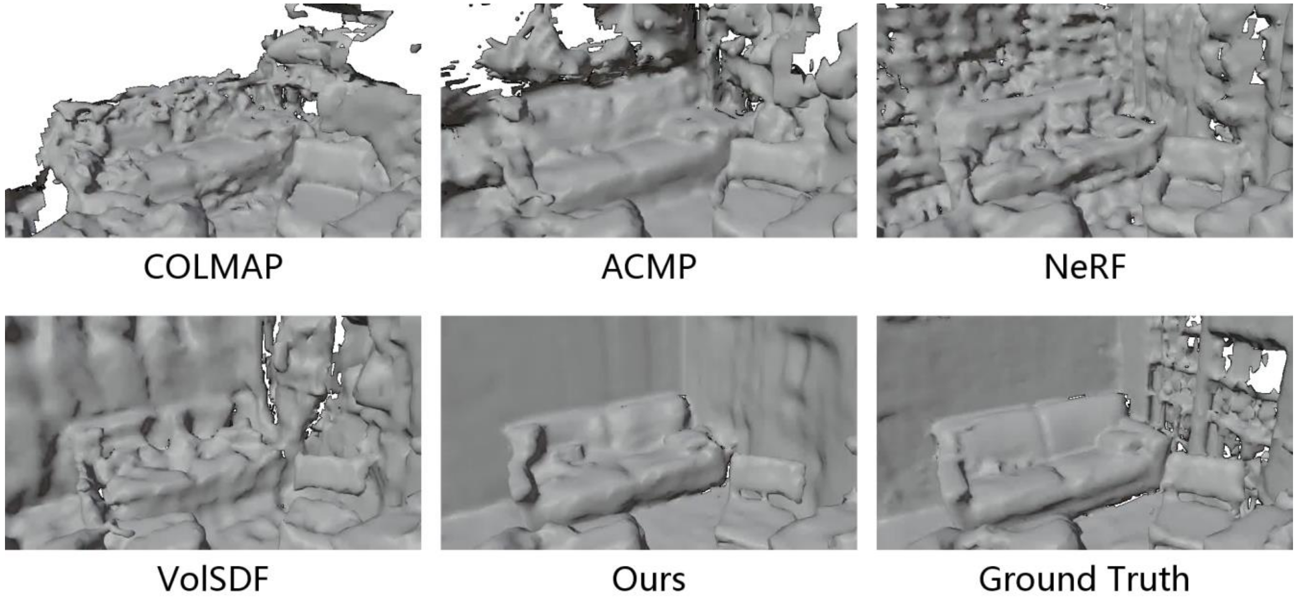
Neural 3D Scene Reconstruction with the Manhattan-world Assumption

