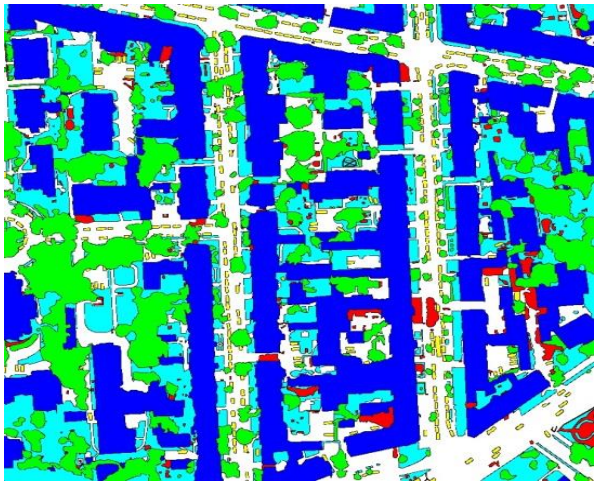
What From Where. In 3D !

Learning Semantic Segmentation from 3D Models

A. Golbert, T. Halperin, D. Arnon

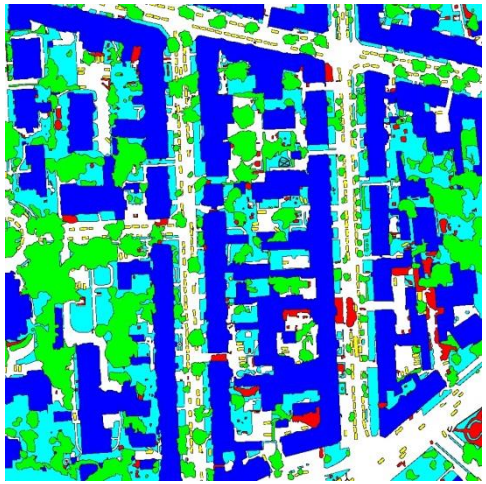
RAFAEL

Image Processing Department



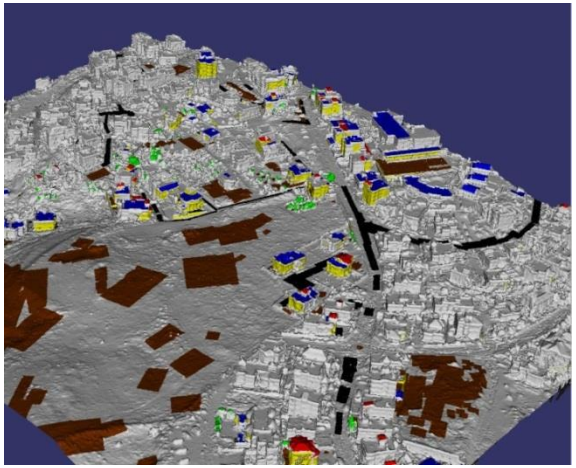
Urban Image Segmentation

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



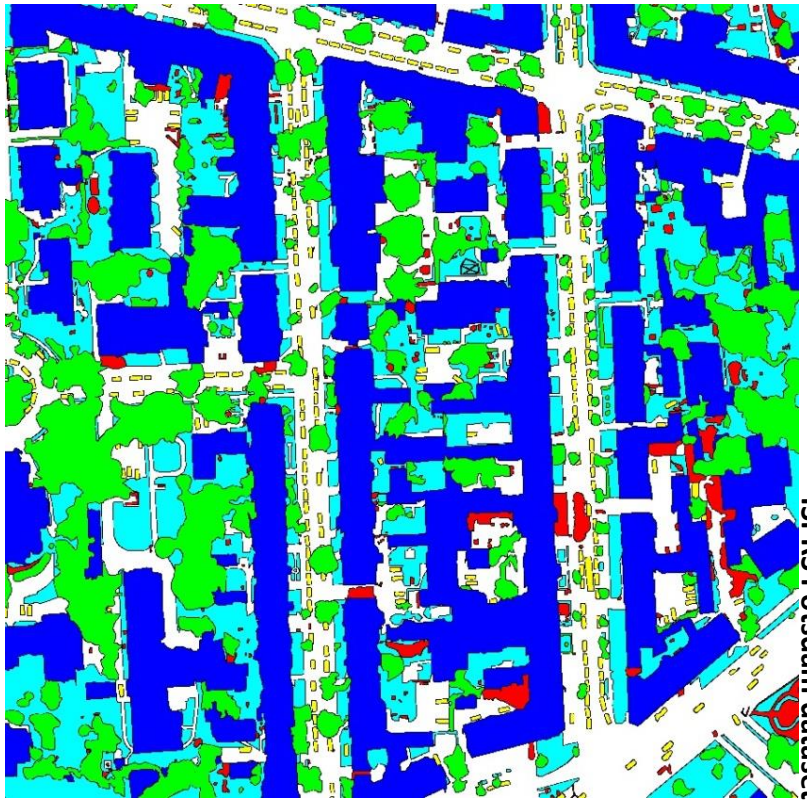
