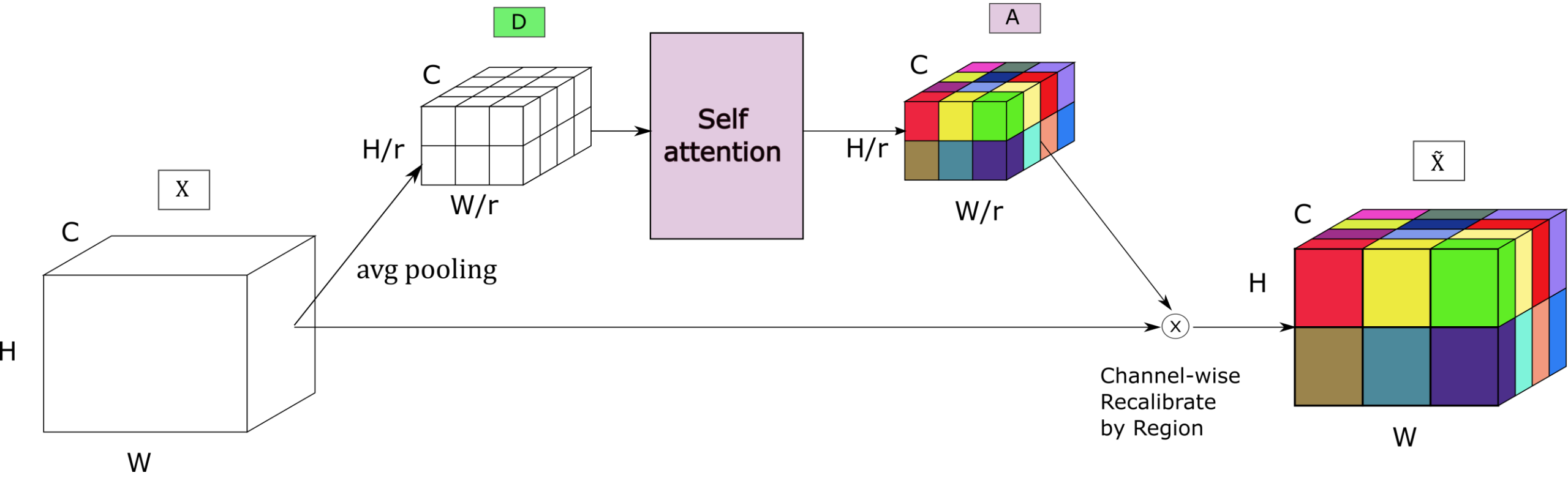
Region-based Self-attention



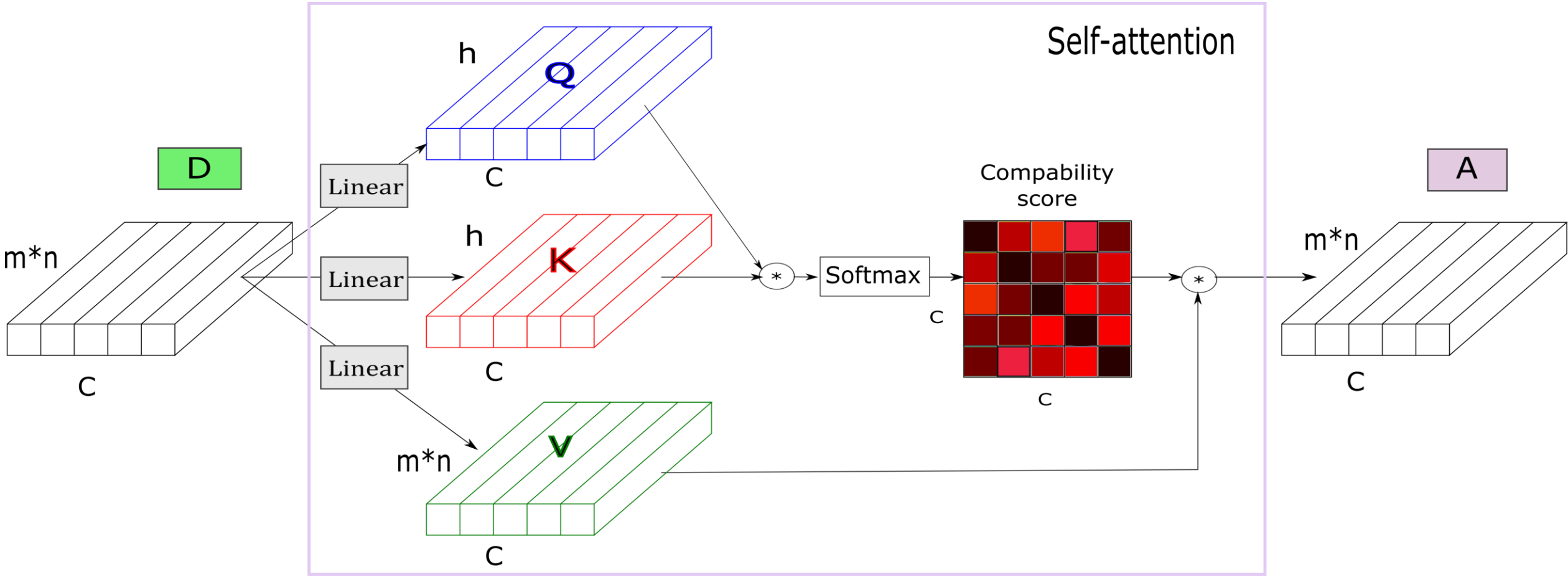
Network Architecture



Region-based Self-attention



Self-attention for Multi-scale fusion



Baseline methods

