What From Where. In 3D!
Learning Semantic Segmentation from 3D Models

A. Golbert, T. Halperin, D. Arnon

RAFAEL
Image Processing Department
Urban Image Segmentation

- Semantic data understanding
  - Mapping / Modeling
  - Urban navigation
  - Autonomous / Assisted Driving
  - Registration to white map
Talk Outline

• 2D Semantic Segmentation
  – ISPRS Benchmark
  – F1 Loss
  – Comparison to SOA
• Minimal labeled Data
  – 3D reconstruction
  – System Outline to leverage “World Experience”
• Results
ISPRS challenge

ISPRS Potsdam dataset
Fully Convolutional Network

- Convert fully Connected layers to convolutions
- Arbitrary image input size
- Output size is smaller than input
  - Shift and stitch
  - Skip layers + deconvolution layers

Fully Convolutional Networks for Semantic Segmentation
Jonathan Long, Evan Shelhamer, Trevor Darrell
ISPRS results

Ground Truth

FCN Labeling

Proprietary of Rafael – Advanced Defense Systems Ltd
ISPRS results
ISPRS results
ISPRS results
ISPRS results

ISPRS Potsdam dataset
ISPRS results
ISPRRS results
Net Details

• F1 Loss
• 130M Parameters
• Data augmentation – crops and mirror
• AdaGrad with Weight Decay
• Train time ~24 hours on TitanX
Net Details

• F1 Loss:

\[ F1(l) = 2 \times \frac{Prec(l) \times Rec(l)}{Prec(l) + Rec(l)} \]

\[ tp(l) = \#\{l_i = 1 \text{ and } gt_i = 1\} \]
\[ fp(l) = \#\{l_i = 1 \text{ and } gt_i = 0\} \]
\[ fn(l) = \#\{l_i = 0 \text{ and } gt_i = 1\} \]

\[ Precision = \frac{tp(l)}{tp(l) + fp(l)} \]
\[ Recall = \frac{tp(l)}{tp(l) + fn(l)} \]

Net Details
## ISPRS results

<table>
<thead>
<tr>
<th>Vaihingen DataSet</th>
<th>Impervious surfaces</th>
<th>Building</th>
<th>Low vegetation</th>
<th>Tree</th>
<th>Car</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paisitkriangkrai et al. (2015)</td>
<td>89.5</td>
<td>93.2</td>
<td>82.3</td>
<td>88.2</td>
<td>63.3</td>
<td>83.3</td>
</tr>
<tr>
<td>RGBD</td>
<td>84.9</td>
<td>88.4</td>
<td>70.1</td>
<td>82.2</td>
<td>74.1</td>
<td>80</td>
</tr>
<tr>
<td>RGBD_F1</td>
<td>87.2</td>
<td>90.6</td>
<td>73.3</td>
<td>84.8</td>
<td>82.6</td>
<td>83.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Potsdam DataSet</th>
<th>Impervious surfaces</th>
<th>Building</th>
<th>Low vegetation</th>
<th>Tree</th>
<th>Car</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sherrah et al (2016)</td>
<td>91.4</td>
<td>95.3</td>
<td>85.1</td>
<td>87.3</td>
<td>88.7</td>
<td>89.6</td>
</tr>
<tr>
<td>RGBD</td>
<td>89</td>
<td>95.3</td>
<td>82.4</td>
<td>85.3</td>
<td>91.9</td>
<td>89</td>
</tr>
<tr>
<td>RGBD_F1</td>
<td>90.3</td>
<td>95.6</td>
<td>83.2</td>
<td>85.8</td>
<td>93.3</td>
<td>89.7</td>
</tr>
</tbody>
</table>

- Trained FCN based on VGGnet trained on Pascal-Voc dataset.
- The network is trained on ~1 Giga of labeled pixels.
Human Learning

- People learn from a lot fewer examples
- How do we solve the impending worldwide Mechanical Turk shortage?
3D Model

- Contains a host of additional information on the scene
Exploiting 3D

3D Model

Sparse Annotation

SVM Classification
In 3D Space:
Exploiting 3D

Model Labeling

3D MODEL + IMAGES

Extract 3D & Texture Features

Sparse User Labeling

SVM Learning

Labeled 3D Model

Net Training

Labeled 3D Model

Projection

Labeled Images

FCNN Training
Results

3D SVM: F1 score 0.71

2D FCN: F1 score 0.7
Results

![Results](image)

3D SVM

2D FCN

Legend:
- Blue: Roof
- Red: Tiles
- Yellow: Wall
- Green: Vegetation
- Black: Road
- Brown: Ground
Results
Results
Church of Annunciation Classification
Church of Annunciation Classification
Church of Annunciation Classification
This work was funded in part by the Omek Consortium and was done in part as a guest researcher at the Deep Vision Lab in TAU headed by Prof. Lior Wolf.

Thanks to Lior Uzan from TAU for invaluable discussion about the F1 Loss.
Questions