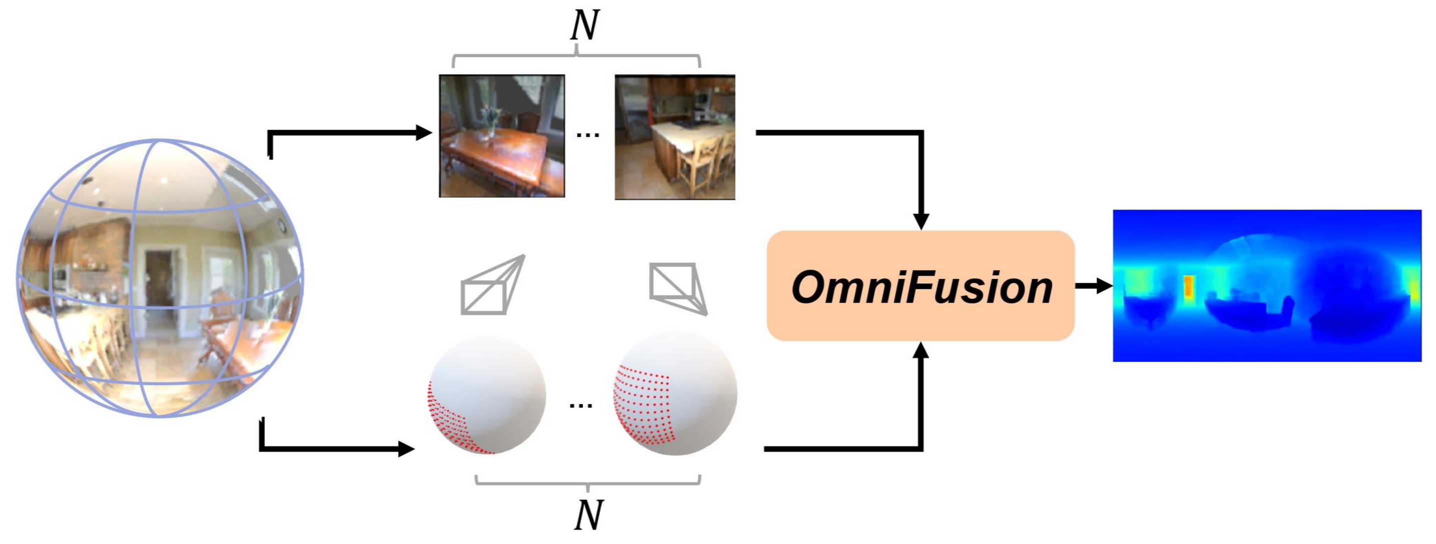
OmniFusion: 360 Monocular Depth Estimation via Geometry-Aware Fusion

Yuyan Li[1], Yuliang Guo[2], Zhixin Yan[2], Xinyu Huang[2], Ye Duan[1], Liu Ren[2]   [1]University of Missouri   [2]Bosch Research North America

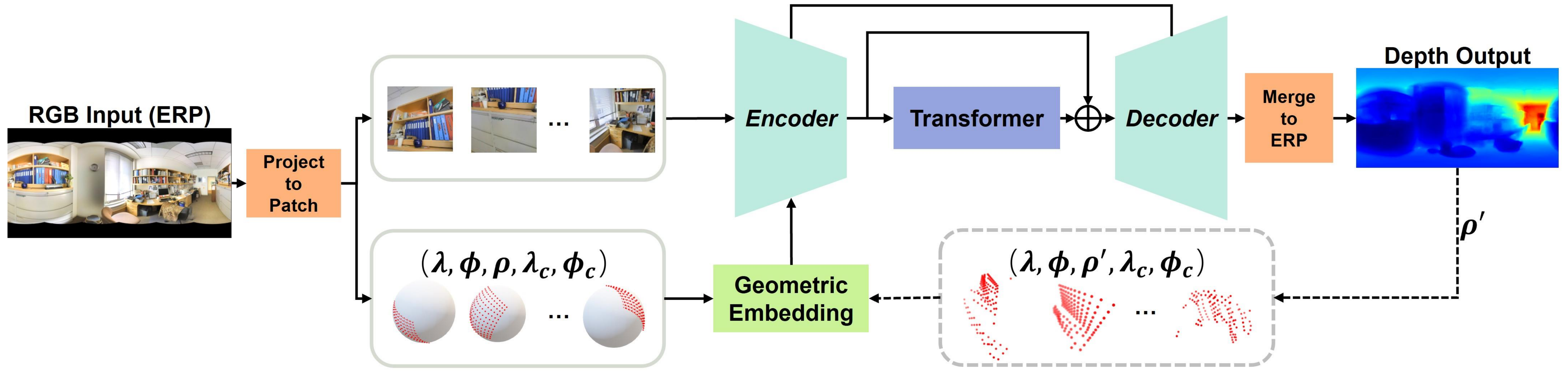
CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

