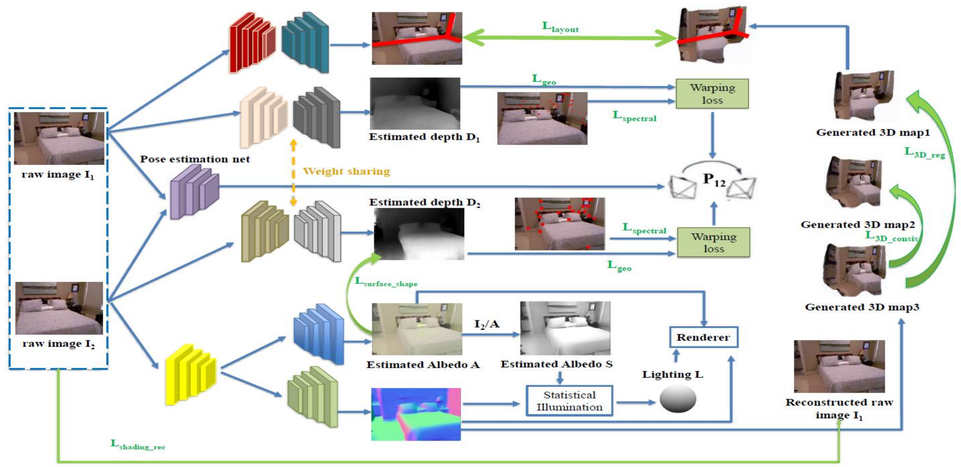
Shading Meets Motion: Self-supervised Indoor 3D Reconstruction Via Simultaneous Shape-from-Shading and Structure-from-Motion

Guoyu Lu (guoyulu62@gmail.com)
Intelligent Vision and Sensing Lab, SUNY Binghamton, NY, USA

Overview:



We propose to train the shading network and structure-from-motion network simultaneously to overcome the potential failure on low-texture scenes. Blue arrows and lines represent the data flow and predictions; Green lines mark the loss constraints for guiding the learning process.

Methodology Details:

A. Shading Navigated Learning Scheme

- Shading network takes an image as the input of the encoder, and learns the representation to generate the estimated albedo map A, surface normal map N, shading map S, and Light resource L.

$$SfS\text{-}Net : \psi(I) \rightarrow \{A, N, (S), (L)\}$$

- With the estimated Albedo A, the estimated surface normal N, the inferred shading S and light source L, the reconstructed image can be rendered by A·S =A·Lf(N)

$$L_{shading\_rec} = \|L * A * B * (I_{raw}) - L * A * B * (I_{rec})\|_1$$

B. Adaptive Brightness and Scale-invariant SfM

- To mitigate the influence of dramatic image intensity variation between two consecutive images, we use two trainable parameters m and b based on the brightness constancy assumption, to model the adjacent frames' brightness. I′= m∗ I +b.
- Apply keypoint and brightness based consistencies as supervisions to guide the learning.

$$L_{spectral} = \frac{1}{N}\sum_{ij} |I'_{t+1}(\omega(\mathbf{Kp}(i,j), T_{t\rightarrow t+1})) - I'_t(\mathbf{Kp}(i,j))|$$

$$L_{geo} = \frac{1}{N}\sum_{ij} |D'_t(\omega(\mathbf{Kp}(i,j), T_{t\rightarrow t+1})) - D_t(\mathbf{Kp}(i,j))|$$

C. Layout Consistency Constraint

- Align the structural layout lines detected from the input image with those extracted from the back-projected image rendered from the reconstructed 3D scene.
- Encourage the reconstructed geometry and camera pose to faithfully preserve dominant layout structures such as wall-floor and wall-wall boundaries, even in partially visible scenes.

$$\mathcal{C} = \left\{ (\ell_i^{\text{img}}, \ell_j^{\text{proj}}) \;\middle|\; \text{match}(i) = j \text{ and } \text{match}^{-1}(j) = i \right\}$$

$$L_{\text{layout}} = \frac{1}{|\mathcal{C}|} \sum_{(\ell^{\text{img}}, \ell^{\text{proj}}) \in \mathcal{C}} \big(\|(x_1, y_1)^{\text{img}} - (x_1, y_1)^{\text{proj}}\|_1 + \|(x_2, y_2)^{\text{img}} - (x_2, y_2)^{\text{proj}}\|_1 + \lambda_s |a^{\text{img}} - a^{\text{proj}}| + \lambda_b |b^{\text{img}} - b^{\text{proj}}| \big)$$

D. Joint Framework Optimization

- **Surface shape consistency**: Depth and normal vectors from the shading network and SfM should be consistent.