Talk Outline

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



ISPRS challenge

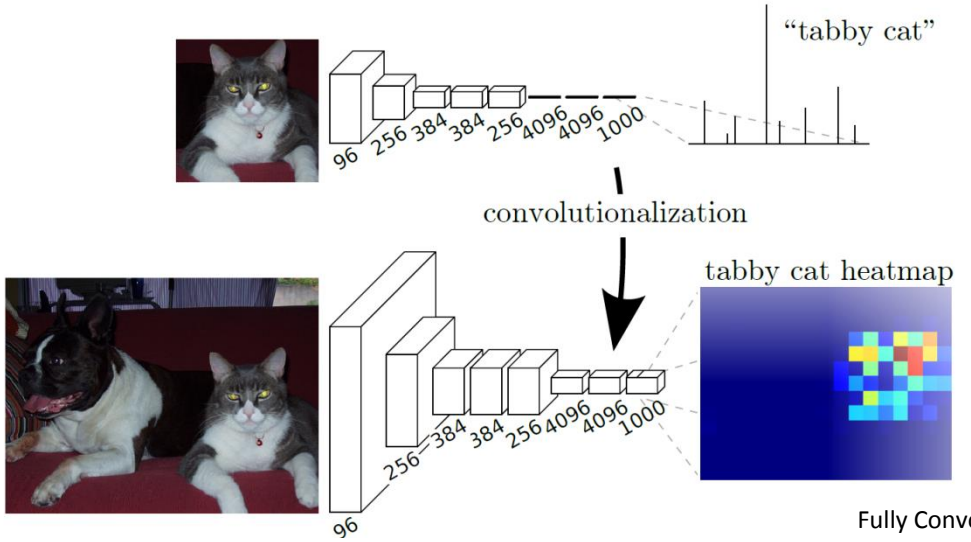
ISPRS Potsdam dataset



ISPRS Potsdam dataset

Fully Convolutional Network

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

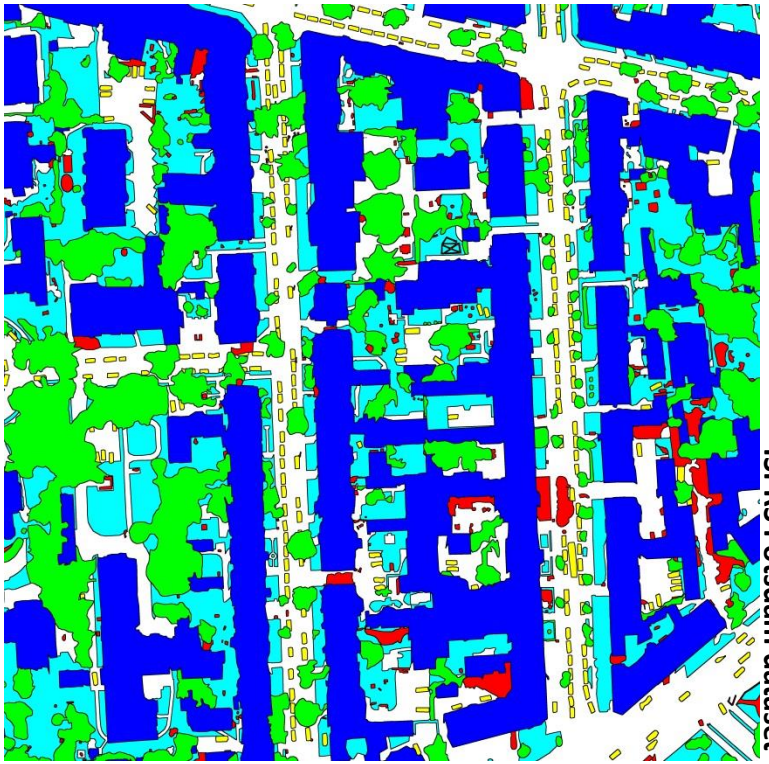


Fully Convolutional Networks for Semantic Segmentation
Jonathan Long, Evan Shelhamer, Trevor Darrell

ISPRS results

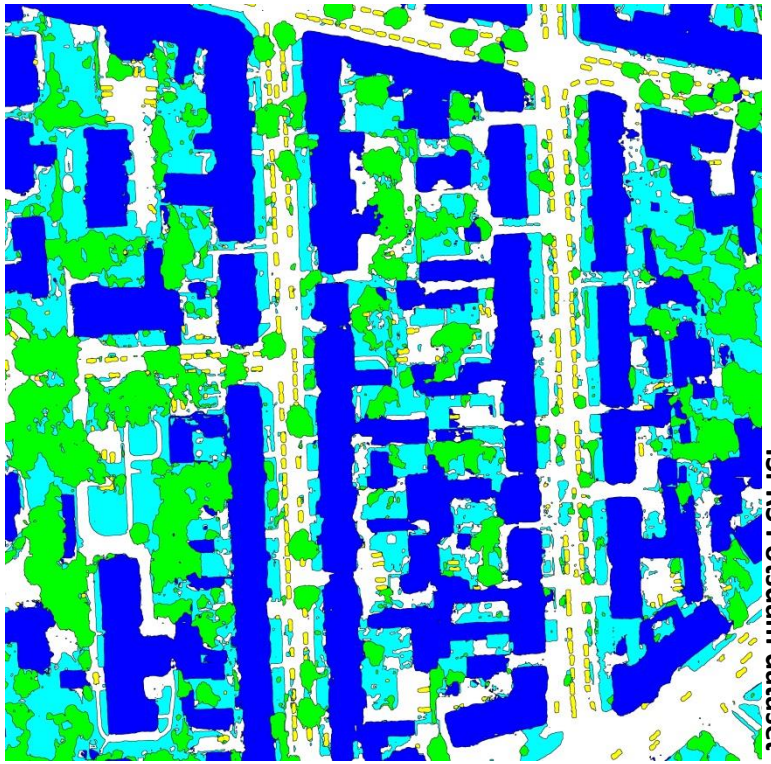
Building
Grass
Surface
Tree
Car
Undefined

