Implicit Neural Representations: From Objects to 3D Scenes

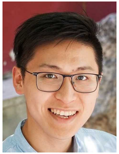
Andreas Geiger

Autonomous Vision Group
University of Tübingen / MPI for Intelligent Systems Tübingen

June 19, 2020

University of Tübingen
MPI for Intelligent Systems
Autonomous Vision Group

Collaborators



Songyou Peng

Michael Oechsle

Carolin Schmitt

Michael Niemeyer

Lars Mescheder

Simon Donne

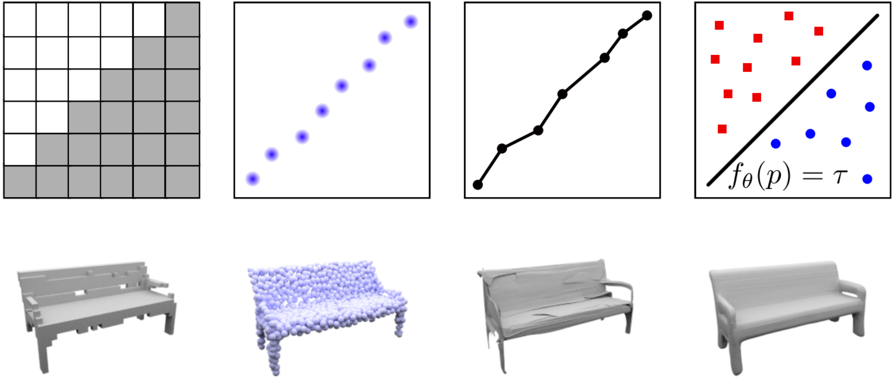
Gernot Riegler

Vladlen Koltun

Marc Pollefeys

Andreas Geiger

3D Representations

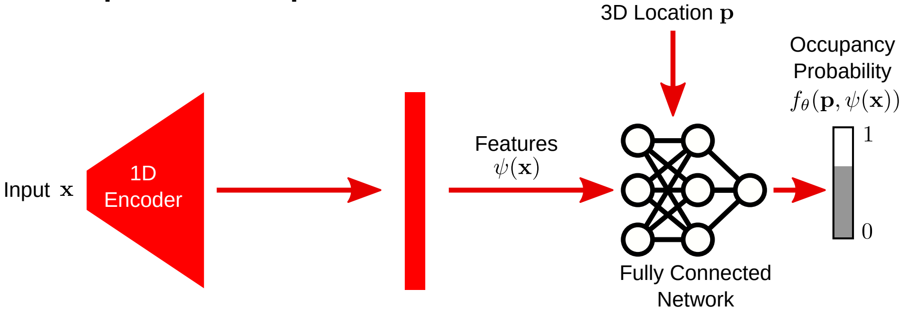


- Traditional Explicit Representations $\Rightarrow$ **Discrete**
- Implicit Neural Representation $\Rightarrow$ **Continuous**

Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019.

Limitations

Structure of implicit neural representations:

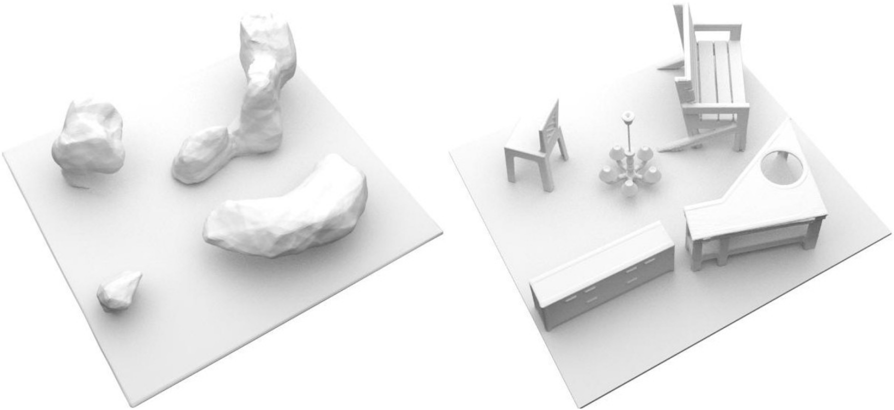


- Global latent code $\Rightarrow$ no local information, overly smooth geometry
- Fully connected architecture $\Rightarrow$ does not exploit translation equivariance

Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019.

Limitations

Implicit models work well for simple objects but **poorly on complex scenes:**

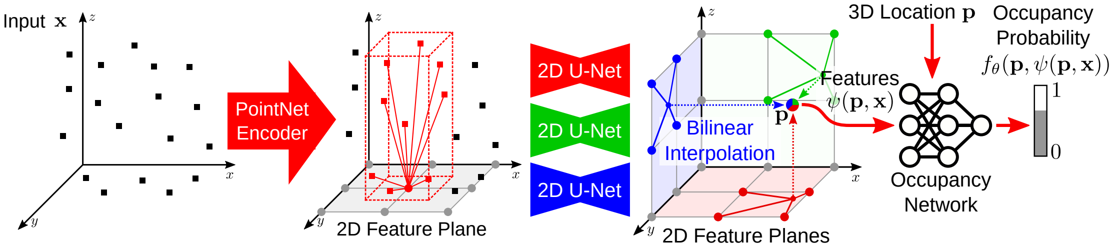


Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019.

