A Two-Stage Masked Autoencoder Based Network for Indoor Depth Completion

Kailai Sun[1,2,*], Zhou Yang[1,*], Qianchuan Zhao[1]
[1]Tsinghua University, [2]National University of Singapore (*Equal contribution)

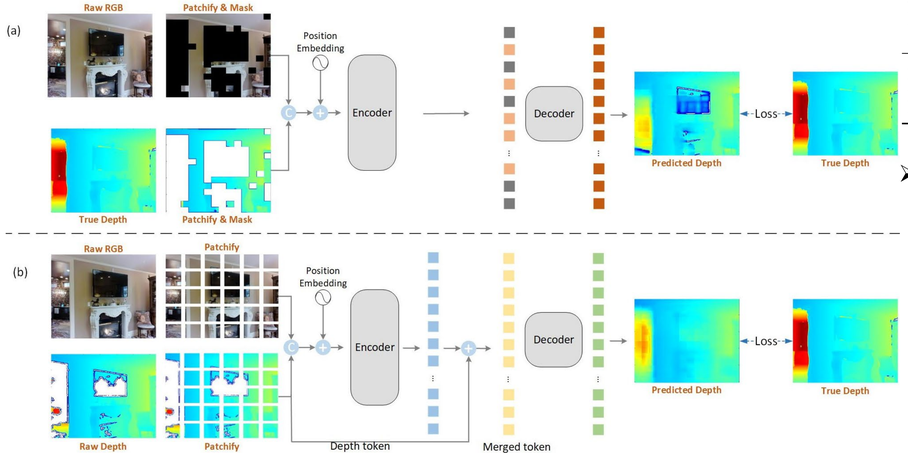
CVPR SEATTLE, WA JUNE 17-21, 2024

- Paper, code, and demo are available:
- kailaisun/Indoor-Depth-Completion

Introduction:

- **Problem:** Depth image captured by RGBD cameras are influenced by illumination or the materials of the objects.
- **Goal:** A novel two-stage approach provide more accurate and complete depth images.
- **Key idea:** Missing depth patches ⇐ Simulate? Masks

Method:



- (a) First stage: Masked Autoencoder (MAE)-based Self-supervision Pre-training. It aims to learn an effective latent representation from the jointly masked RGB and depth images. Unlike MAE, we only mask the missing depth areas; in loss function, we only consider the parts where the depth value exceeds zero.

$$RMSE(D, DT) = \sqrt{\frac{1}{|O|}\sum_{p \in O} \|D(p) - DT(p)\|^2}$$

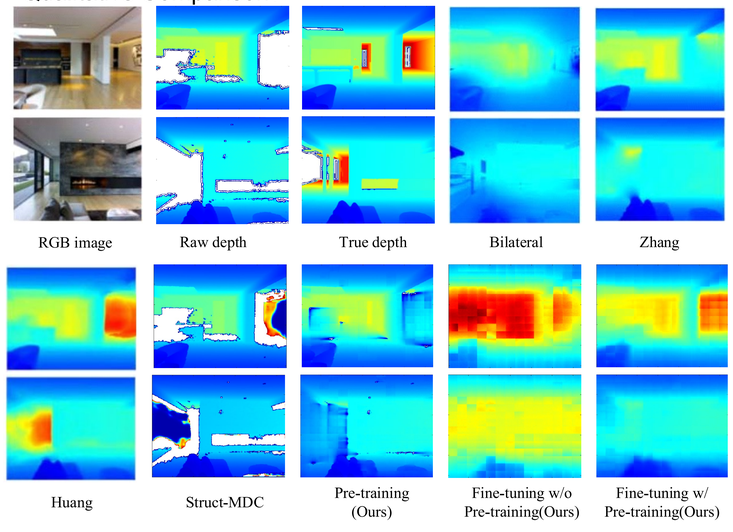
- (b) Second stage: Indoor Depth Completion Based on Supervised Fine-tuning. It aims to learn a decoder based on token fusion to complete (reconstruct) the full depth from an incomplete depth image.

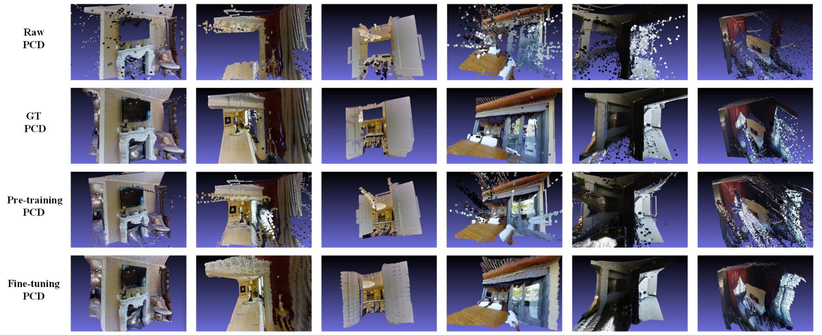
Experiments:

- Quantitative Comparison:

Methods	RMSE↓	ME↓	SSIM↑	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑
Joint Bilateral Filter	1.978	0.774	0.507	0.613	0.689
MRF[1]	1.675	0.618	0.692	0.651	0.780
AD[2]	1.653	0.610	0.696	0.663	0.792
FCN	1.262	0.517	0.605	0.681	0.808
Zhang[3]	1.316	0.461	0.762	0.781	0.851
Huang[4]	1.092	0.342	**0.799**	0.850	0.911
Struct-MDC[5]	1.060	0.503	0.534	0.656	0.713
Pre-training	1.216	0.675	0.642	0.705	0.800
Fine-tuning w/o Pre-training	**0.660**	0.243	0.654	0.794	0.904
Fine-tuning w/ Pre-training	**0.690**	**0.206**	**0.765**	**0.852**	**0.912**

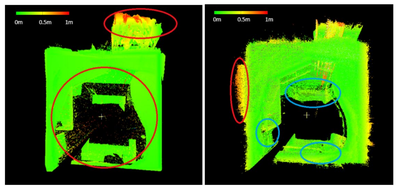
- Qualitative Comparison:





Additional Applications:

- Indoor 3D Reconstruction with completed depth images performs better than uncompleted depth images
- ORB SLAM 3 method was used to reconstruct indoor scene on ICL-NUIM dataset.



Reconstruction Error Before/After Depth Completion

Methods	Mean (m)↓	Median (m)↓	Standard Deviation (m)↓	Minimum (m)	Maximum (m)↓
Depth uncompletion	0.138	0.053	0.200	0.0	1.106
Depth completion	**0.086**	0.057	**0.101**	**0.0**	**1.100**

References

Jinwoo Jeon, Hyunjun Lim, Dong-Uk Seo, and Hyun Myung. Struct-mdc: Mesh-refined unsupervised depth completion leveraging structural regularities from visual slam. IEEE Roboticsand Automation Letters,7(3):6391–6398,2022.2

Huang YK, WuTH, Liu YC, et al. Indoor depth completion with boundary consistency and self-attention. ICCV. 2019.