박지호

Contents

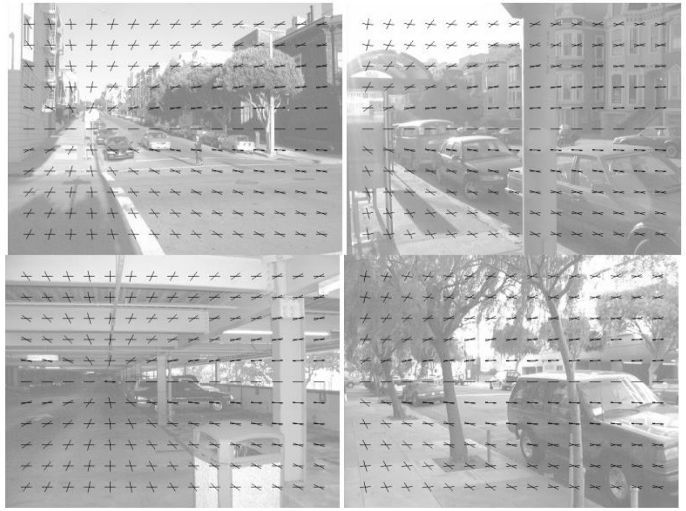


1. Task

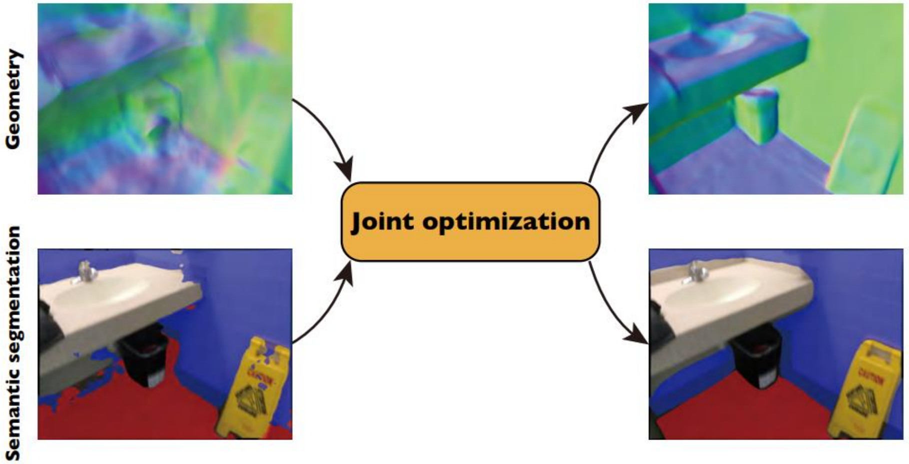
- Overcoming difficulty in Indoor Scene reconstruction
- Handling Planar Regions(low-textured)

2. Concept

- Appling Planar Constraint Regularization on wall and floor using 2D semantic segmentation
- Jointly Optimize Geometry and Semantics



Planar Constraint in Manhattan Assumption:
- Floor = $< 0, 0, 1 >$
- Wall = $< \pm 1, 0, 0 >$ or $< 0, \pm 1, 0 >$



Joint Optimization

3. Scene Representation

Prediction: Volume Rendering with Signed Distance Function
(considering normal vector)

MLP Prediction

$$(d(\mathbf{x}), \mathbf{z}(\mathbf{x})) = F_d(\mathbf{x})$$

$$\mathbf{c}(\mathbf{x}) = F_{\mathbf{c}}(\mathbf{x}, \mathbf{v}, \mathbf{n}(\mathbf{x}), \mathbf{z}(\mathbf{x}))$$

$d(x)$: *signed distance*
v: *view direction*
$n(x)$: *normal vector* $= \nabla d(x)$
$z(x)$: *geometry feature*

Volume Rendering

$$\sigma(\mathbf{x}) = \begin{cases} \frac{1}{\beta}\left(1 - \frac{1}{2}\exp\left(\frac{d(\mathbf{x})}{\beta}\right)\right) & \text{if } d(\mathbf{x}) < 0, \\ \frac{1}{2\beta}\exp\left(-\frac{d(\mathbf{x})}{\beta}\right) & \text{if } d(\mathbf{x}) \geq 0, \end{cases}$$

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{K} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i$$

β: *learnable parameter*

3. Scene Representation

Optimization: Losses for scene reconstruction

Photometric Loss:

$$\mathcal{L}_{\text{img}} = \sum_{\mathbf{r} \in \mathcal{R}} \left\| \hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|$$

Eikonal Loss:
(Geometric Regularization)

$$\mathcal{L}_E = \sum_{\mathbf{y} \in \mathcal{Y}} (\|\nabla_{\mathbf{y}} d(\mathbf{y})\|_2 - 1)^2$$

Depth Loss:
(Assists Learning)

$$\mathcal{L}_d = \sum_{\mathbf{r} \in \mathcal{D}} \left| \hat{D}(\mathbf{r}) - D(\mathbf{r}) \right|$$

4. Handling Planar Region

4-1. Planar Constraint Regularization

x_r: surface intersection point(by 2D semantic segmentation)
$n_f = < 0, 0, 1 >$
$n_w = < 1, 0, 0 >$

Floor Regularization Loss: $$\mathcal{L}_f(\mathbf{r}) = |1 - \mathbf{n}(\mathbf{x_r}) \cdot \mathbf{n}_f|$$

Wall Regularization Loss: $$\mathcal{L}_w(\mathbf{r}) = \min_{i \in \{-1,0,1\}} |i - \mathbf{n}(\mathbf{x_r}) \cdot \mathbf{n}_w|$$

Geometric Loss(sum): $$\mathcal{L}_{\text{geo}} = \sum_{r \in \mathcal{F}} \mathcal{L}_f(\mathbf{r}) + \sum_{r \in \mathcal{W}} \mathcal{L}_w(\mathbf{r})$$

