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

Kailai Sun, Zhou Yang, Qianchuan Zhao

Email: skl23@nus.edu.sg

2024.6.18

Project:

WeChat:

Opening

- Scan to BIM: a workflow or process that translates scanned, point-cloud digital models into building information modeling (BIM) platforms

- Indoor 3D reconstruction [1]: create a 3D digital spatial information representation of the interior of a building

A point cloud is displayed in Autodesk ReCap

[1] Shayan Nikoohemat, et al. Indoor 3D reconstruction from point clouds for optimal routing in complex buildings to support disaster management. Automation in Construction, 113, 2020.

Challenge

- Depth completion: an important task focuses on using part of the depth data measured in the real scene to obtain more dense and complete depth data.

- Cause: illumination or the materials of the scene objects, limited distance

- However, the latest methods often suffer from sensitivity to dynamic environmental lighting conditions.

Matterport3D dataset

15% depth values are missing

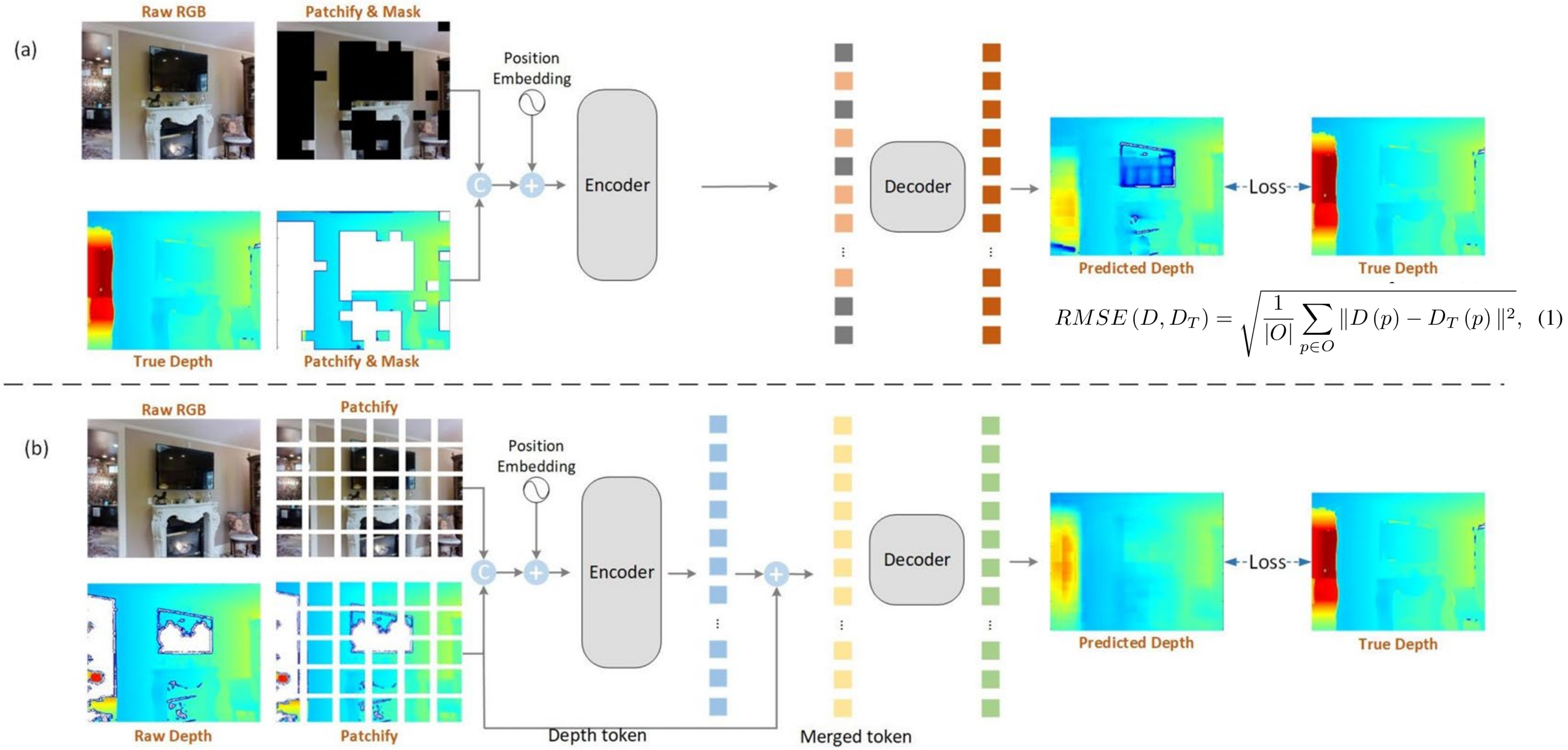
Insight

- Masked Autoencoder only apply partial observation to reconstruct the entire image, learning robust features and improving the generalization ability.
- We consider: Missing depth patches $\xleftarrow{\text{simulate}}$? Masks.