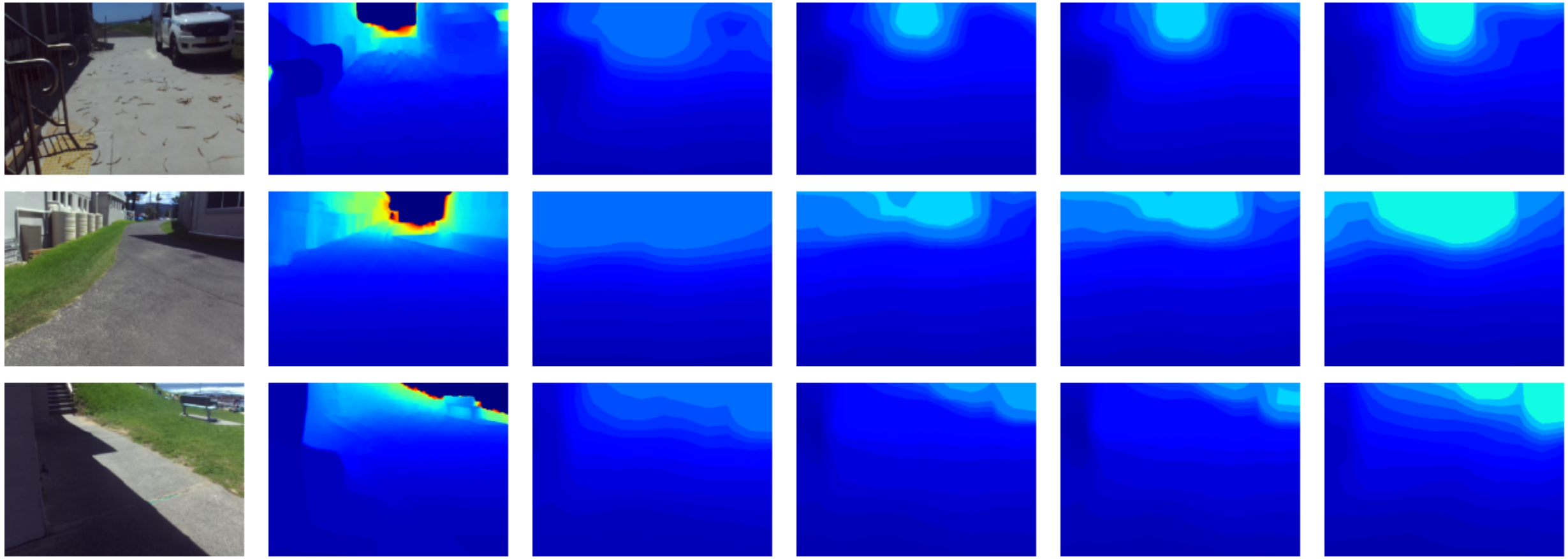
Method	Description
DORN[1]	Deep ordinal regression network
DORN + cSE[2]	Channel-wise squeeze-excitation
DORN + sSE[3]	Spatial squeeze-excitation
DORN + scSE[3]	Spatial-channel squeeze-excitation
DORN + rSA	Proposed method