How to reconstruct large-scale 3D scenes with implicit neural representations?

Convolutional Occupancy Networks

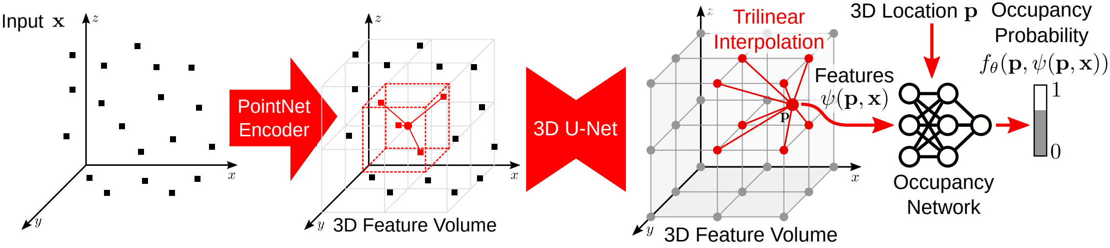
Convolutional Occupancy Networks



- **2D Plane Encoder:** Local PointNet processes input, project onto canonical plane
- **2D Plane Decoder:** Processed by U-Net, query features via bilinear interpolation
- **Occupancy Readout:** Shallow occupancy network $f_\theta(\cdot)$

Peng, Niemeyer, Mescheder, Pollefeys and Geiger: Convolutional Occupancy Networks. arXiv, 2020.

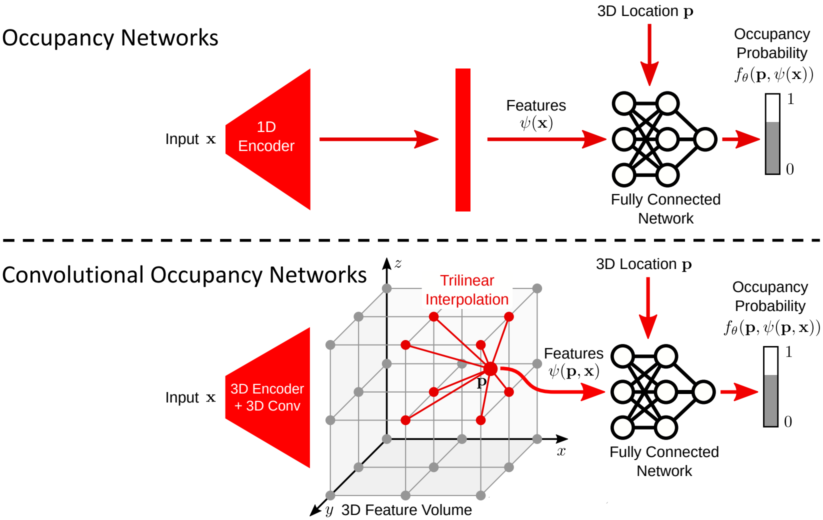
Convolutional Occupancy Networks



- **3D Volume Encoder:** Local PointNet processes input, volumetric feature encoding
- **3D Volume Decoder:** Processed by 3D U-Net, query features via trilinear interp.
- **Occupancy Readout:** Shallow occupancy network $f_\theta(\cdot)$

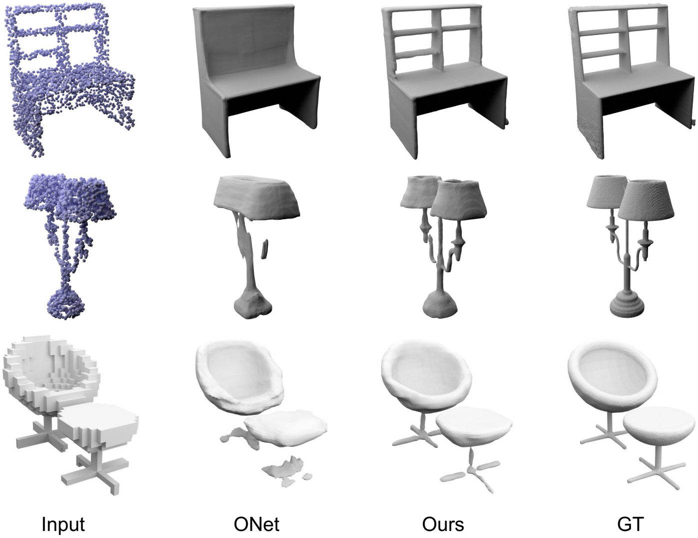
Peng, Niemeyer, Mescheder, Pollefeys and Geiger: Convolutional Occupancy Networks. arXiv, 2020.

