EdgeNet: Semantic Scene Completion from a Single RGB-D Image

Aloisio Dourado, Teofilo Emidio de Campos
University of Brasilia
Brasilia, Brazil
aloisio.dourado.bh@gmail.com, t.decampos@st-annes.oxon.org

Hansung Kim
University of Southampton
Southampton, UK
H.Kim@soton.ac.uk

Adrian Hilton
University of Surrey
Surrey, UK
a.hilton@surrey.ac.uk

ICPR 2020
Milan, January 2021
Semantic Scene Completion

Introduced by Song et al. [1] in 2017

Trained a 3D CNN that jointly deals with completion and semantic segmentation

Previous Works

**Depth maps only**

- **SSCNET**: Song et al. [1]

---

Previous Works

**Depth maps only**

- **SSCNET: Song et al. [1]**
  - Encoder-decoder network architecture

---

Previous Works

**Depth maps only**

- **SSCNET: Song et al. [1]**
  - Encoder-decoder network architecture
  - Proposed F-TSDF encoding

---

Previous Works

**Depth maps only**

- **SSCNET: Song et al. [1]**
  - Encoder-decoder network architecture
  - Proposed F-TSDF encoding
  - Introduced SUNCG Dataset

---

Previous Works

**Depth maps only**

- SSCNET: Song et al. [1]
- Guo and Tong [2]:
  - 2D features projected to 3D


Previous Works

**Depth maps only**

- SSCNET: Song et al. [1]
- Guo and Tong [2]:
  - 2D features projected to 3D

Neglect the RGB channels from the input data


Previous Works

Depth maps plus RGB

• Guedes et al.[3]

Previous Works

Depth maps plus RGB

- Guedes et al.[3]

Previous Works

Depth map plus 2D segmentation

- Two stream 3D semantic scene completion: Garbade et al.[4]

Previous Works

**Depth map plus 2D segmentation**

- Two stream 3D semantic scene completion: Garbade *et al.*[4]
- TNetFusion: Liu *et al.*[5]

---


Previous Works

**Depth map plus 2D segmentation**

- Two stream 3D semantic scene completion: Garbade *et al.*[4]
- TNetFusion: Liu *et al.*[5]

---


Requires a complex two step training procedure
Our Approach: EdgeNet

- We extract boundary information from RGB data and project to 3D...
Our Approach: EdgeNet

- ...then, we apply F-TSDF to the projected edge volume
Network Architecture

Conv3D (channels, size, strides) + BatchNorm + ReLU
Maxpooling3D (size, strides)
Conv3D (channels, size, strides, dilation=2) + BatchNorm + ReLU
Upsample3D (size)
Conv3D (channels, size, strides) + Categorical Cross Entropy Loss
Network Architecture - Fusion Schemes
Network Architecture - Fusion Schemes

Mid Fusion Scheme

Depth

Edges

Input Branch

Input Branch

Concat.

Encoding Branch

Latent Features

Decoding Branch

ch=4 ch=8

ch=32 ch=64

ch=128

ch=64
Network Architecture - Fusion Schemes

Late Fusion Scheme

Depth

Edges

Input Branch

Enc. Branch

Latent Features

Decoding Branch

ch=4 ch=8

ch=16 ch=32

ch=128

ch=64
Network Architecture - Fusion Schemes
Network Architecture - Fusion Schemes

11 GB NVIDIA GTX1080-TI
Datasets

- **SUNCG***

(a) SUNCG dataset  
(b) 3D Scene  
(c) Synthetic depth and volumetric ground truth

- **NYUDv2**

---

---
Quantitative Results

• SUNCG
  • New state-of-the-art result (70.3% avg. IoU)

• NYUD-V2
  • Our solution surpassed previous end-to-end approaches (33.7% avg. IoU)
  • EdgeNet's results are similar to non end-to-end solutions, with a much simpler training pipeline.

• Best Fusion Scheme: Mid Fusion (EdgeNet-MF)
Qualitative Results

Ground Truth  
SSCNet  
EdgeNet-MF

Higher overall accuracy
Qualitative Results

Ground Truth

SSCNet

EdgeNet-MF

Hard-to-detect classes
Qualitative Results

NYU Ground Truth errors
Conclusions

- A new end-to-end network architecture
- A new strategy to encode data from RGB channels
- Visually perceptible improvements in 3D
- Improvement over the state-of-the-art result on SUNCG
- We surpassed other end-to-end approaches on NYUDv2
Thank you!