4. Handling Planar Region

4-2. Joint Optimization

Optimize 3D Semantic Label, in case of 2D semantic segmentation is wrong,

Label Prediction with Volume Rendering

$$\mathbf{s}(\mathbf{x}) = F_{\mathbf{s}}(\mathbf{x})$$

$$\hat{\mathbf{S}}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{s}_i$$

$s(x)$: semantic logit of x

Joint Loss

$$\mathcal{L}_{\text{joint}} = \sum_{\mathbf{r} \in \mathcal{F}} \hat{p}_f(\mathbf{r}) \mathcal{L}_f(\mathbf{r}) + \sum_{\mathbf{r} \in \mathcal{W}} \hat{p}_w(\mathbf{r}) \mathcal{L}_w(\mathbf{r})$$

Cross Entropy Loss:
(avoid $\hat{p}_f, \hat{p}_w$ to be vanished)

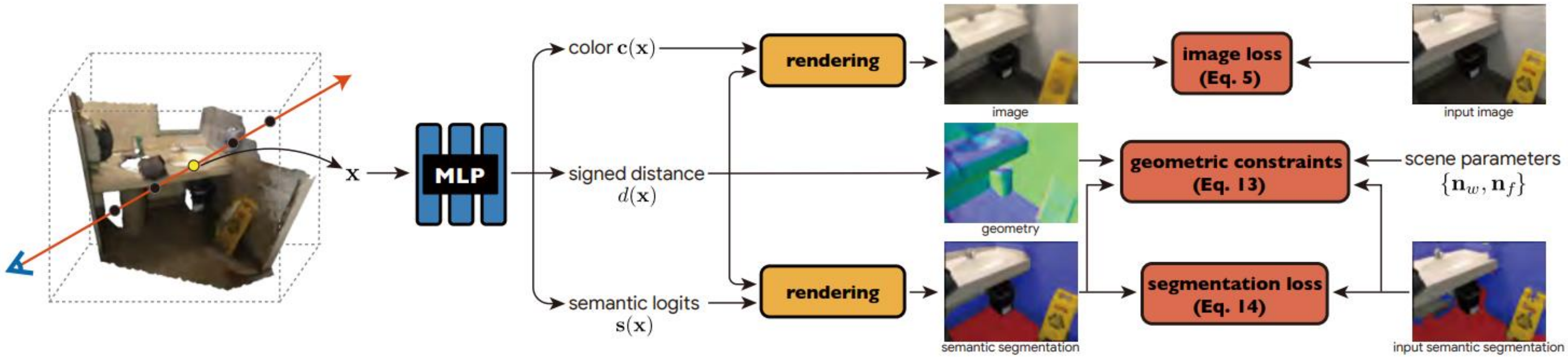
$$\mathcal{L}_{\mathbf{s}} = -\sum_{\mathbf{r} \in \mathcal{R}} \sum_{k \in \{f, w, b\}} p_k(\mathbf{r}) \log \hat{p}_k(\mathbf{r})$$