- We propose a Vision Transformer-based two-stage network for indoor depth completion:

(1) an MAE-based self-supervision pre-training encoder to learn an effective latent representation from the jointly masked RGB and depth images;

(2) a decoder based on token fusion to complete (reconstruct) the full depth from an incomplete depth image.

Method



(a)
Raw RGB
Patchify & Mask
True Depth
Patchify & Mask
Position Embedding
C
Encoder
Decoder
Predicted Depth
Loss
True Depth
$RMSE\left(D,D_T\right)=\sqrt{\frac{1}{|O|}\sum_{p\in O}\left\|D\left(p\right)-D_T\left(p\right)\right\|^2}, \quad (1)$
(b)
Raw RGB
Patchify
Raw Depth
Patchify
Position Embedding
C
Encoder
Depth token
Merged token
Decoder
Predicted Depth
Loss
True Depth

Result

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**

$$ME\left(D, D_T\right) = \frac{1}{|O|} \sum_{p \in O} \left\| D\left(p\right) - D_T\left(p\right) \right\|$$

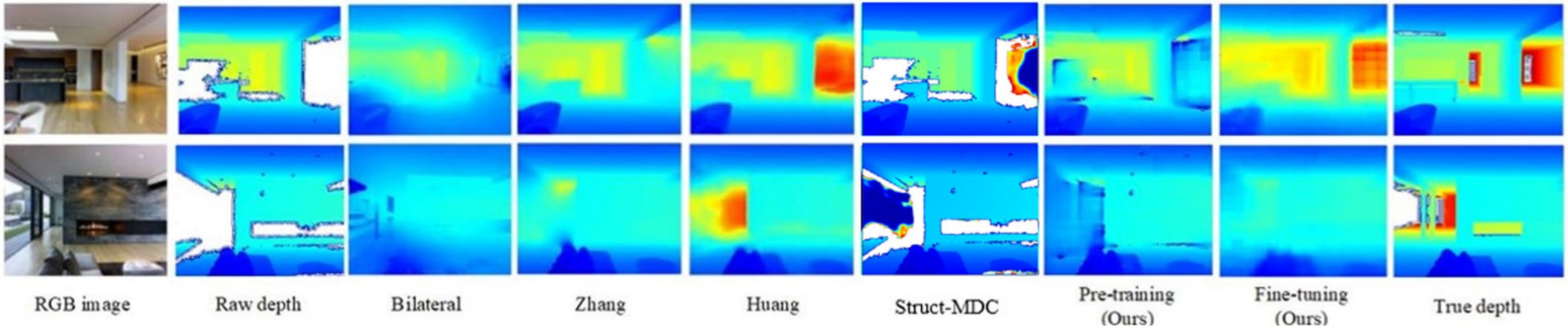
$$SSIM\left(D, D_T\right) = \frac{\left(2\mu_{D_T}\mu_D + c_1\right)\left(2\sigma_{D_T D} + c_2\right)}{\left(\mu_{D_T}^2 + \mu_D^2 + c_1\right)\left(\sigma_{D_T}^2 + \sigma_D^2 + c_2\right)}$$

$$Max\left(\frac{D\left(p\right)}{D_T\left(p\right)}, \frac{D_T\left(p\right)}{D\left(p\right)}\right) < t,$$

The pre-training model is better than the traditional methods (e.g., Joint bilateral filter and MRF) and the method of Zhang on RMSE.