Comparison


Occupancy Networks
3D Location $\mathbf{p}$
Occupancy
Probability
$f_\theta(\mathbf{p}, \psi(\mathbf{x}))$
Input $\mathbf{x}$
1D
Encoder
Features
$\psi(\mathbf{x})$
1
0
Fully Connected
Network
Convolutional Occupancy Networks
z
Trilinear
Interpolation
3D Location $\mathbf{p}$
Occupancy
Probability
$f_\theta(\mathbf{p}, \psi(\mathbf{p}, \mathbf{x}))$
Features
$\psi(\mathbf{p}, \mathbf{x})$
$\mathbf{p}$
Input $\mathbf{x}$
3D Encoder
+ 3D Conv
1
0
x
Fully Connected
Network
y 3D Feature Volume

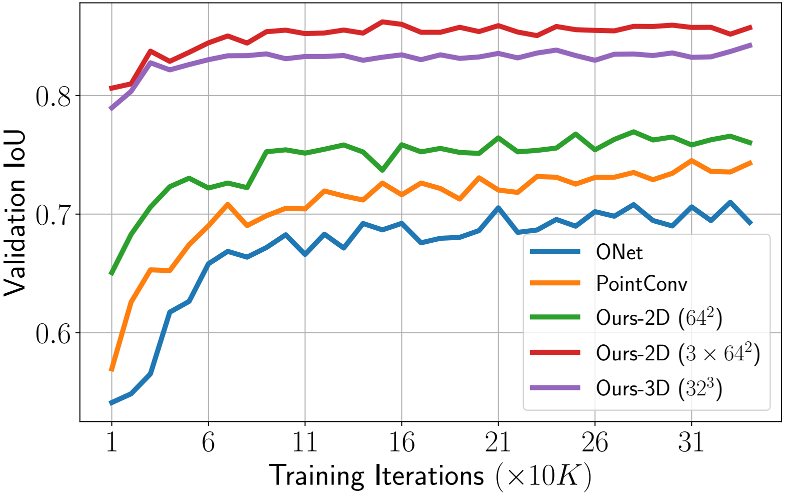
Results

Object-Level Reconstruction

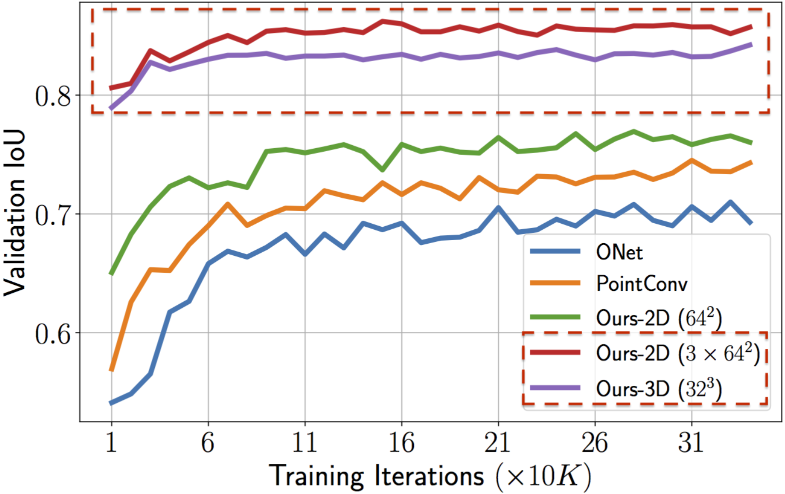


Peng, Niemeyer, Mescheder, Pollefeys and Geiger: Convolutional Occupancy Networks. arXiv, 2020.

Training Speed

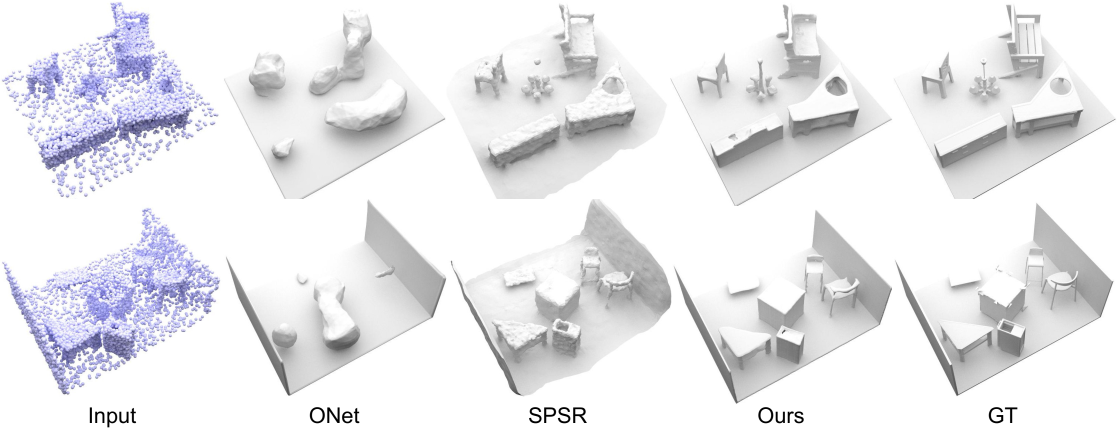


Training Speed



Peng, Niemeyer, Mescheder, Pollefeys and Geiger: Convolutional Occupancy Networks. arXiv, 2020.

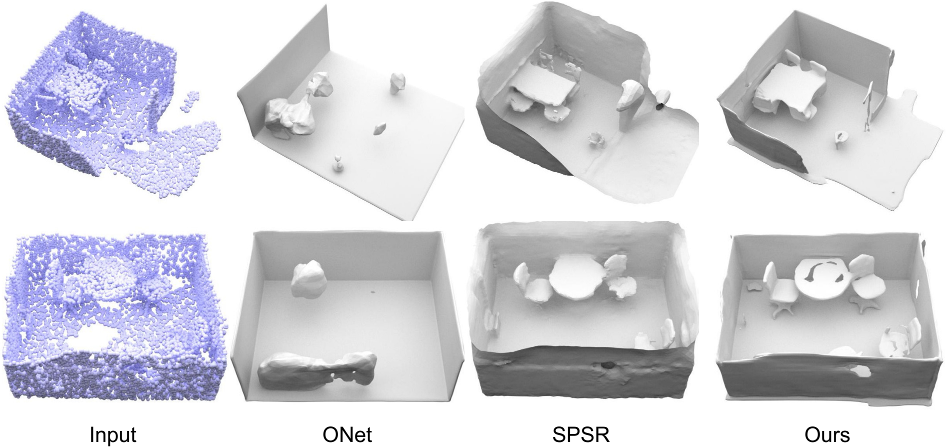
Scene-Level Reconstruction



- Trained and evaluated on synthetic rooms

Peng, Niemeyer, Mescheder, Pollefeys and Geiger: Convolutional Occupancy Networks. arXiv, 2020.

Scene-Level Reconstruction

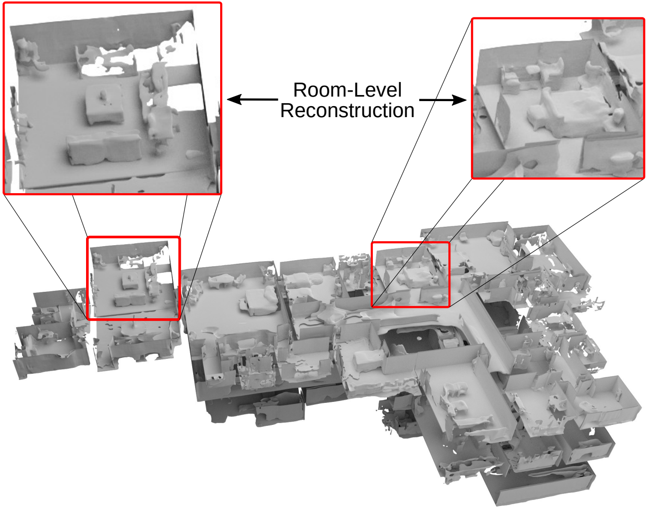


- Trained on synthetic rooms, evaluated on **ScanNet**

Large-Scale Reconstruction

Results on Matterport3D

- Fully convolutional model
- Trained on synthetic crops
- Sliding window evaluation
- Scales to any scene size



Peng, Niemeyer, Mescheder, Pollefeys and Geiger: Convolutional Occupancy Networks. arXiv, 2020.

Key Insights

Key Insights:

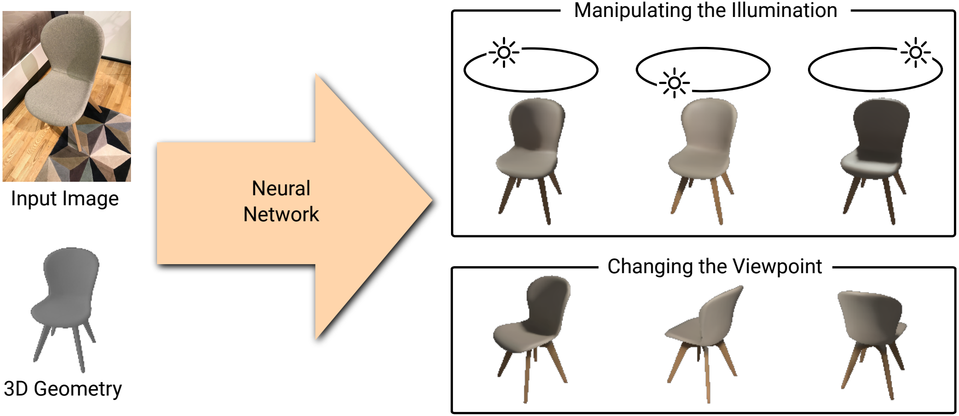
- Convolutional models allow for scaling implicit models to larger scenes
- Convolutional models train faster than fully implicit models
- Convolutional models allow for incorporating local feature information
- For objects, the 3-plane model has the best accuracy/memory trade-off
- For scenes, the volumetric representation performs best
- Models transfer from synthetic to real scenes

Peng, Niemeyer, Mescheder, Pollefeys and Geiger: Convolutional Occupancy Networks. arXiv, 2020.

How to capturing the visual appearance of objects?

Conditional Surface Light Fields

Problem Definition

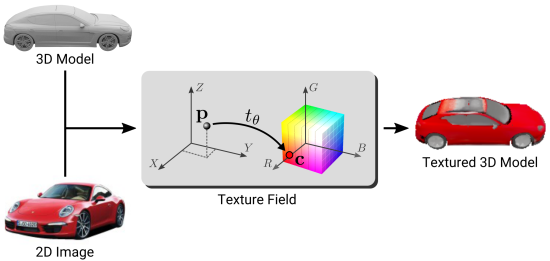


Oechsle, Niemeyer, Mescheder, Strauss and Geiger: Learning Implicit Surface Light Fields. arXiv, 2020.