[1] Fu et al., CVPR, 2018.
[2] Hu et al., CVPR, 2018.
[3] Roy et al., MICCAI, 2018.

Dataset

Name	Train set	Test set	Environment	Resolution
NYU-Depth-v2	120K	694	Indoor	480 x 640
Make3D	400	134	Outdoor	2272 x 1704
UoW	3087	1323	Outdoor	1920 x 1080

Qualitative Evaluation



(a) RGB image (b) Ground-truth (c) DORN (d) DORN+cSE (e) DORN+scSE (f) Proposed method

Fig. 1. **Visual results of depth estimation on the UOW dataset by several methods.**

Quantitative Evaluation Metrics

Metrics	Formulation
Threshold (δ_i)	% of $\hat{d}_i$ such that $\max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) = \delta_i < 1.25^i$
Root mean squared error (RMSE)	$\sqrt{\frac{1}{N}\sum_i (d_i - \hat{d}_i)^2}$
Relative error	$\frac{1}{N}\sum_i \lvert d_i - \hat{d}_i \rvert / d_i$
Error in log scale (log10)	$\frac{1}{N}\sum_i \lvert \log_{10} d_i - \log_{10} \hat{d}_i \rvert$

NYU Depth v2

Method	RMSE	log10	Rel	δ_1	δ_2	δ_3
DORN[1]	0.632	0.075	0.175	74.7	93.8	98.2
DORN+cSE[2]	0.734	0.086	0.192	66.1	89.4	97.6
DORN+sSE[3]	0.824	0.102	0.213	54.3	92.4	98.0
DORN+scSE[3]	0.704	0.086	0.206	67.3	90.0	97.0
DORN+rSA	0.623	0.071	0.155	77.1	95.2	98.8

[1] Fu et al., CVPR, 2018.
[2] Hu et al., CVPR, 2018.
[3] Roy et al., MICCAI, 2018.

UoW Dataset

Method	RMSE	log10	Rel	δ_1	δ_2	δ_3
DORN[1]	2.701	0.260	0.515	27.7	50.8	67.3
DORN+cSE[2]	2.449	0.214	0.363	38.6	61.2	75.4
DORN+sSE[3]	2.502	0.221	0.437	35.0	60.9	74.3
DORN+scSE[3]	2.355	0.203	0.369	39.5	63.6	77.8
DORN+rSA	1.607	0.121	0.316	57.5	80.4	90.1

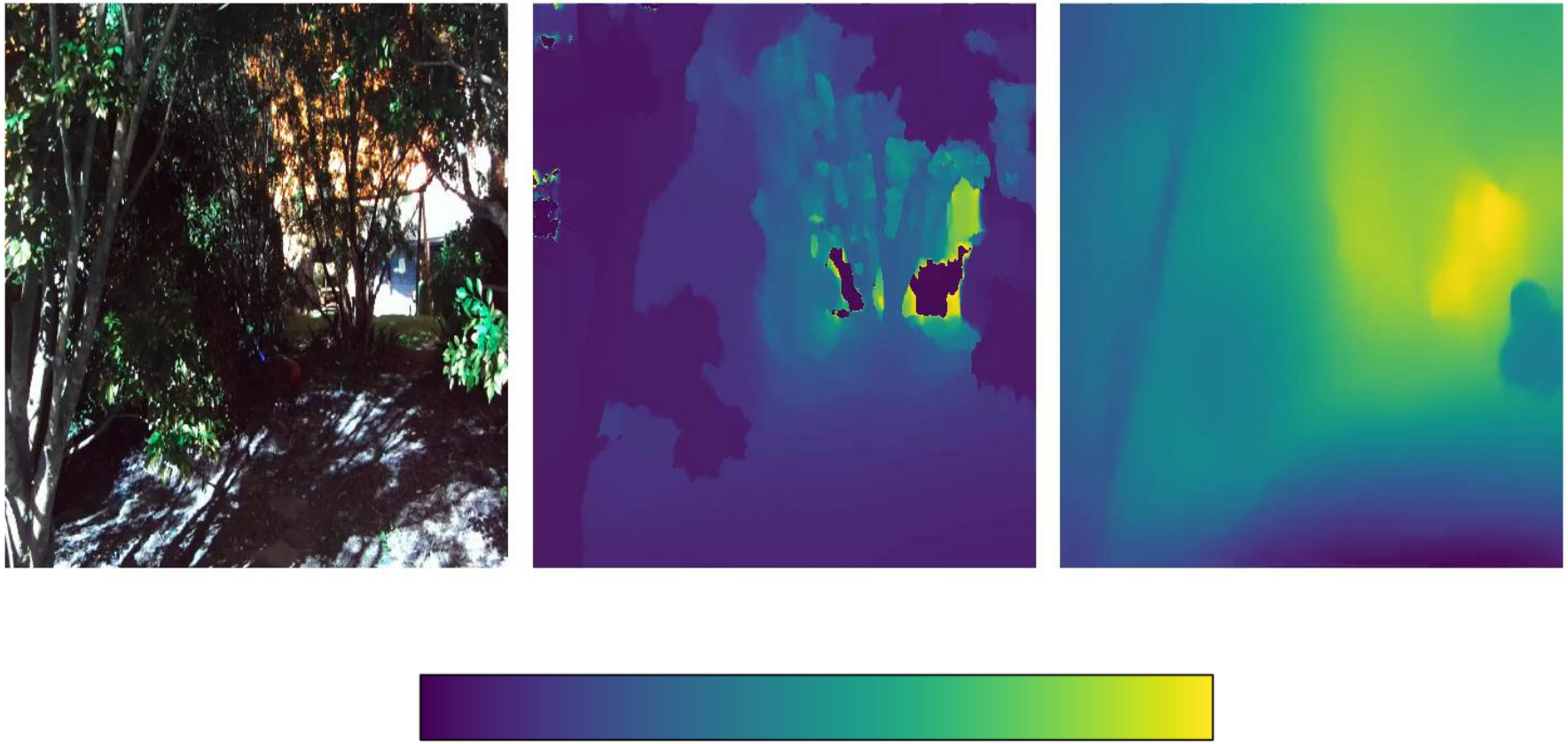
[1] Fu et al., CVPR, 2018.
[2] Hu et al., CVPR, 2018.
[3] Roy et al., MICCAI, 2018.

