How do NeRF and CLIP advance 3D Scene Reconstruction and Understanding?

Songyou Peng
ETH Zurich and Max Planck Institute for Intelligent Systems

Chinese University of Hong Kong (CUHK), Shenzhen
March 07, 2023
Who Am I?

• 4th Year PhD Student
  • Marc Pollefeys
  • Andreas Geiger

• Internships during PhD
  • 2021: Michael Zollhoefer
  • 2022: Tom Funkhouser

• Open to chat!

pengsongyou.github.io
My PhD Topics: Neural Scene Representations
for 3D reconstruction, novel view synthesis, and SLAM

Convolutional Occupancy Networks
ECCV 2020 (Spotlight)

KiloNeRF
ICCV 2021

UNISURF
ICCV 2021 (Oral)

Shape As Points
NeurIPS 2021 (Oral)

MonoSDF
NeurIPS 2022

NICE-SLAM
CVPR 2022
How do NeRF and CLIP advance 3D Scene Reconstruction and Understanding?
How does NeRF advance 3D Scene Reconstruction?

How does CLIP advance 3D Scene Understanding?
How does NeRF advance 3D Scene Reconstruction?

How does CLIP advance 3D Scene Understanding?
NeRF is awesome!

Some problems still exist...

😢 Poor underlying geometry
😢 Camera poses needed

😊 MonoSDF
😊 NICE-SLAM

Mildenhall*, Srinivasan*, Tancik* et al: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. ECCV 2020
MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, Andreas Geiger
Neural Implicit Surfaces with Volume Rendering

Fails with sparse input views
Poor results in large-scale indoor scenes

Neural Implicit Surfaces with Volume Rendering

- Manage to reconstruct with sparse views
- Nice 3D reconstruction in large-scale indoor scenes

Shape-Appearance Ambiguity

There exists an infinite number of photo-consistent explanations for input images!

Zhang, Riegler, Snavely, Koltun: NeRF++: Analyzing and Improving Neural Radiance Fields. ArXiv, 2020
Shape-Appearance Ambiguity

There exists an infinite number of photo-consistent explanations for input images!

Exploit monocular geometric priors

Zhang, Riegler, Snavely, Koltun: NeRF++: Analyzing and Improving Neural Radiance Fields. ArXiv, 2020
Depth Map Prediction from a Single Image

Omnidata

OmniData

MonoSDF
MonoSDF
MonoSDF
MonoSDF

Neural Implicit Scene Representation
MonoSDF

Neural Implicit Scene Representation

MLP

Dense SDF Grid

Single-res Feature Grid

Multi-res Feature Grids

Input Views

Ray Distance

σ
MonoSDF

Neural Implicit Scene Representation

Volume Rendering

Input Views

Ray Distance

\( \sigma \)
MonoSDF
MonoSDF
MonoSDF

Neural Implicit Scene Representation

Volume Rendering

Input Views

C (r)  \hat{D}(r)  \hat{N}(r)

Ray Distance
MonoSDF

MLP

Dense SDF Grid

Single-res Feature Grid

Multi-res Feature Grids

Neural Implicit Scene Representation

Volume Rendering

\( \hat{C}(r) \)

\( \hat{D}(r) \)

\( \hat{N}(r) \)

Input Views

Pretrained Omnidata Model

\( \sigma \)

Ray Distance

Monocular Geometric Cues

\( f_0 \rightarrow \hat{s} \)

\( x \rightarrow \hat{s} \)
## Ablation Study

<table>
<thead>
<tr>
<th></th>
<th>Normal C.↑</th>
<th>Chamfer-$L_1$ ↓</th>
<th>F-score ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MLP</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Cues</td>
<td>86.48</td>
<td>6.75</td>
<td>66.88</td>
</tr>
<tr>
<td>Only Depth</td>
<td>90.56</td>
<td>4.26</td>
<td>76.42</td>
</tr>
<tr>
<td>Only Normal</td>
<td>91.35</td>
<td>3.19</td>
<td>85.84</td>
</tr>
<tr>
<td>Both Cues</td>
<td><strong>92.11</strong></td>
<td><strong>2.94</strong></td>
<td><strong>86.18</strong></td>
</tr>
<tr>
<td><strong>Multi-Res. Grids</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Cues</td>
<td>87.95</td>
<td>5.03</td>
<td>78.38</td>
</tr>
<tr>
<td>Only Depth</td>
<td>90.87</td>
<td>3.75</td>
<td>80.32</td>
</tr>
<tr>
<td>Only Normal</td>
<td>89.90</td>
<td>3.61</td>
<td>81.28</td>
</tr>
<tr>
<td>Both Cues</td>
<td><strong>90.93</strong></td>
<td><strong>3.23</strong></td>
<td><strong>85.91</strong></td>
</tr>
</tbody>
</table>

![Graph showing F-score over iterations](image)

- Monocular cues improve reconstruction results significantly
- Combining **depth** & **normal** leads to best performance
- Monocular cues can improve convergence speed
Baseline Comparisons on ScanNet

Ours
Multi-Res. Feature Grids with High-Res. Cues
Baseline Comparisons on DTU (3-views)
Take-home Message

Monocular cues improve reconstruction results and speed up optimization

Inspire Haiwen & Dan’s ICLR 2023 paper GOOD 😊

Limitation: Still require camera poses given :( 

https://niujinshuchong.github.io/monosdf/
NICE-SLAM
Neural Implicit Scalable Encoding for SLAM
CVPR 2022

Zihan Zhu*  Songyou Peng*  Viktor Larsson  Weiwei Xu  Hujun Bao  Zhaopeng Cui  Martin R. Oswald  Marc Pollefeys