$$P(p) = \left[\tfrac{x-c_x}{f} D(p), \tfrac{y-c_y}{f} D(p), D(p)\right] \qquad \frac{\partial P}{\partial x} = \begin{bmatrix} \frac{-1}{f}(x-c_x)\frac{\partial D}{\partial x} - \frac{1}{f}D \\ \frac{-1}{f}(y-c_y)\frac{\partial D}{\partial x} - \frac{1}{f}D \\ \frac{\partial D}{\partial x} \end{bmatrix} \qquad \frac{\partial P}{\partial y} = \begin{bmatrix} \frac{-1}{f}(x-c_x)\frac{\partial D}{\partial y} - \frac{1}{f}D \\ \frac{-1}{f}(y-c_y)\frac{\partial D}{\partial y} - \frac{1}{f}D \\ \frac{\partial D}{\partial y} \end{bmatrix}$$

$$L_{cross-normal} = \frac{1}{N}\sum_i (1 - \frac{< \tilde{d}_i, \tilde{n}_i >}{||\tilde{d}_i|| \cdot ||\tilde{n}_i||})$$

- **Scene registration consistency**: We establish a global scene registration consistency between the two networks for better shape completion.

$$L_{3D\_reg} = \sum_{P^{t+1}_{SfM}} P^t_{SfS} ||T \cdot P^{t+1}_{SfM} - P^t_{SfS}||^2_2 + \sum_{P^t_{SfS}} P^{t+1}_{SfM} ||T^{-1} \cdot P^t_{SfS} - P^{t+1}_{SfM}||^2_2$$

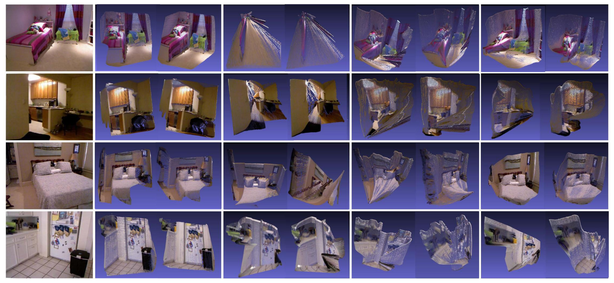
$$L_{3D\_consis} = \sum_{P^t_{SfM}} P^t_{SfS} ||P^t_{SfM} - P^t_{SfS}||^2_2 + \sum_{P^t_{SfS}} P^t_{SfM} ||P^t_{SfS} - P^t_{SfM}||^2_2$$

Acknowledgement: This work is supported by NSF Awards NO. 2340882, 2334624, 2334246, and 2334690.

Contribution:

1. We propose Shading-SfM-Net, leveraging the benefits from both SfS and SfM networks simultaneously, to aid scene reconstruction
2. The network introduces an adaptive keypoint and layout-based consistency and a curved surface consistency between Albedo and depths, which firmly constrain the learning process on low and non-textured regions
3. The 3D point clouds from SfM at different time are further optimized by registering with the point cloud from shape-from-shading network, to build a 3D geometric registration loss, which exploits their mutual information and overcomes the spatial confusion of SfM systems in low or no texture regions of challenging indoor images.
4. The unsupervised manner in which the entire framework is designed mitigates the training requirement while enhancing the robust generalization capability.

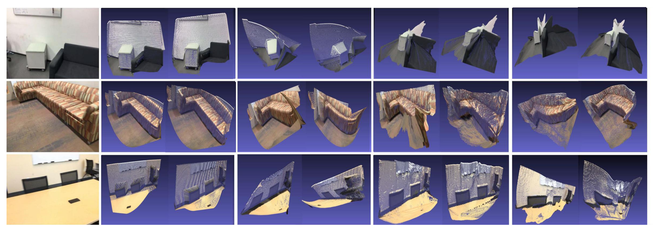
Experiments



Visual comparisons of 3D reconstruction on NYUv2 dataset. From left to right: input (1st column); result from our method (2nd column), Lite-mono (3rd column), HR-depth (4th column), and RA-Depth (5th column). Each reconstructed 3D scene is shown from two perspectives for comparison.