Make3D Dataset

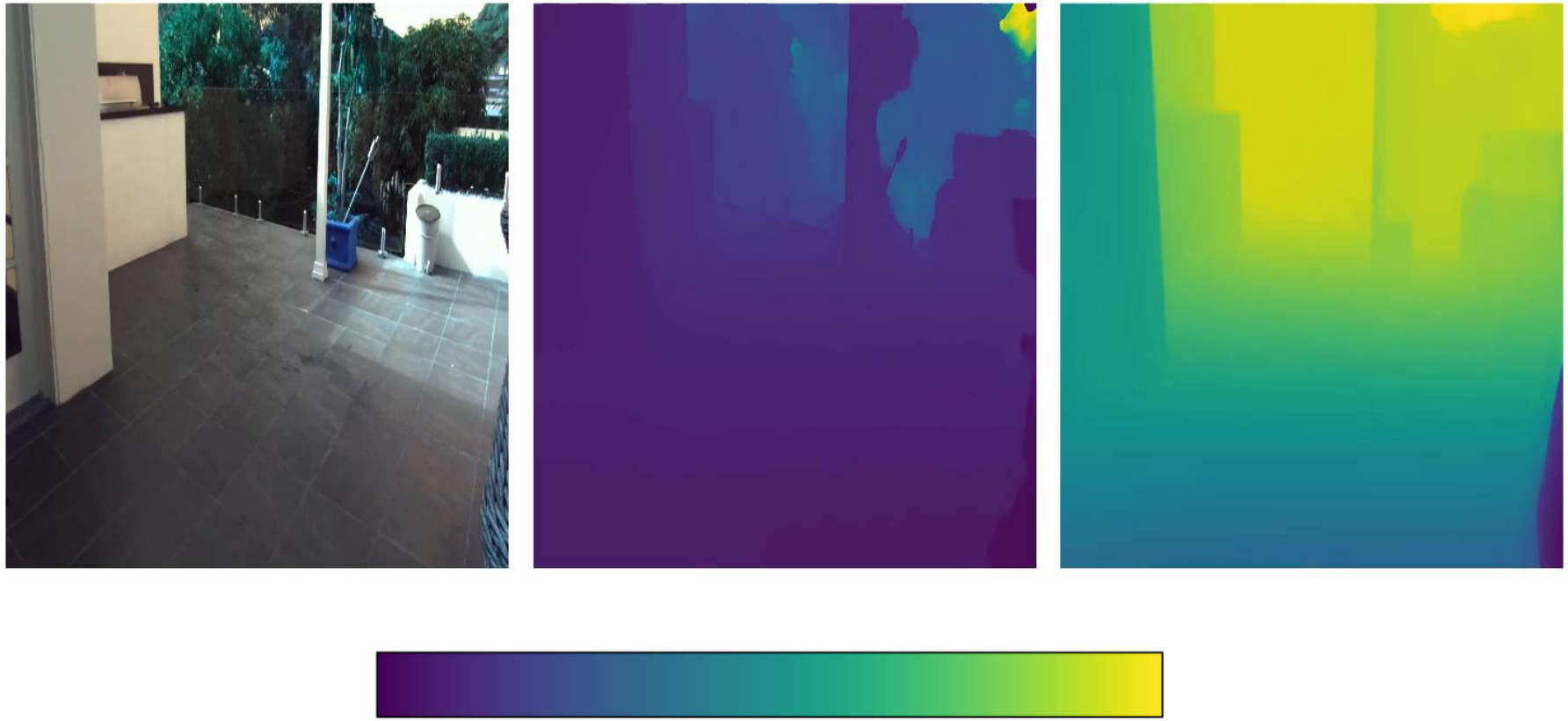
Method	RMSE	log10	Rel
DORN[1]	14.32	0.125	0.289
DORN+cSE[2]	12.62	0.118	0.265
DORN+sSE[3]	15.71	0.141	0.318
DORN+scSE[3]	12.11	0.113	0.267
DORN+rSA	10.15	0.094	0.256

[1] Fu et al., CVPR, 2018.
[2] Hu et al., CVPR, 2018.
[3] Roy et al., MICCAI, 2018.