* Equal Contributions
RGB-D Sequences

40x Speed
iMAP
[Sucar et al., ICCV'21]

First neural implicit-based **online** SLAM system
iMAP
[Sucar et al., ICCV'21]

Fail when scaling up to larger scenes
Global update $\rightarrow$ Catastrophic forgetting
Slow convergence

A single MLP

Predicted Poses
GT Poses
NICE-SLAM

Feature grids + tiny MLPs

Applicable to **large-scale scenes**
Local update $\rightarrow$ **No forgetting problem**
**Fast** convergence

- Predicted Poses
- GT Poses
Pipeline

Mapping

Tracking

Input Depth

Input RGB

Coarse-level Occupancy

Hierarchical Feature Grid

Tri-linear Interpolation

Ray -> Point Sampler

Camera Pose
Results
iMAP* (our re-implementation of iMAP)

4x Speed
iMAP*  
(our re-implementation of iMAP)

10x Speed
Take-home Message

• A NICE online implicit SLAM system for indoor scenes
• Hierarchical feature grids + a tiny MLP seems to be a trend!
  • Instant-NGP [SIGGRAPH’22 Best Paper]

Limitations

• Requires depths as input
• Only bounded scenes
• Still not real-time
NICER-SLAM: Neural Implicit Scene Encoding for RGB SLAM

Zihan Zhu\textsuperscript{1}\textsuperscript{*} Songyou Peng\textsuperscript{1,2}\textsuperscript{*} Viktor Larsson\textsuperscript{3} Zhaopeng Cui\textsuperscript{4}
Martin R. Oswald\textsuperscript{1,5} Andreas Geiger\textsuperscript{6} Marc Pollefeys\textsuperscript{1,7}

\textsuperscript{1}ETH Zürich \hspace{1cm} \textsuperscript{2}MPI for Intelligent Systems, Tübingen \\
\textsuperscript{3}Lund University \hspace{1cm} \textsuperscript{4}State Key Lab of CAD&CG, Zhejiang University \\
\textsuperscript{5}University of Amsterdam \hspace{1cm} \textsuperscript{6}University of Tübingen, Tübingen AI Center \\
\textsuperscript{7}Microsoft

---

https://arxiv.org/abs/2302.03594
How does NeRF advance 3D Scene Reconstruction?

How does CLIP advance 3D Scene Understanding?
How does NeRF advance 3D Scene Reconstruction?

How does CLIP advance 3D Scene Understanding?
Input 3D Geometry

Traditional Semantic Segmentation

Only train and test on a few common classes
Input 3D Geometry

3D Scene Understanding Tasks w/o Labels

- Affordance prediction
- Material identification
- Physical property estimation
- Rare object retrieval
- Activity site prediction
- Fine-grained semantic segmentation
- Many more…
Key Idea: Co-embed 3D features with CLIP features
Key Idea: Co-embed 3D features with CLIP features

3D Geometry

CLIP Text Features
(visualize with T-SNE)

RGB Images

Note: bold word embeddings are approximate
How to Learn Such Text-Image-3D Co-Embeddings?
Step 1: Multi-view Feature Fusion

3D Geometry

Per-pixel Features
(visualize with PCA)

RGB Images

$\mathcal{E}^{2D}$

OpenSeg [1]
LSeg [2]

Step 2: 3D Distillation

\[ \mathcal{L} = 1 - \cos(f^{2D} - f^{3D}) \]
Step 3: 2D-3D Ensemble

3D Geometry

2D-3D Ensemble Features

Choose the feature with the highest max score among all prompts

\[ s_{n}^{2D} = \cos(f^{2D}, t_n) \]
\[ s_{n}^{3D} = \cos(f^{3D}, t_n) \]
Open-Vocabulary, Zero-shot

3D Semantic Segmentation
Our Zero-shot 3D Segmentation
(20 classes)
Our Zero-shot 3D Segmentation
(160 classes)
Comparison

Matterport3D Top 160 Classes (ranked by number of instances in training set)

Most Common Classes

<table>
<thead>
<tr>
<th>Range</th>
<th>Fully supervised</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 20</td>
<td>64.5</td>
<td>42.9</td>
</tr>
<tr>
<td>21 - 40</td>
<td>33.2</td>
<td>33.3</td>
</tr>
<tr>
<td>41 - 60</td>
<td>15.6</td>
<td>21.9</td>
</tr>
<tr>
<td>61 - 80</td>
<td>13.2</td>
<td>24.2</td>
</tr>
<tr>
<td>81 - 100</td>
<td>4.5</td>
<td>19.4</td>
</tr>
<tr>
<td>101 - 120</td>
<td>7.8</td>
<td>16.0</td>
</tr>
<tr>
<td>121 - 140</td>
<td>8.6</td>
<td>19.7</td>
</tr>
<tr>
<td>141 - 160</td>
<td>1.7</td>
<td>16.3</td>
</tr>
</tbody>
</table>

Rarest Classes
Interactive Demo

Open-vocabulary 3D Scene Exploration
Take-home Message

• We enable a wide range of applications by open-vocabulary queries

• This can hopefully influence how people train 3D scene understanding systems in the future

• The project can be improved in many aspects
  • Better feature fusion strategy than simple averaging
  • Combine CLIP features with NeRF/SLAM
    • concept-fusion.github.io
How does NeRF advance 3D Scene Reconstruction?

[NeurIPS'22] MonoSDF
github.com/autonomousvision/monosdf

[CVPR’22] NICE-SLAM
pengsongyou.github.io/nice-slam

How does CLIP advance 3D Scene Understanding?

[CVPR’23] OpenScene
pengsongyou.github.io/openscene