Ground Truth



ISPRS Potsdam dataset

FCN Labeling

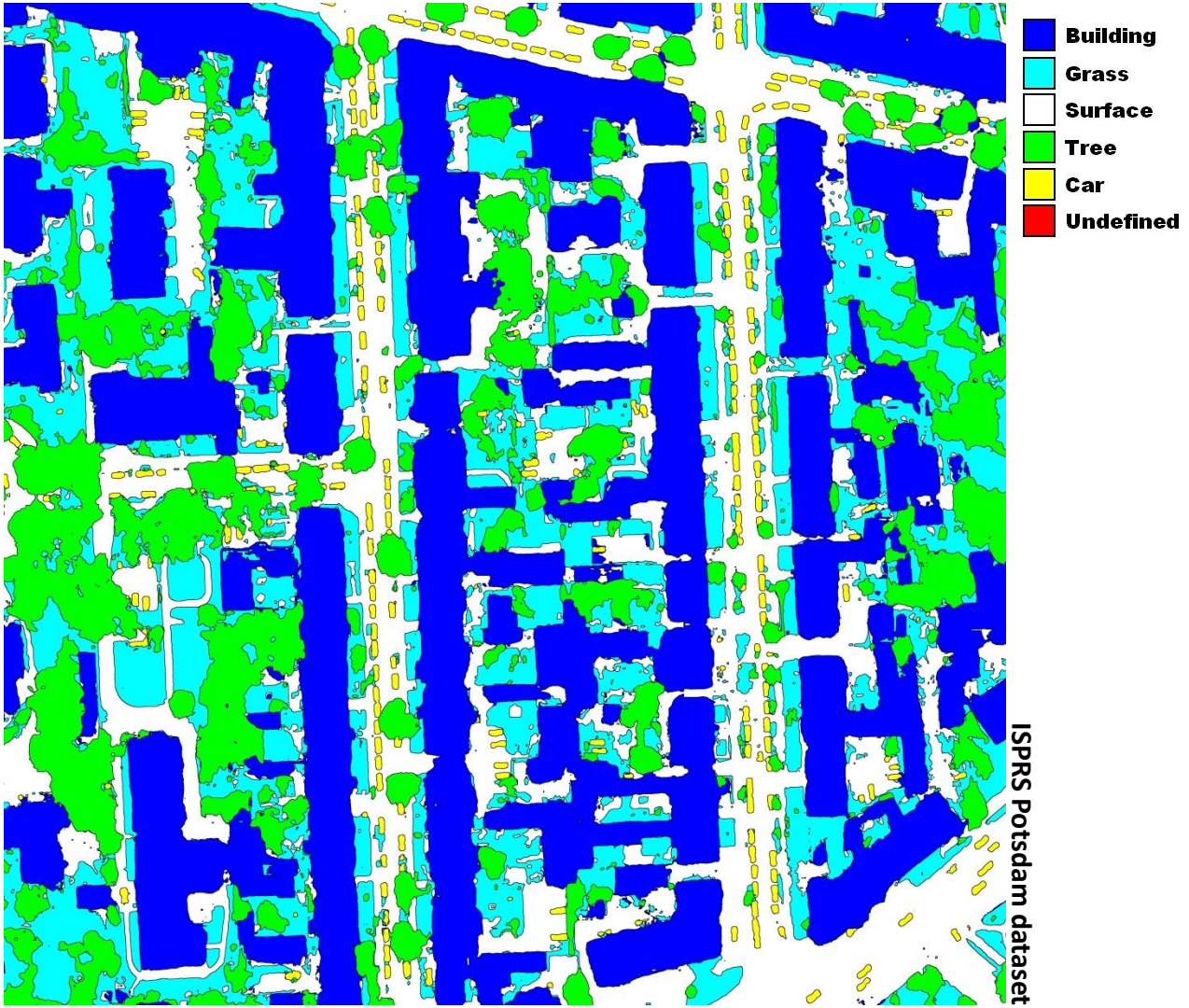


ISPRS Potsdam dataset

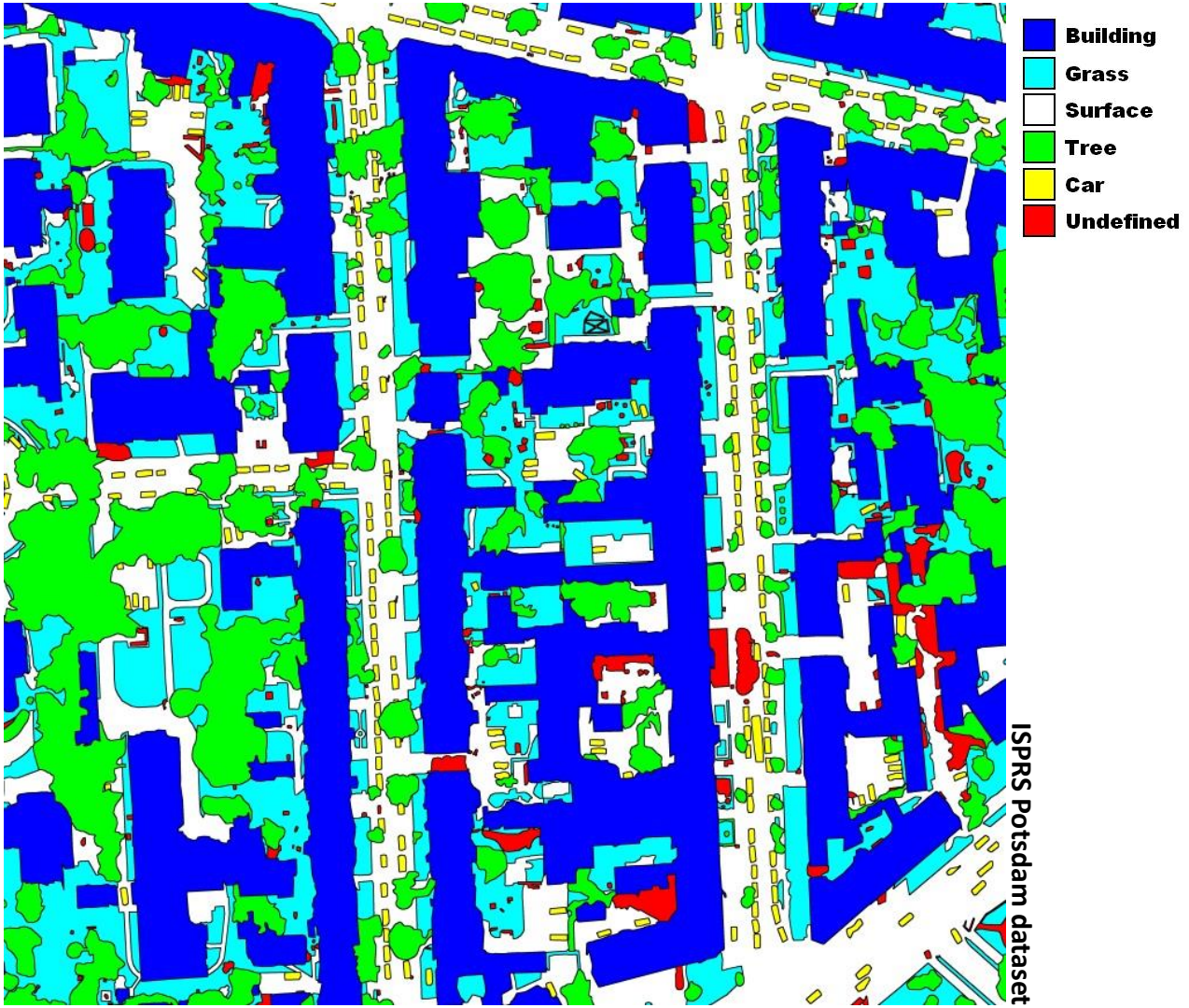
ISPRS results

ISPRS Potsdam dataset

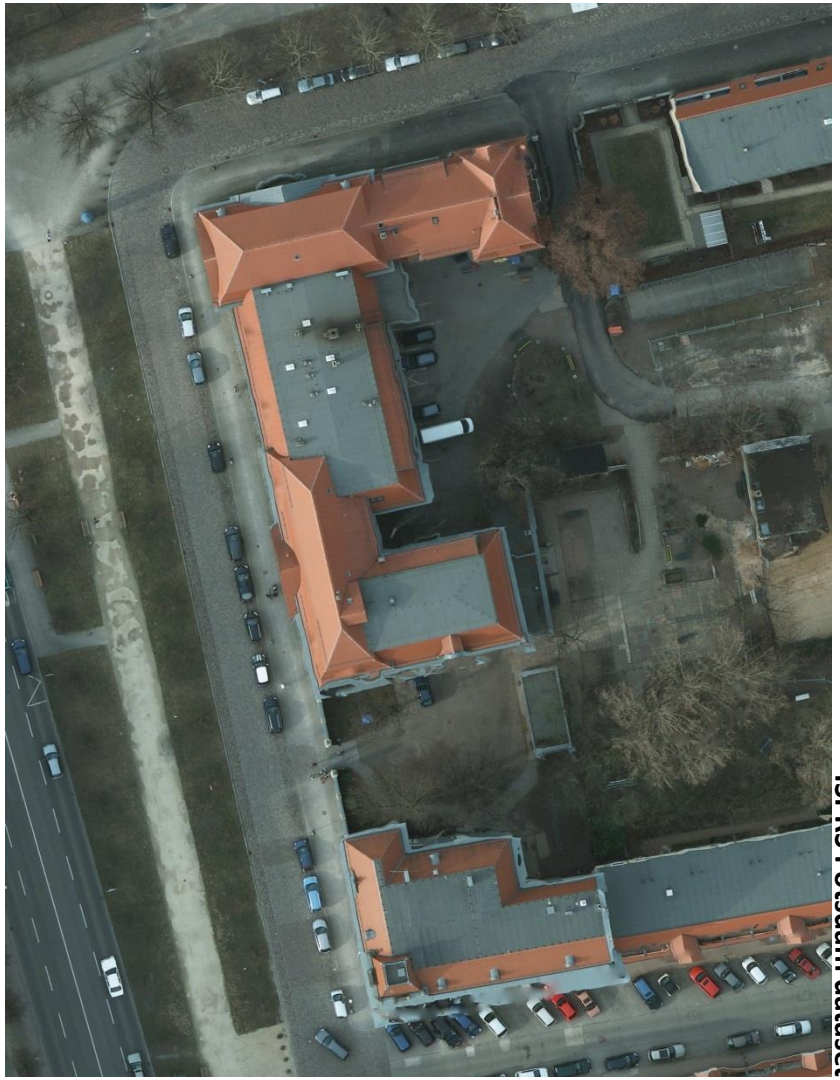
ISPRS results



ISPRS results



ISPRS results

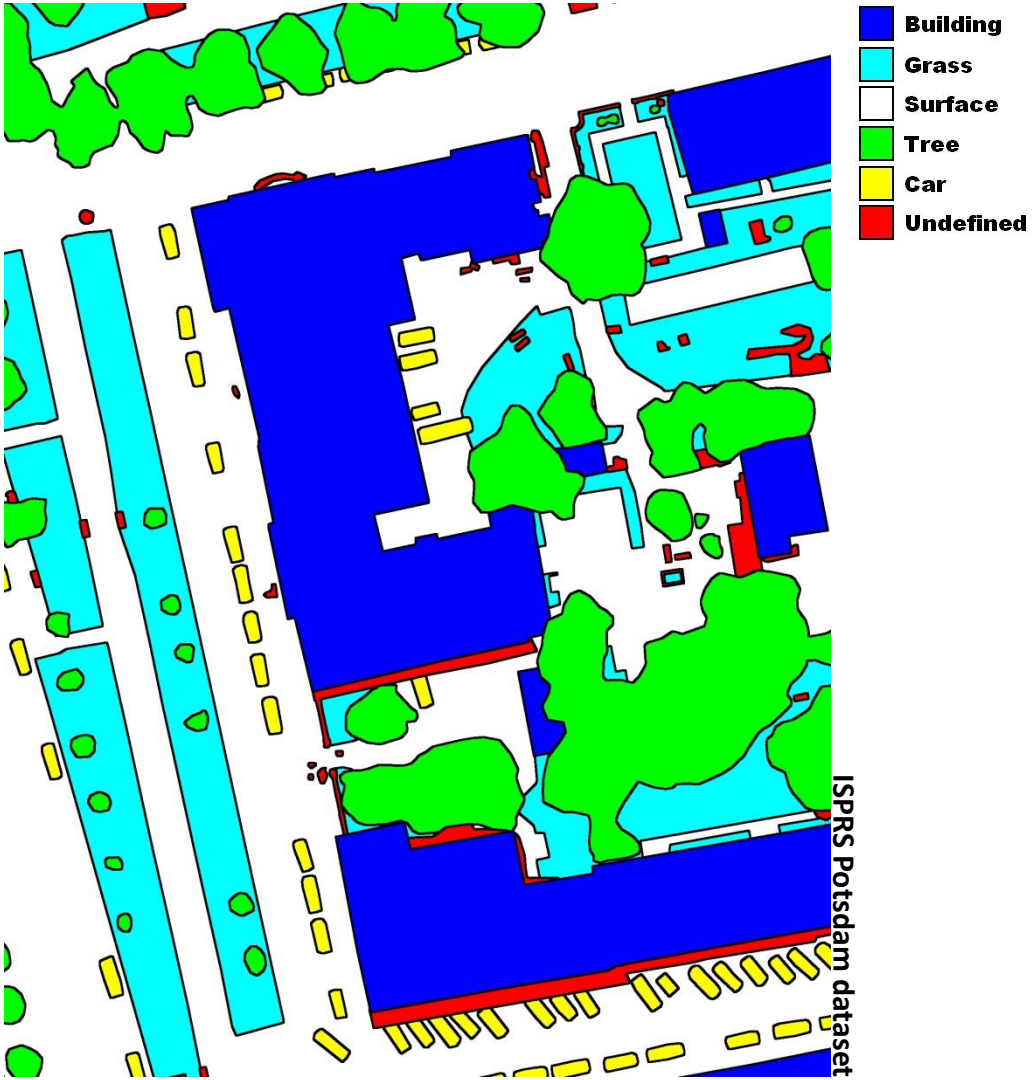


Building
Grass
Surface
Tree
Car
Undefined

ISPRS Potsdam dataset

ISPRS results