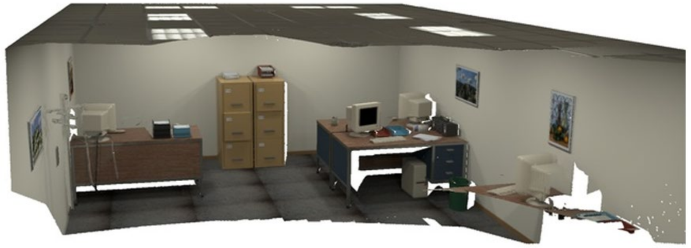
Our fine-tuning model achieves superior performance on the Matterport3D dataset, and performs best on δ and ME.

Result


RGB image
Raw depth
Bilateral
Zhang
Huang
Struct-MDC
Pre-training (Ours)
Fine-tuning (Ours)
True depth


Raw PCD
GT PCD
Pre-training PCD
Fine-tuning PCD

Application: Indoor 3D Reconstruction



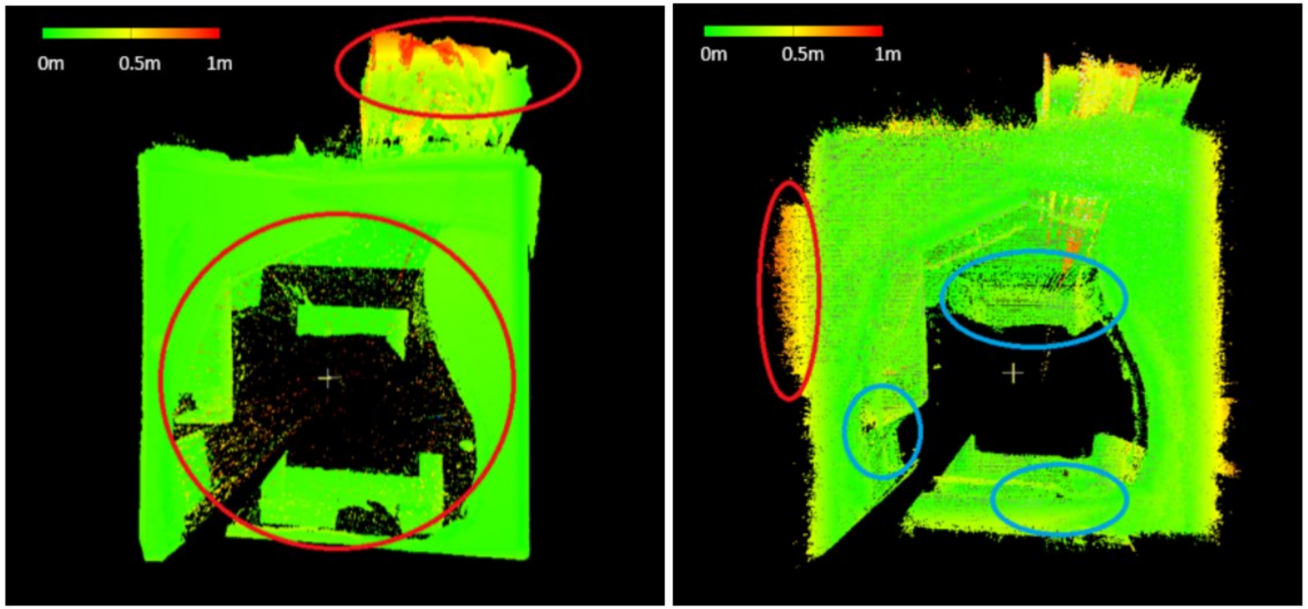
ICL-NUIM dataset

↓

Depth completion

↓

ORB-SLAM3



Reconstruction errors before/after depth completion

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

Thank you.

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

Kailai Sun, Zhou Yang, Qianchuan Zhao

Email: skl23@nus.edu.sg

2024.6.18

Project:

WeChat: