RayNet: Learning Volumetric 3D Reconstruction with Ray Potential

Despoina Paschalidou[1,5]   Ali Osman Ulusoy[2]   Carolin Schmitt[1]   Luc van Gool[3]   Andreas Geiger[1,4]

[1]Autonomous Vision Group, MPI for Intelligent Systems Tübingen   [2]Microsoft   [3]Computer Vision Lab, ETH Zürich & KU Leuven
[4]Computer Vision and Geometry Group, ETH Zürich   [5]Max Planck ETH Center for Learning Systems

Motivation

Problem: Reconstruct a dense 3D model using RGB images captured from different views.

Contribution: Combine the benefits of a **CNN-based** approach with the strengths of a **structured** model with **ray potentials**.

	Ray potentials	CNN-based	Raynet
Data driven		✓	✓
View invariant representations		✓	✓
Large receptive field		✓	✓
Physics of perspective geometry	✓		✓
Occlusion relationships	✓		✓
Consistency across views	✓		✓

Our Solution: Integrates a Multi-view CNN that learns view-invariant feature representations with an MRF that explicitly encodes the physics of perspective projection and occlusion.

RayNet

Multi-View CNN / Unrolled Message Passing / Markov Random Field (MRF) / Depth Prediction & Loss

Inference:

1. Multi-view CNN estimates depth distributions
2. Fuse $s_i^r \in \mathcal{S}$ using message passing in MRF
3. Predict a globally consistent depth distribution $p(d_r = d_i^r)$

Training:

- Naïve backpropagation through the unrolled MRF is **intractable** due to the large number of messages that need to be stored
- Use **stochastic ray sampling scheme** to produce mini-batches

Multi-view Stereopsis with Ray Potentials

Given depth probabilities $\{s_i^r\}$ for every ray/pixel $r \in \mathcal{R}$ we encourage the predicted depth at pixel r to coincide with the first occupied voxel along the ray using ray potentials $\psi_r(\mathbf{o_r})$.

- **Occupancy variable**

  $o_i \in \{0, 1\}$ for every voxel $i \in \mathcal{X}$

- **Unary potentials**

  $$\varphi_i(o_i) = \underbrace{\gamma^{o_i}(1-\gamma)^{1-o_i}}_{\text{Bernoulli prior on } o_i}$$

- **Ray potentials**

  $$\psi_r(\mathbf{o_r}) = \underbrace{\sum_{i=1}^{N_r}}_{\substack{\text{sum over} \\ \text{voxels} \\ \text{along } r}} o_i^r \underbrace{\prod_{j<i}(1-o_j^r)}_{\substack{= 1 \text{ for first} \\ \text{occupied voxel}}} \underbrace{s_i^r}_{\substack{\text{depth} \\ \text{probability}}}$$

Qualitative Evaluation

DTU Dataset: 124 indoor scenes with ~ 50 images

Image — Ground-truth — SurfaceNet — RayNet — Image — Ground-truth — SurfaceNet — RayNet

Aerial Dataset: ~ 200 images with LIDAR ground-truth

Image — ZNCC — Ulusoy et al. 2015 — Hartmann et al. 2017 — Multi-View CNN — RayNet

Quantitative Evaluation

Methods	Aerial Dataset				DTU Dataset - BUDDHA				DTU Dataset - BIRD			
	Accuracy		Completeness		Accuracy		Completeness		Accuracy		Completeness	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Ulusoy et al. 2015	0.0790	0.0167	**0.0088**	**0.0065**	4.784	3.552	0.953	0.402	6.024	4.623	2.996	0.898
Hartmann et al. 2017	0.0907	0.0285	0.0209	0.0209	**0.637**	**0.206**	1.057	0.475	**1.881**	**0.271**	4.16711	1.044
SurfaceNet (LR) (Ji et al. 2017)	-	-	-	-	2.034	1.677	1.453	1.141	2.887	2.468	2.330	1.556
RayNet	**0.0611**	**0.0160**	0.0125	0.0085	0.663	0.470	**0.558**	**0.251**	2.618	1.680	**0.983**	**0.668**