ISPRS 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:

$$F1(l) = 2 * \frac{Prec(l) * Rec(l)}{Prec(l) + Rec(l)}$$

$$Precision = \frac{tp(l)}{tp(l) + fp(l)}$$

$$Recall = \frac{tp(l)}{tp(l) + fn(l)}$$

$$tp(l) = \#\{l_i = 1 \; and \; gt_i = 1\} = \sum_i l_i gt_i$$
$$fp(l) = \#\{l_i = 1 \; and \; gt_i = 0\} = \sum_i l_i(1 - gt_i)$$
$$fn(l) = \#\{l_i = 0 \; and \; gt_i = 1\} = \sum_i (1 - l_i) gt_i$$

ISPRS results

Vaihingen DataSet	Impervious surfaces	Building	Low vegetation	Tree	Car	Mean
Paisitkriangkrai et al. (2015)	89.5	93.2	82.3	88.2	63.3	83.3
RGBD	84.9	88.4	70.1	82.2	74.1	80
RGBD_F1	87.2	90.6	73.3	84.8	82.6	**83.7**

Potsdam DataSet	Impervious surfaces	Building	Low vegetation	Tree	Car	Mean
Sherrah et al (2016)	91.4	95.3	85.1	87.3	88.7	89.6
RGBD	89	95.3	82.4	85.3	91.9	89
RGBD_F1	90.3	95.6	83.2	85.8	93.3	**89.7**

x20 larger dataset

- 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?