Existing Representation

Texture Fields

- 3D consistent
- Generalize across objects
- View-point independent
- Do not model lighting



[Oechsle et al., ICCV 2019]

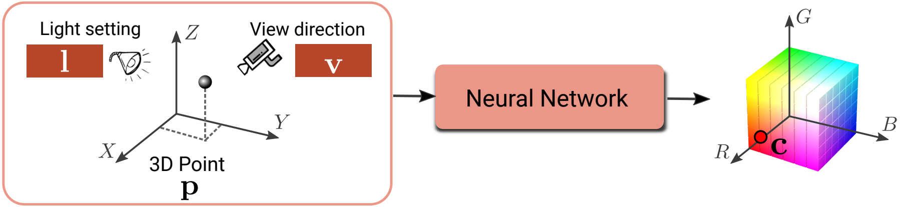
Conditional Surface Light Field

Rendering equation:

$$L(\mathbf{p}, \mathbf{v}, \mathbf{l}, \mathbf{n}) = \int_{\Omega} \text{svBRDF}(\mathbf{p}, \mathbf{r}, \mathbf{v}) \cdot \mathbf{l}(\mathbf{r}) \cdot (\mathbf{n}^T \mathbf{r}) \, d\mathbf{r}$$

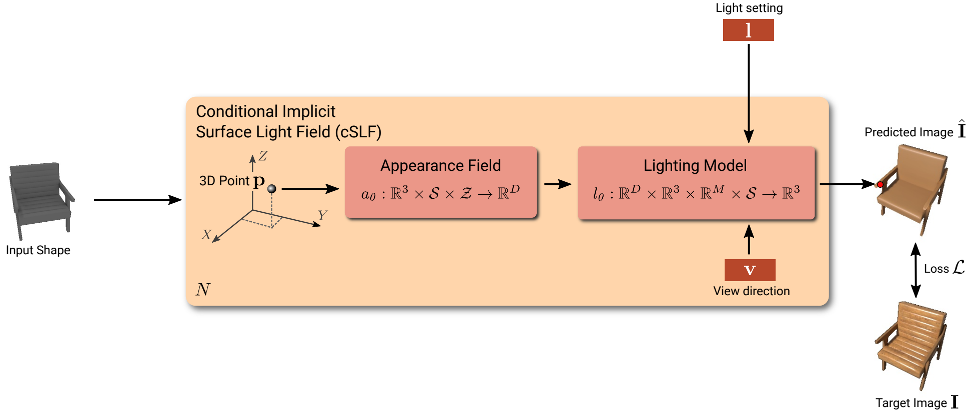
Conditional surface light field:

$$L_{\text{cSLF}}(\mathbf{p}, \mathbf{v}, \mathbf{l}) : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^M \rightarrow \mathbb{R}^3$$



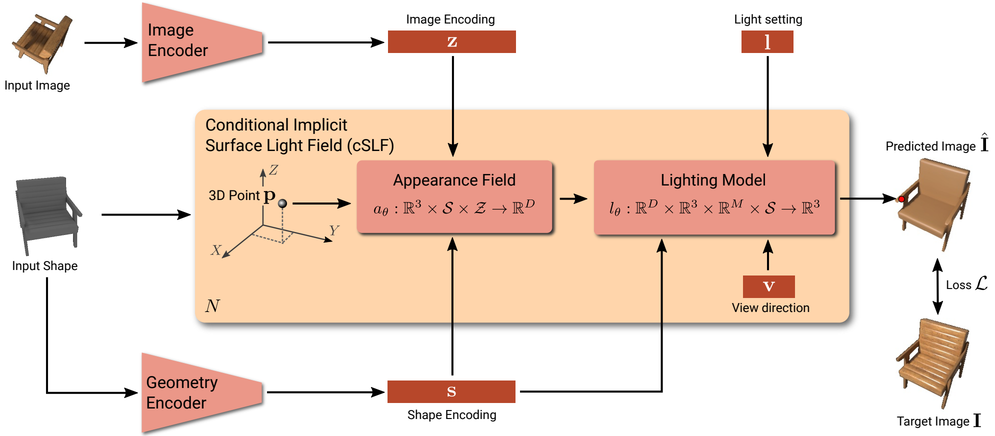
Oechsle, Niemeyer, Mescheder, Strauss and Geiger: Learning Implicit Surface Light Fields. arXiv, 2020.