INTRODUCTION



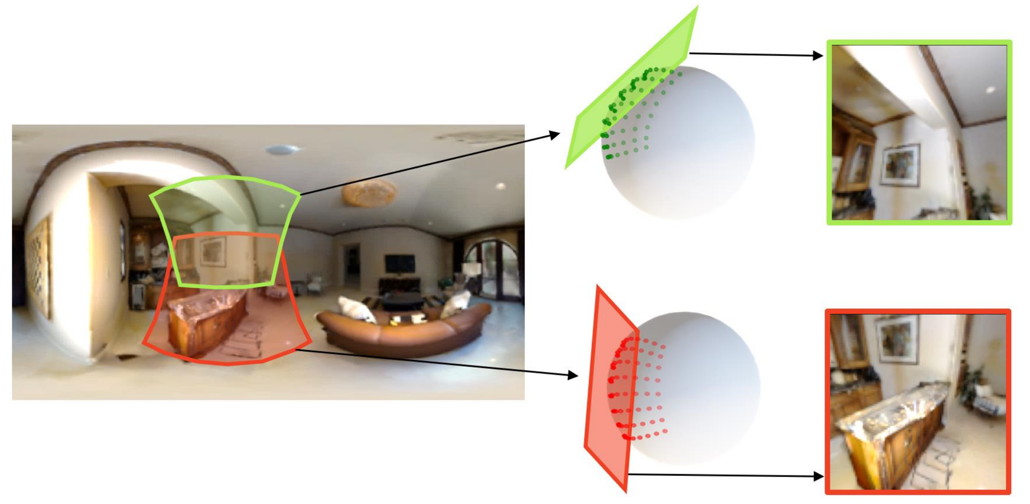
- We present a framework, ***OmniFusion***, for producing high-quality 360 depth from a monocular 360 RGB input.

OMNIFUSION PIPELINE



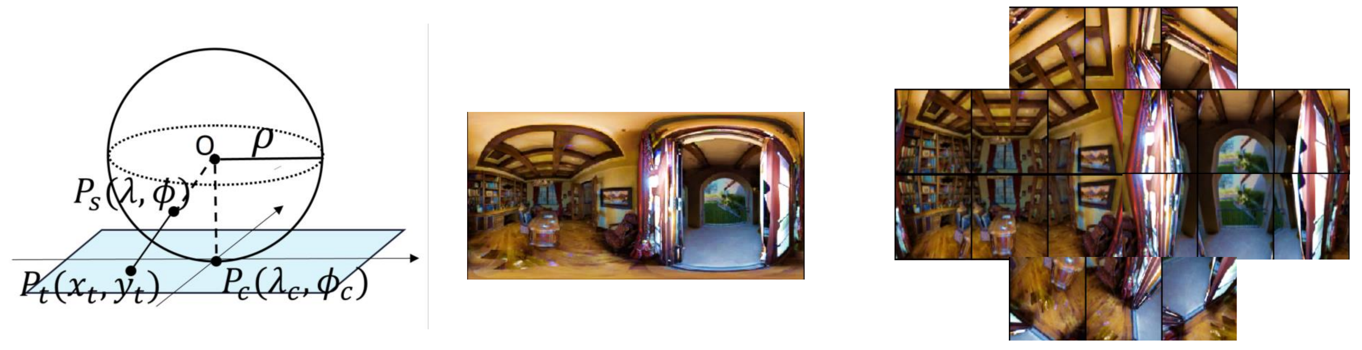
- A 360 monocular depth prediction pipeline that addresses the distortion issue.
- A geometric embedding network to provide 3D geometric features to mitigate the discrepancy.
- A self-attention transformer to globally aggregate information which enhances the depth scale estimation.
- An iterative mechanism to further improve the depth estimation with structural details.