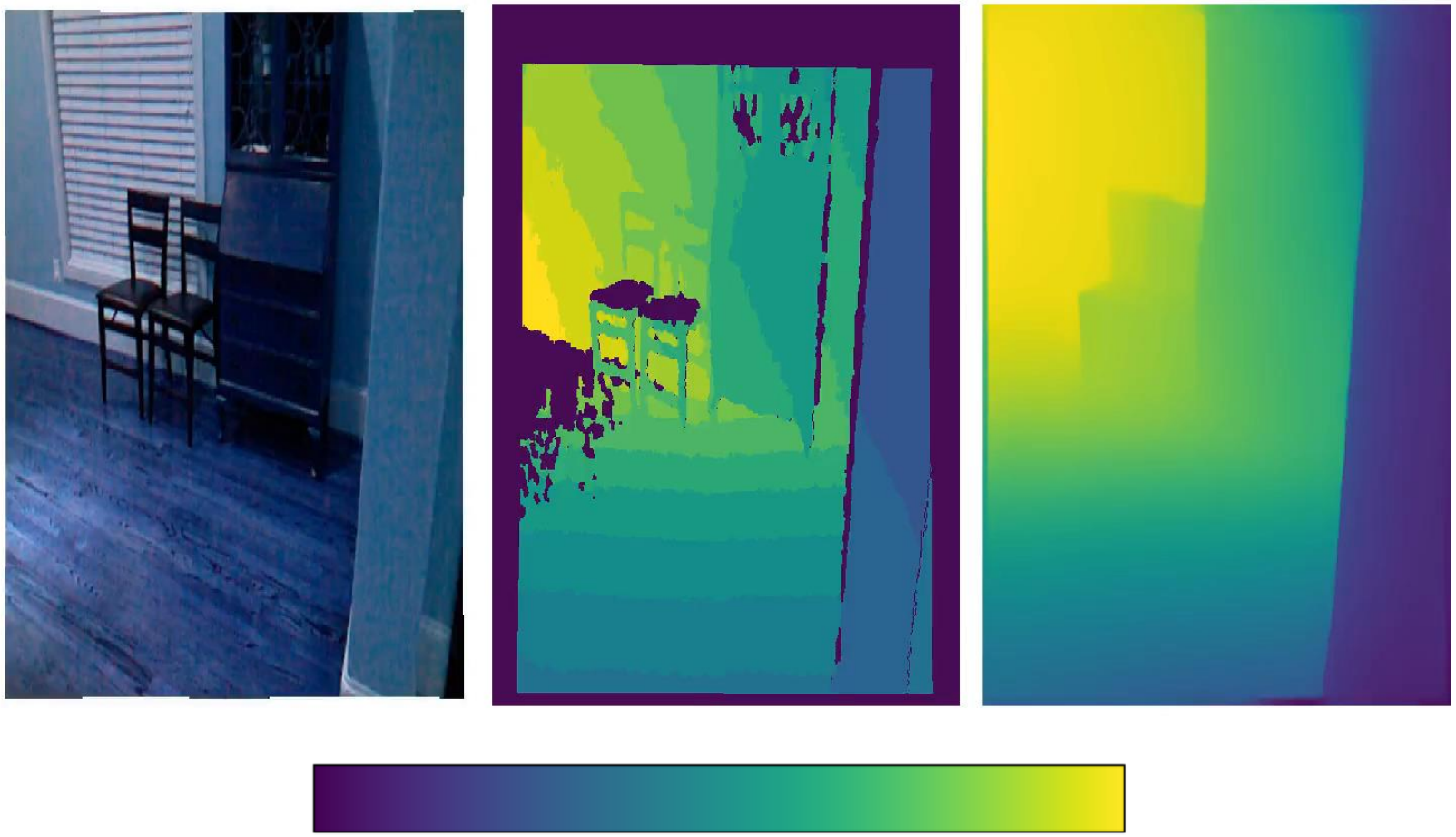
Demo of Depth Estimation



Demo of Depth Estimation



Demo of Depth Estimation



Conclusion

- Propose self-attention module that calibrates features at different regions.
- Enhance the ability to fuse multi-scale features.
- Improve the performance of Make3D and UoW dataset by 29.12% and 40.50%.
- Identify weak features and prune the networks to produce faster computation while retaining the accuracy.