Overfitting to Single Objects



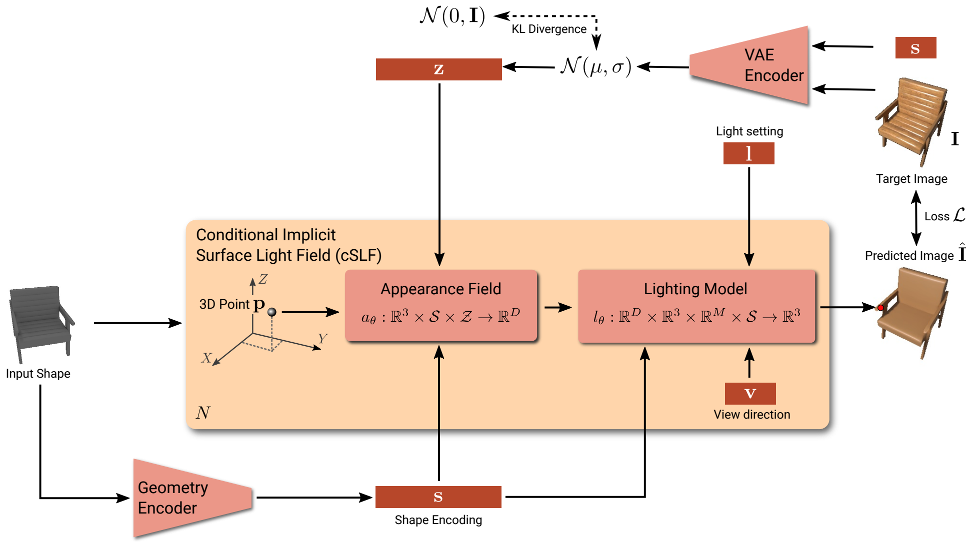
Oechsle, Niemeyer, Mescheder, Strauss and Geiger: Learning Implicit Surface Light Fields. arXiv, 2020.

Single-Image Appearance Prediction



Oechsle, Niemeyer, Mescheder, Strauss and Geiger: Learning Implicit Surface Light Fields. arXiv, 2020.

Generative Model



Oechsle, Niemeyer, Mescheder, Strauss and Geiger: Learning Implicit Surface Light Fields. arXiv, 2020.

How to obtain training data with materials?

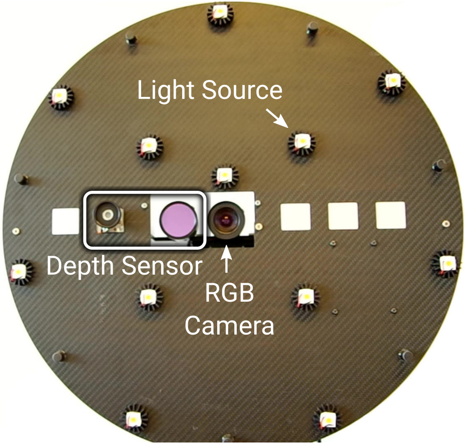
Joint Estimation of Pose, Geometry and svBRDF

Joint Estimation of Pose, Geometry and svBRDF

Goal: Dataset of 3D indoor scenes
captured with high accuracy
from a handheld mobile sensor.

Custom built sensor rig:

- Custom IR depth sensor similar to Microsoft Kinect
- Active illumination + RGB camera for material estimation



Schmitt, Donné, Riegler, Koltun and Geiger: On Joint Estimation of Pose, Geometry and svBRDF From a Handheld Scanner. CVPR, 2020.

Joint Estimation of Pose, Geometry and svBRDF

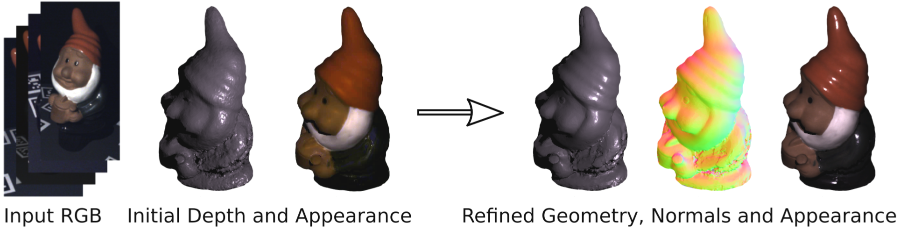
Schmitt, Donné, Riegler, Koltun and Geiger: On Joint Estimation of Pose, Geometry and svBRDF From a Handheld Scanner. CVPR, 2020.

Materials $\longleftrightarrow$ Geometry

$\longrightarrow$ Accurate geometry reconstruction requires known appearance properties

$\longleftarrow$ Accurate appearance estimation requires very well known geometry

$\longleftrightarrow$ Joint estimation requires **only a rough initialization** for both

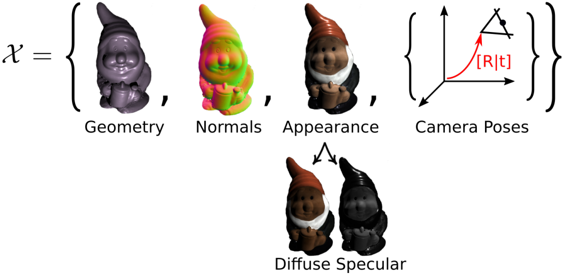


Schmitt, Donné, Riegler, Koltun and Geiger: On Joint Estimation of Pose, Geometry and svBRDF From a Handheld Scanner. CVPR, 2020.