Methods	Type	Error			Acc		
		AbsRel	RMS	Log10	δ_1	δ_2	δ_3
Wang et al. [69]	S	0.220	0.824	-	0.605	0.890	0.970
Eigen et al. [16]	S	0.158	0.641	-	0.769	0.950	0.988
Liu et al. [44]	S	0.213	0.759	0.087	0.650	0.906	0.976
Li et al. [40]	S	0.143	0.635	0.063	0.788	0.958	0.991
Xu et al. [73]	S	0.143	0.613	-	0.789	0.946	0.984
DORN [21]	S	0.115	0.509	0.051	0.828	0.965	0.992
SharpNet [61]	S	0.139	0.502	0.047	0.836	0.966	0.993
Adabins [8]	S	0.103	0.364	0.044	0.903	0.984	0.997
P3Depth [57]	S	0.104	0.356	0.043	0.898	0.981	0.996
NeWCRFs [80]	S	0.095	0.334	0.041	0.922	**0.992**	**0.998**
Trap Attention [56]	S	**0.090**	**0.329**	**0.039**	**0.927**	0.991	**0.998**
GeoNet [77]	U	\	\	\	\	\	\
SC-SfMLearner [10]	U	0.239	0.832	0.098	0.638	0.897	0.964
Monodepth2 [24]	U	0.176	0.639	0.074	0.734	0.937	0.983
MovingIndoor [88]	U	0.208	0.712	0.086	0.674	0.900	0.968
TrianFlow [87]	U	0.201	0.708	0.085	0.687	0.903	0.968
PackNet-SfM [25]	U	0.167	0.608	0.071	0.749	0.943	0.986
Johnston et al. [31]	U	0.164	0.602	0.070	0.760	0.943	0.987
P^2Net [79](3-frame)	U	0.159	0.599	0.068	0.772	0.942	0.984
HR-Depth [54]	U	0.161	0.596	0.069	0.762	0.944	0.983
Lite-Mono [82]	U	0.158	0.588	0.068	0.766	0.945	0.984
RA-Depth [27]	U	0.152	0.576	0.063	0.775	0.949	0.987
Ours full	U	**0.129**	**0.520**	**0.051**	**0.828**	**0.967**	**0.992**

Quantitative depth comparisons between our proposed Shading-SfM scheme and state-of-the-art learning-based depth estimation methods on NYUv2 Depth dataset. 'S' and 'U' indicate supervised and unsupervised training types, and \ indicates training collapsed.

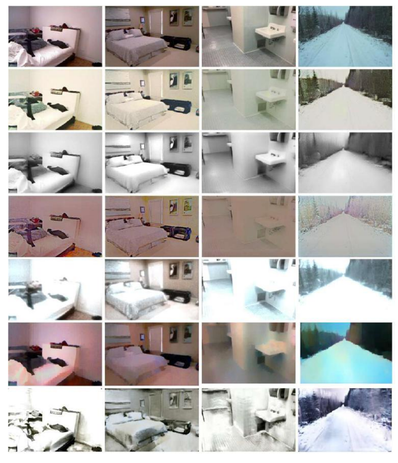


Visual comparisons of 3D reconstruction on ScanNet dataset. From left to right: input (1st column); result from our method (2nd column), Lite-mono (3rd column), HR-depth (4th column), and RA-Depth (5th column).

Methods	Type	Error			Acc		
		AbsRel	RMS	Log10	δ_1	δ_2	δ_3
DORN [21]	S	**0.140**	**0.290**	0.054	0.798	0.960	**0.991**
SharpNet [61]	S	0.167	0.358	**0.052**	**0.801**	**0.961**	**0.991**
Monodepth2 [24]	U	0.196	0.461	0.072	0.692	0.935	0.982
MovingIndoor [88]	U	0.238	0.570	-	0.631	-	-
P^2Net [79]	U	0.181	0.440	0.068	0.717	0.937	0.983
HR-Depth [54]	U	0.174	0.442	0.063	0.725	0.942	0.983
RA-Depth [27]	U	0.169	0.423	0.061	0.737	0.946	0.985
Lite-Mono [82]	U	0.172	0.434	0.062	0.729	0.944	0.984
Ours full	U	**0.151**	**0.352**	**0.058**	**0.798**	**0.956**	**0.990**

Quantitative depth comparison on ScanNet dataset.

With input (1st row), our depth estimation results (2nd row) on NYUv2 (the first three columns) and ScanNet (the last one column) datasets in comparison with RA-Depth (3rd row) and HR-Depth (4th row).



With input image (1st row), visual comparison of the estimated albedo and shading of our results (2-3rd row) on indoor NYUv2 (the first three columns) and outdoor Finnforest (the last one column) datasets, with Unsupervised Intrinsic Decomposition [37] (4-5th row) and Inverserendernet [78] (6-7th row).

w/ SfS-Net	w/ SSC	w/ SRC	w/ KPC	w/ LPC	AbsRel	δ_1
-	-	-	-	-	0.178	0.733
✓	-	-	-	-	0.161	0.758
✓	✓	-	-	-	0.155	0.770
✓	✓	✓	-	-	0.149	0.781
✓	✓	✓	✓	-	0.138	0.805
✓	✓	✓	✓	✓	**0.129**	**0.828**

Ablation study of depth estimation on different combinations of network components on NYUv2 dataset.