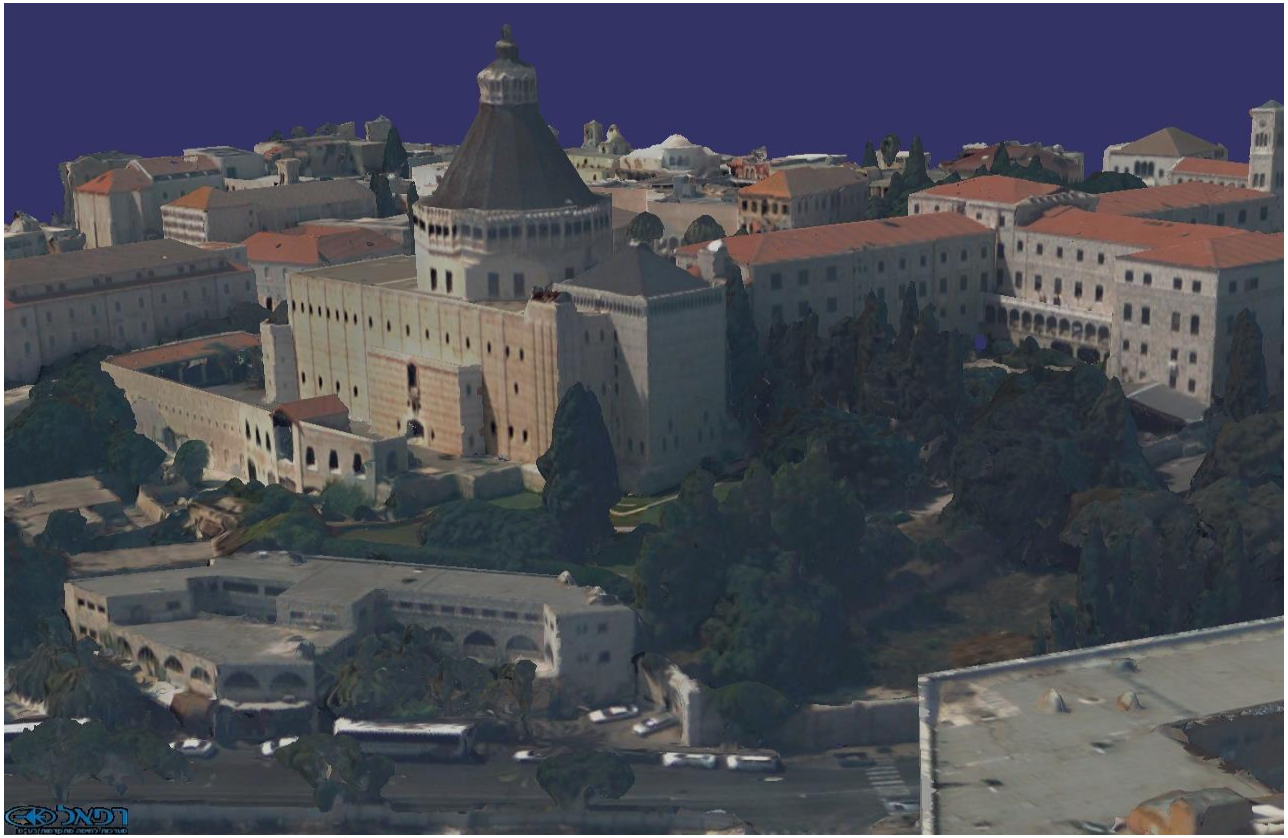
DOG?

DOG?

DOG?