Joint Estimation of Pose, Geometry and svBRDF

Contributions:

- **Joint** formulation
- **Single objective function** minimized using off-the-shelf gradient-based solvers
- **Meaningful segmentation** differentiably part of the optimization
- **Accurate geometry** with very fine details



$$\mathcal{X}^* = \underset{\mathcal{X}}{\operatorname{argmin}} \quad \mathcal{L}(\mathcal{X})$$

Joint Estimation of Pose, Geometry and svBRDF

Contributions:

- **Joint** formulation
- **Single objective function** minimized using off-the-shelf gradient-based solvers
- **Meaningful segmentation** differentiably part of the optimization
- **Accurate geometry** with very fine details

Reconstruction

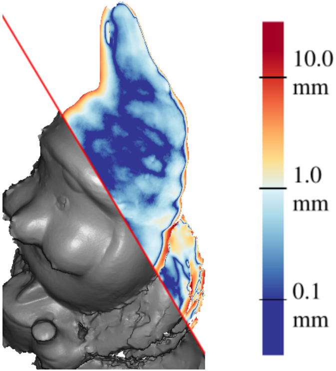
Segmentation

Schmitt, Donné, Riegler, Koltun and Geiger: On Joint Estimation of Pose, Geometry and svBRDF From a Handheld Scanner. CVPR, 2020.

Joint Estimation of Pose, Geometry and svBRDF

Contributions:

- **Joint** formulation
- **Single objective function** minimized using off-the-shelf gradient-based solvers
- **Meaningful segmentation** differentiably part of the optimization
- **Accurate geometry** with very fine details



Geometric error

Qualitative Results

Relighting

Novel Viewpoint

Conclusion:

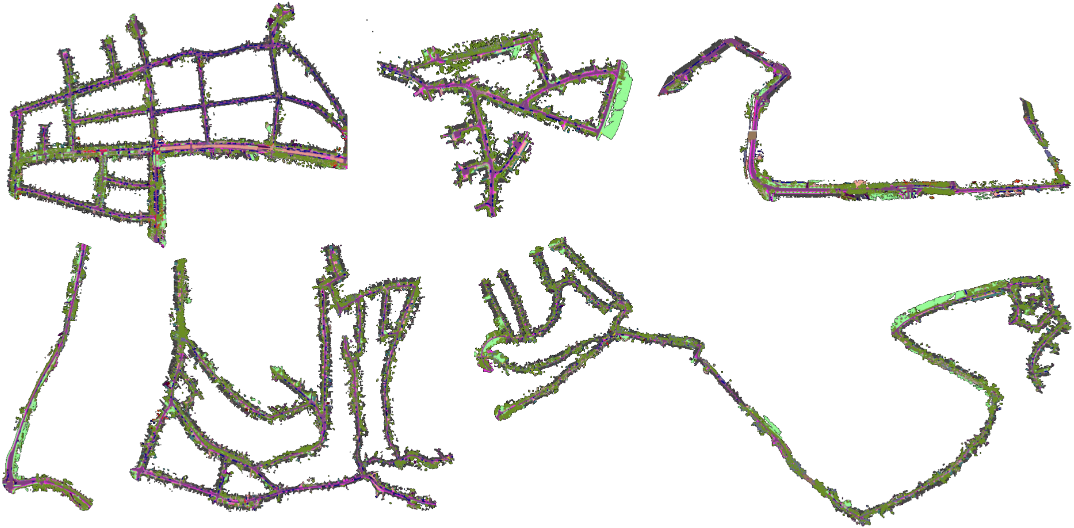
- Joint estimation helps
- This is only a first step
- Object-level reconstruction remains challenging with limited observations
- Scaling to larger scenes
- Scaling to scenes with external illumination

Schmitt, Donné, Riegler, Koltun and Geiger: On Joint Estimation of Pose, Geometry and svBRDF From a Handheld Scanner. CVPR, 2020.

How to obtain training data with semantic labels?

KITTI-360

KITTI-360



Xie, Kiefel, Sun, Geiger: Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. CVPR, 2016

KITTI-360

Sensors:

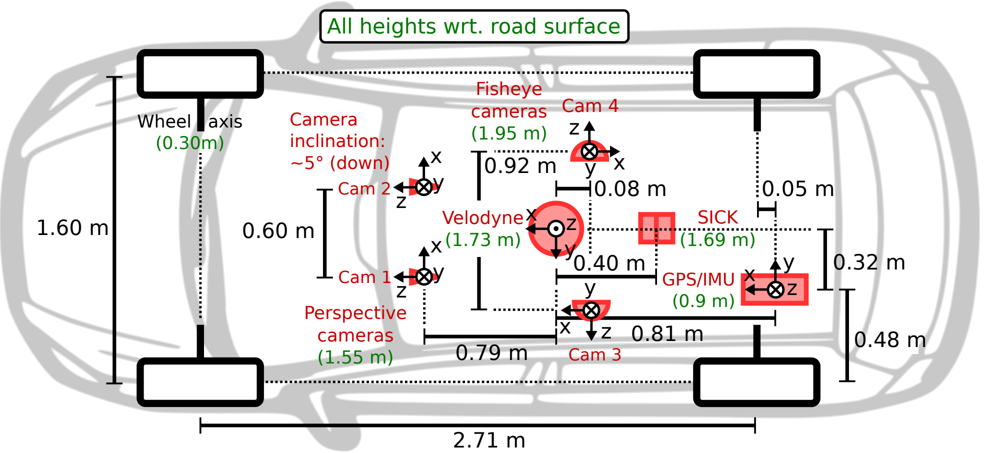
- Front-facing stereo camera
- 360° fisheye cameras
- Velodyne HDL 64 laser scanner
- SICK pushbroom laser scanner
- IMU/GPS localization system