$\hat{p}_f, \hat{p}_w, \hat{p}_b$: softmax result after label prediction

5. Final Network

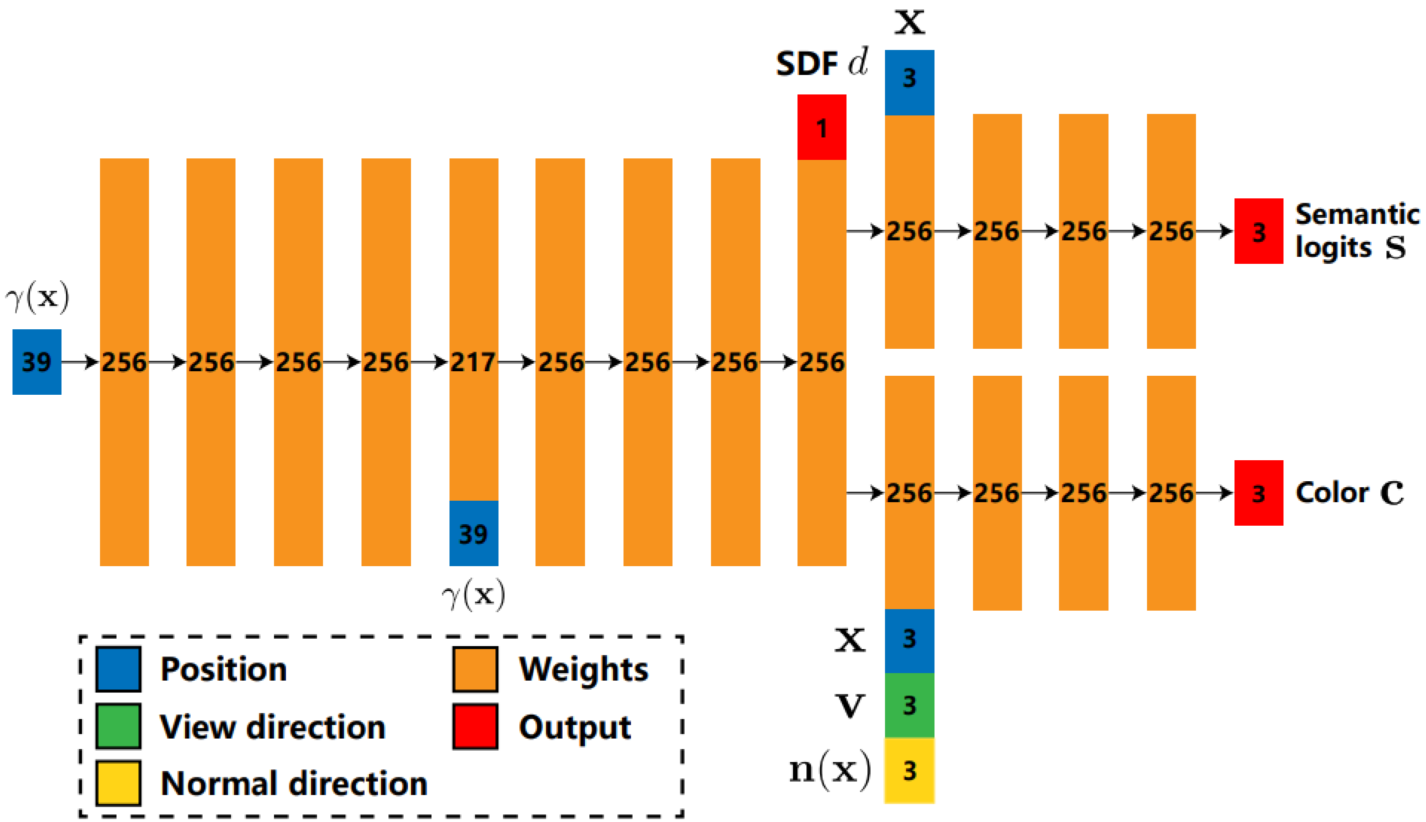
Jointly Optimizing

1. Photometric information
2. Geometric Constraints of Planar(using semantic segmentation)
3. 3D Semantic information



5. Final Network

Network Architecture:



6. Results

Ablation Study: Comparing Depth loss, Geometric loss, Joint loss

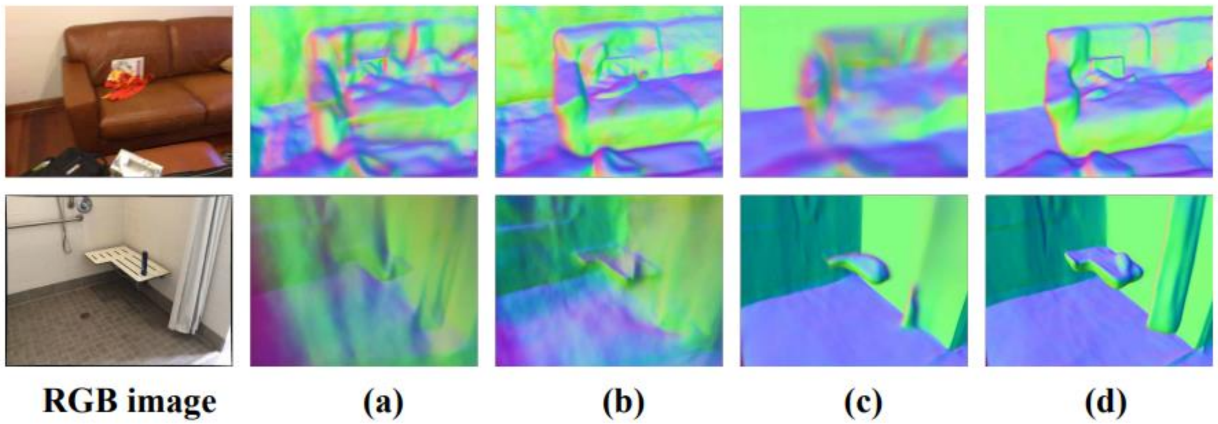
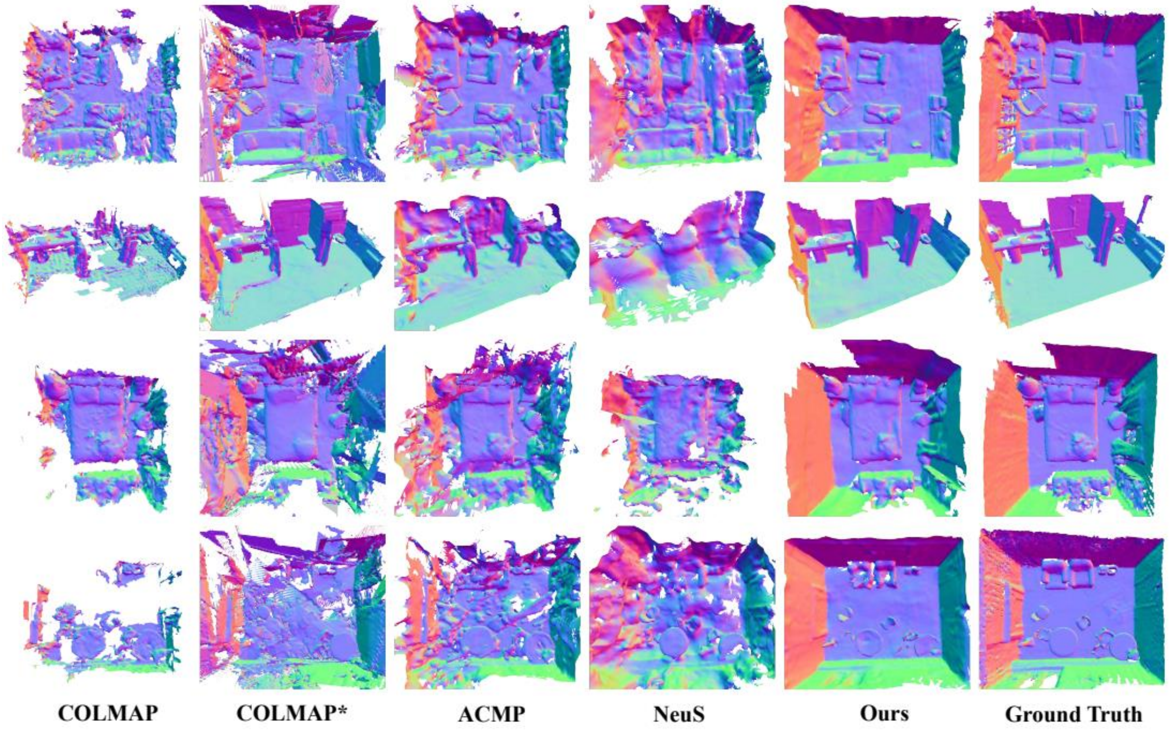


Figure 3. **Qualitative ablations.** (a) Training with only images. (b) Adding $\mathcal{L}_d$. (c) Adding $\mathcal{L}_{\text{geo}}$. (d) Replacing $\mathcal{L}_{\text{geo}}$ with $\mathcal{L}_{\text{joint}}$.

6. Results

Comparing Scene Reconstruction Performance



Method	ScanNet				
	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**
COLMAP	**0.047**	0.235	**0.711**	0.441	0.537
COLMAP*	0.396	0.081	0.271	**0.595**	0.368
ACMP	0.118	0.081	0.531	0.581	0.555
NeRF	0.735	0.177	0.131	0.290	0.176
UNISURF	0.554	0.164	0.212	0.362	0.267
NeuS	0.179	0.208	0.313	0.275	0.291
VolSDF	0.414	0.120	0.321	0.394	0.346
Ours	0.072	**0.068**	0.621	0.586	**0.602**
Method	**7-Scenes**				
	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**
COLMAP	**0.069**	0.417	**0.536**	0.202	0.289
COLMAP*	0.670	0.215	0.116	0.215	0.149
ACMP	0.293	0.194	0.350	0.269	0.299
NeRF	0.573	0.321	0.159	0.085	0.083
UNISURF	0.407	0.136	0.195	0.301	0.231
NeuS	0.151	0.247	0.313	0.229	0.262
VolSDF	0.285	0.140	0.220	0.285	0.246
Ours	0.112	**0.133**	0.351	**0.326**	**0.336**

7. Conclusion

Proposed Methods

- Utilizing semantic information in planar regions(floor, wall) to guide geometry reconstruction
- Learning 3D semantic from 2D segmentation
- Joint Loss

Limitation

- Manhattan world assumption is not general enough.

Conclusion

- Joint Optimization improves the robustness against inaccurate 2D segmentation.