DISCREPANCY



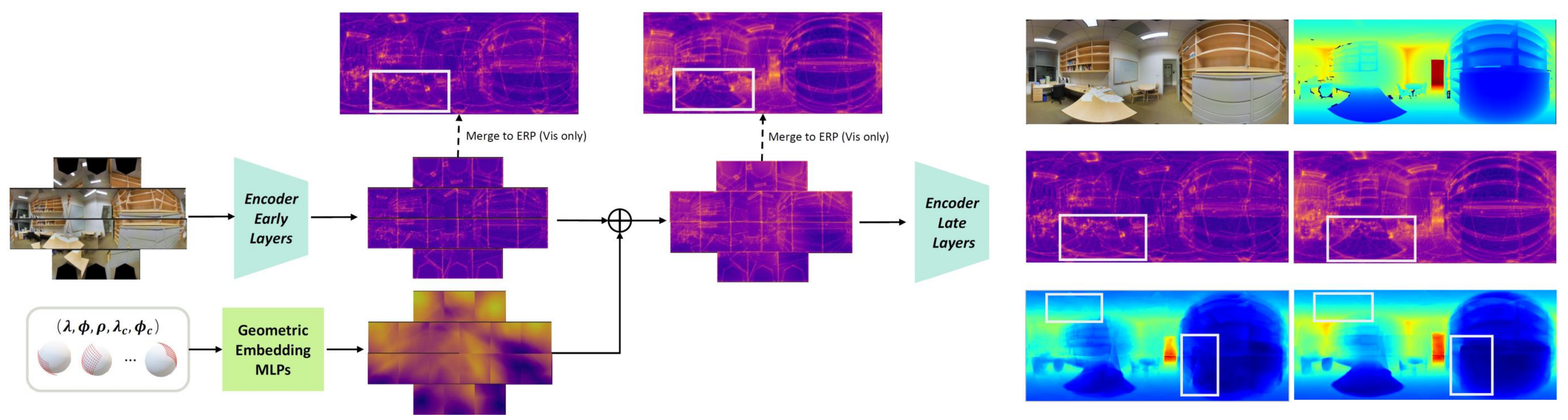
- There are overlapping areas between two neighboring patches. The same object may appear differently in different patches.
- Discrepancy could greatly harm the quality of depth fusion from multiple patches.

TANGENT IMAGE



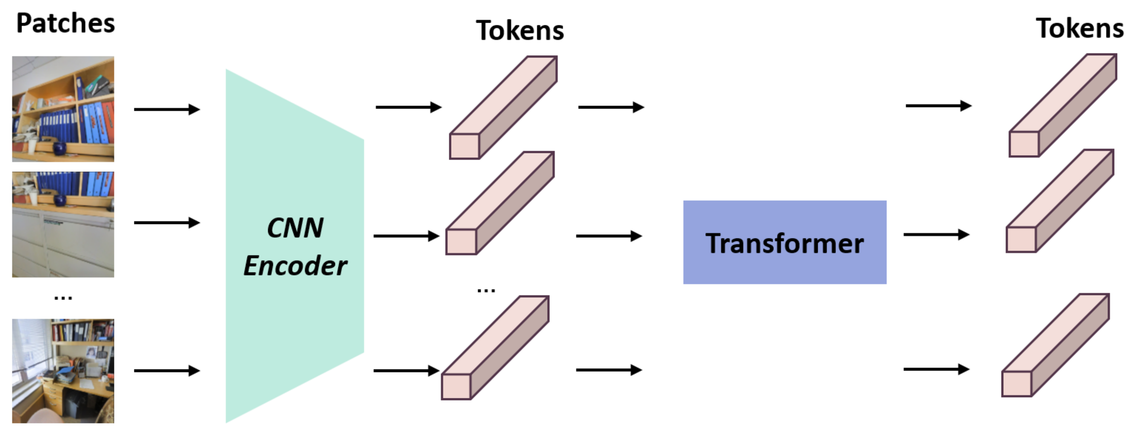
- A tangent image is generated via gnomonic projection of a sphere surface onto a flat, rectangular plane.
- We generate 18 tangent patches from a single ERP input.

GEOMETRIC FUSION



- 3D coordinates are encoded into geometric feature maps.
- 2D image features are fused with the patch-wise geometric features.
- With geometric fusion, more structural and cleaner depth is produced.

TRANSFORMER



- The self-attention transformer builds global patch-to-patch relationships.
- Depth scales across patches are more consistent.

EXPERIMENTAL RESULTS

- Study 1: patch size and number of patches

Encoder	#iters	FPS↑	Abs Rel↓	Sq Rel↓	RMSE↓
ResNet18	1	**9.8**	0.1037	0.0589	0.3686
ResNet18	2	4.6	0.0979	0.0539	0.3702
ResNet18	3	3.1	**0.0981**	0.0521	**0.3699**
ResNet18	4	1.5	0.0983	**0.0519**	0.3700
ResNet34	1	**9.2**	0.0961	0.0543	0.3715
ResNet34	2	4.6	0.0950	0.0491	**0.3474**
ResNet34	3	2.9	**0.0894**	**0.0482**	0.3498
ResNet34	4	1.4	0.0899	0.0485	0.3491

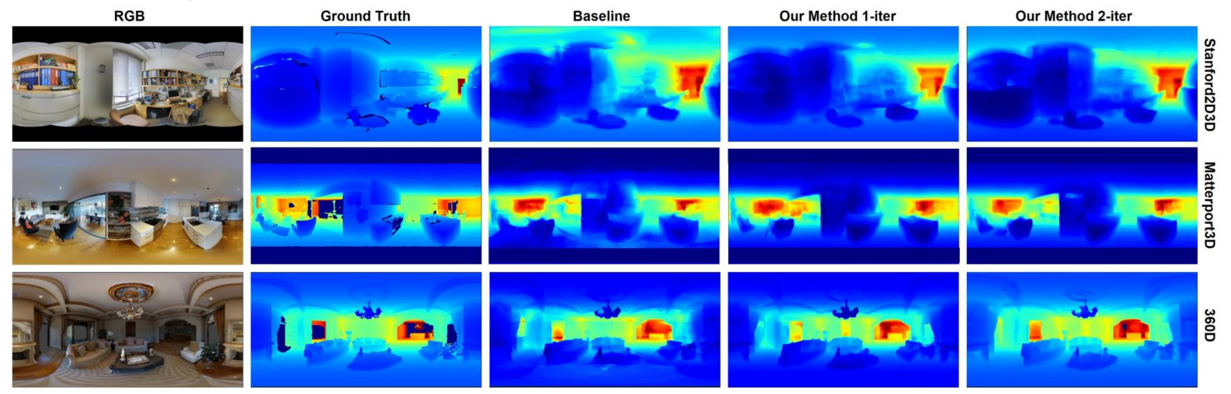
- Study 2: variations of encoders and iterations

#patch	Patch size	Patch FoV	Abs Rel↓	Sq Rel↓	RMSE↓
10	256x256	120	0.1067	0.0571	0.3788
18	128x128	80	0.1178	0.0666	0.4018
18	**256x256**	**80**	**0.1037**	**0.0589**	**0.3686**
26	256x256	60	0.1104	0.0679	0.3955
46	128x128	50	0.1181	0.0680	0.4101

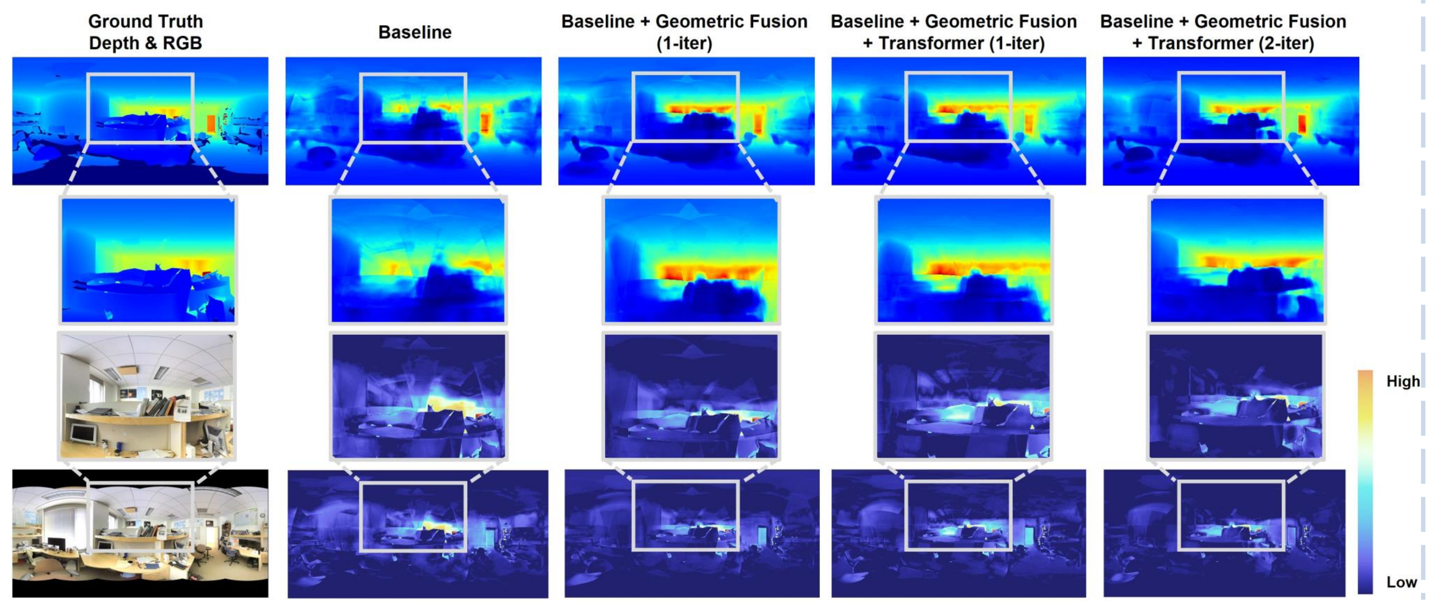
- Study 3: individual components

Methods	#Params	FPS↑	Abs Rel↓	Sq Rel↓	RMSE↓
Baseline	23.5M	9.4	0.1136	0.0638	0.3894
Baseline + geometric fusion (1-iter)	23.5M (+1.3K)	9.3	0.1026	0.0588	0.3812
Baseline + geometric fusion + transformer (1-iter)	42.3M (+18.8M)	9.2	0.0961	0.0543	0.3715
Baseline + geometric fusion + transformer (2-iter)	42.3M (+18.8M)	4.6	**0.0950**	**0.0491**	**0.3474**

- Qualitative results on different datasets



- Qualitative results regarding individual components



- Qualitative comparisons with other existing methods