Features:

- Driving distance: **73.7 km**   Frames: **4** $\times$ **83,000**
- All frames accurately **geolocalized** ($\Rightarrow$ OpenStreetMap)
- Semantic label definition consistent with Cityscapes, **19 classes** for evaluation
- Each instance assigned with a **consistent instance ID** across all frames

Xie, Kiefel, Sun, Geiger: Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. CVPR, 2016

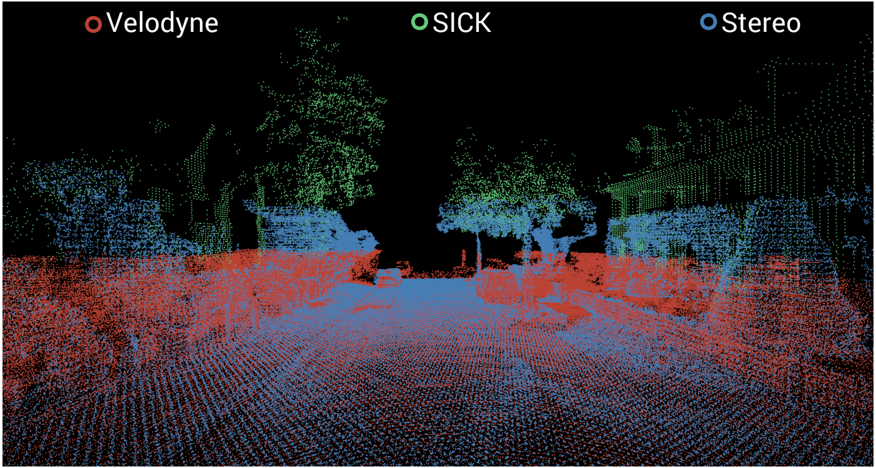
Sensors



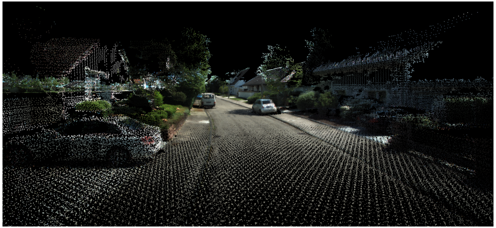
All heights wrt. road surface
Wheel axis
(0.30m)
1.60 m
Camera inclination: ~5° (down)
Cam 2
Cam 1
0.60 m
Perspective cameras
(1.55 m)
Fisheye cameras
(1.95 m)
Cam 4
Cam 3
0.92 m
0.08 m
0.05 m
Velodyne
(1.73 m)
SICK
(1.69 m)
0.40 m
0.32 m
GPS/IMU
(0.9 m)
0.81 m
0.79 m
0.48 m
2.71 m
x
y
z

360° 2D Sensors

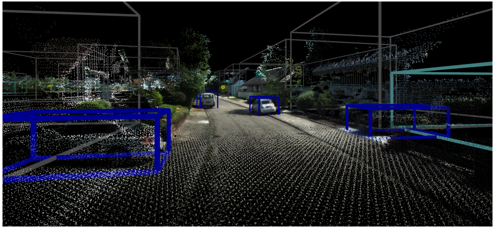
360° 3D Sensors



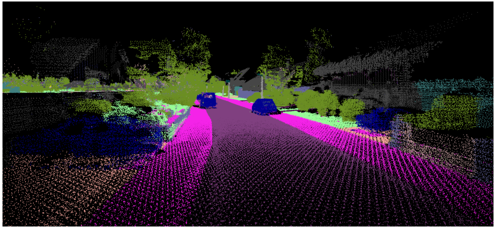
3D Annotations



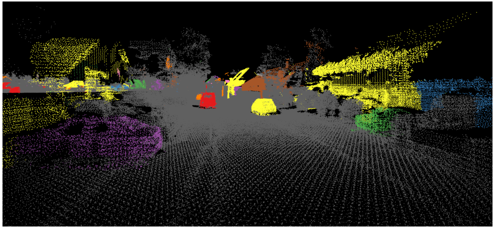
RGB



Bounding Box



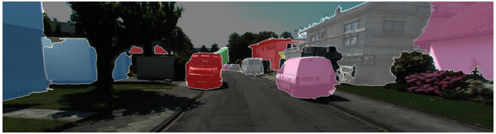
Semantic



Instance

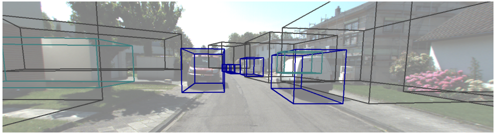
2D Annotations

Semantic



Instance

Confidence



Bounding Box

Thank you!

http://autonomousvision.github.io

European Research Council
Established by the European Commission

Federal Ministry of Education and Research

intel

Microsoft Research

NVIDIA