Learning Neural Scene Representations for 3D Reconstruction and Understanding

Songyou Peng
ETH Zurich and Max Planck Institute for Intelligent Systems

Shanghai AI Lab
June 15, 2023
Who Am I?

• Final-year PhD Student
  • Marc Pollefeys
  • Andreas Geiger

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

• Before PhD, worked in Singapore, and interned at INRIA and TUM
Motivation

Input Images

3D Reconstruction
Motivation

3D Reconstruction

3D Scene Understanding
My PhD Topics: Neural Scene Representations for 3D reconstruction and 3D scene understanding

- Convolutional Occupancy Nets
  ECCV 2020 (Spotlight)

- Shape As Points
  NeurIPS 2021 (Oral)

- KiloNeRF
  ICCV 2021

- UNISURF
  ICCV 2021 (Oral)

- MonoSDF
  NeurIPS 2022

- NICE-SLAM
  CVPR 2022

- OpenScene
  CVPR 2023

- NICER-SLAM
  arXiv 2023
My PhD Topics: Neural Scene Representations for 3D reconstruction and 3D scene understanding

MonoSDF
NeurIPS 2022

NICE-SLAM
CVPR 2022

OpenScene
CVPR 2023
NeRF is awesome!

Some existing problems...

😊 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

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
MonoSDF

MLP

x → \(f_0\) → \(\hat{s}\)

Dense SDF Grid

x → \(\hat{s}\)

Single-res Feature Grid

x → \(f_0\) → \(\hat{s}\)

Multi-res Feature Grids

x → \(\ldots\) → \(f_0\) → \(\hat{s}\)

Neural Implicit Scene Representation

\(\sigma\)

Ray Distance

Input Views
MonoSDF

Neural Implicit Scene Representation

Volume Rendering

Input Views

Ray Distance
MonoSDF

Neural Implicit Scene Representation

MLP

Dense SDF Grid

Single-res Feature Grid

Multi-res Feature Grids

Volume Rendering

Input Views

C

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

\( L_{rgb} \)

\( \sigma \)

Ray Distance
MonoSDF
MonoSDF
MonoSDF

Monocular Geometric Cues
# 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>

- Monocular cues improve reconstruction results significantly
- Combining **depth & normal** leads to best performance
- Monocular cues can improve **convergence speed**
Baseline Comparisons on ScanNet
Multi-Res. Feature Grids with High-Res. Cues
Baseline Comparisons on DTU (3-views)
Monocular cues improve reconstruction results and speed up optimization

Inspire applications in other fields [GOOD, ICLR 2023]

Limitation: Still require camera poses given :(
RGB-D Sequences

40x Speed
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
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
**NICE-SLAM**

- Feature grids + tiny MLPs

- Applicable to **large-scale scenes**
- Local update → **No forgetting problem**
- **Fast** convergence
Results
iMAP*  
(our re-implementation of iMAP)

4x Speed

NICE-SLAM
iMAP* (our re-implementation of iMAP)  NICE-SLAM

10x Speed
Note: Runtime evaluation setting from iMAP paper, not the best-performing setting.
Take-home Message

• A NICE NeRF-based 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\(^1\)*  
Songyou Peng\(^1,2\)*  
Viktor Larsson\(^3\)  
Zhaopeng Cui\(^4\)  
Martin R. Oswald\(^1,5\)  
Andreas Geiger\(^6\)  
Marc Pollefeys\(^1,7\)

\(^1\)ETH Zürich  
\(^2\)MPI for Intelligent Systems, Tübingen  
\(^3\)Lund University  
\(^4\)State Key Lab of CAD&CG, Zhejiang University  
\(^5\)University of Amsterdam  
\(^6\)University of Tübingen, Tübingen AI Center  
\(^7\)Microsoft

NICE-SLAM  
Vox-Fusion  
COLMAP  
DROID-SLAM  
NICER-SLAM  
GT

\(^{RGB-D input}\)  
\(^{RGB input}\)

https://arxiv.org/abs/2302.03594
Input 3D Geometry
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...
3D Scene Understanding with Open Vocabularies

Songyou Peng  Kyle Genova  Chiyu “Max” Jiang  Andrea Tagliasacchi  Marc Pollefeys  Tom Funkhouser
Key Idea: Co-embed 3D features with CLIP features

CLIP: Contrastive Language-Image Pre-Training

Radford et al.: Learning Transferable Visual Models From Natural Language Supervision. ICML 2021
Key Idea: Co-embed 3D features with CLIP features

3D Geometry

CLIP Text Features
(visualize with T-SNE)

RGB Images
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**

3D Geometry

**Per-pixel Features**

**RGB Images**


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

$$s_{2D}^n = \cos(f_{2D}^n, t_n)$$

$$s_{3D}^n = \cos(f_{3D}^n, t_n)$$

(visualize with PCA)
Open-Vocabulary, Zero-shot

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

Most Common Classes

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

- Fully supervised
- Ours

mAcc (%)
Comparison

Rarest Classes

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

The bar chart shows the mIoU (%) for two datasets: ScanNet and Matterport3D, under different methods:

- **2D Fusion**
- **3D Distillation**
- **2D-3D Ensemble**

For ScanNet:
- 2D Fusion: 41.4%
- 3D Distillation: 46.0%
- 2D-3D Ensemble: 47.5%

For Matterport3D:
- 2D Fusion: 41.3%
- 3D Distillation: 32.4%
- 2D-3D Ensemble: 42.6%
Image-based 3D Scene Query
Our Segmentation
Input 3D Geometry
Image Queries

Given 3D Geometry
Image Queries
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

• Our real-time demo already shows the **possibility to directly apply to AR/VR**
My PhD Topics: Neural Scene Representations for 3D reconstruction and 3D scene understanding

- **MonoSDF**
  NeurIPS 2022

- **NICE-SLAM**
  CVPR 2022

- **OpenScene**
  CVPR 2023
Learning Neural Scene Representations for 3D Reconstruction and Understanding
Songyou Peng

Ours

MonoSDF
NeurIPS 2022
niujinshuchong.github.io/monosdf/

NICE-SLAM
CVPR 2022
pengsongyou.github.io/nice-slam

OpenScene
CVPR 2023
pengsongyou.github.io/openscene

Thank you!