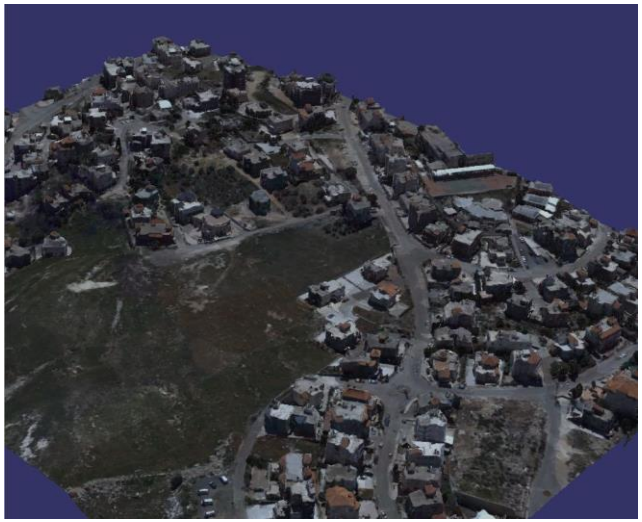
3D Model

- Contains a host of additional information on the scene

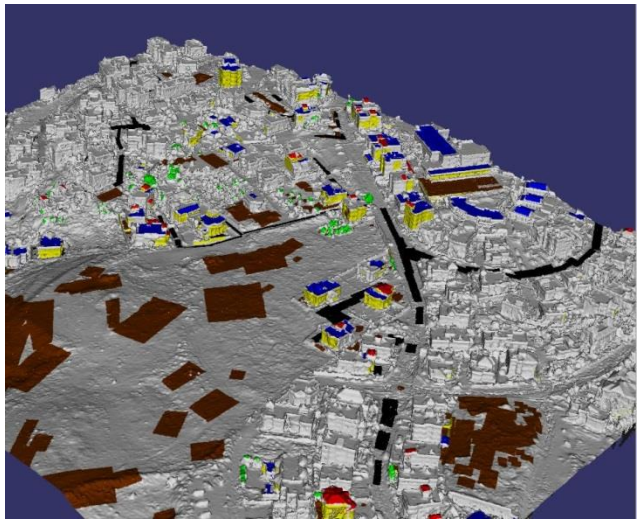


Exploiting 3D

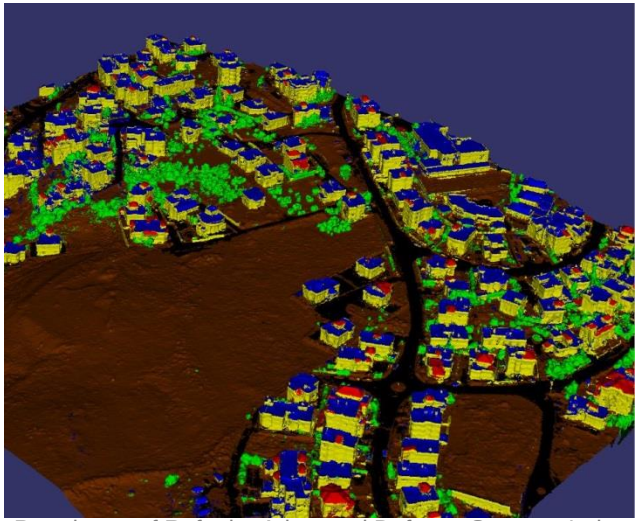
3D Model



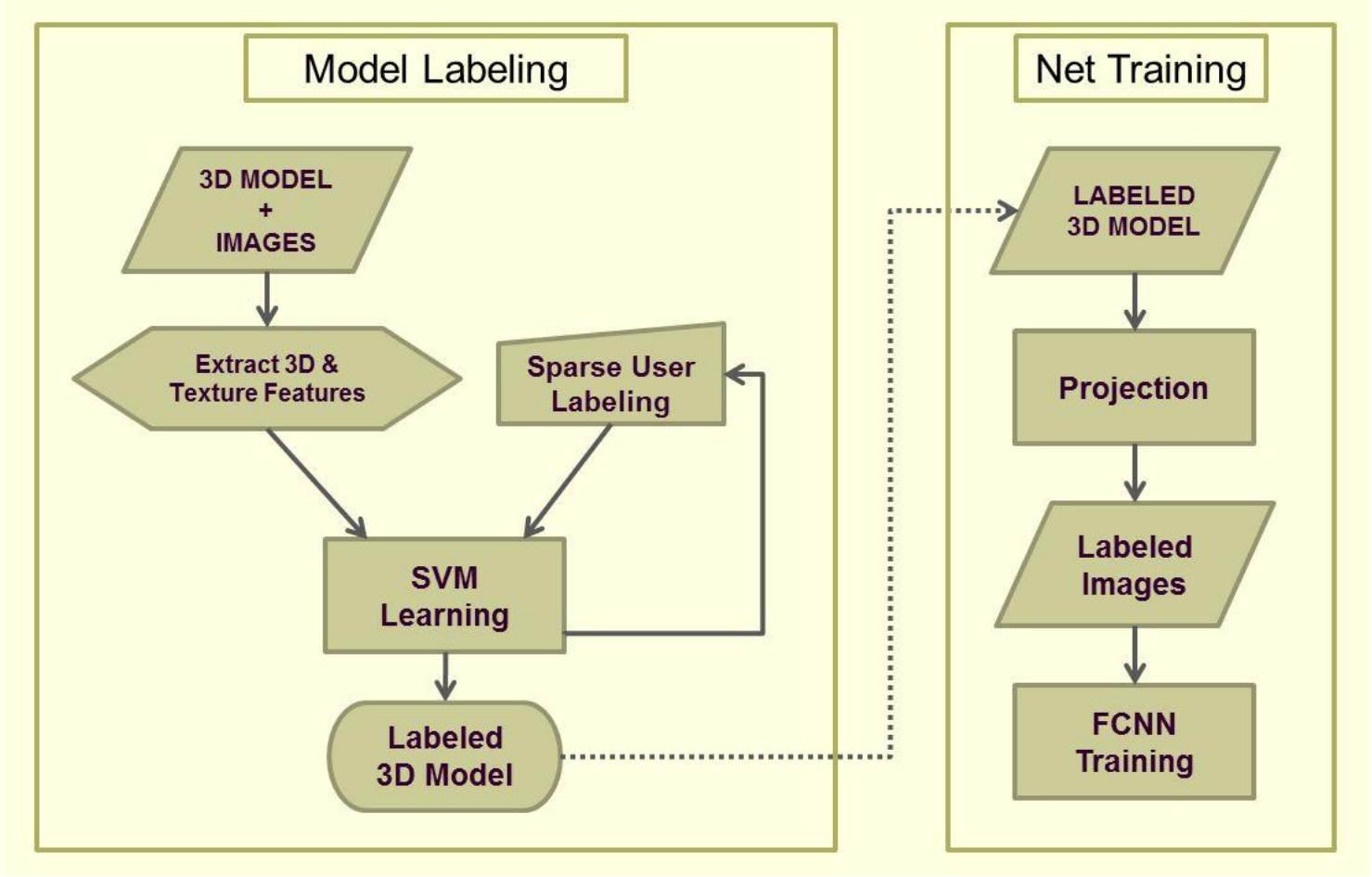
Sparse Annotation



SVM Classification
In 3D Space:

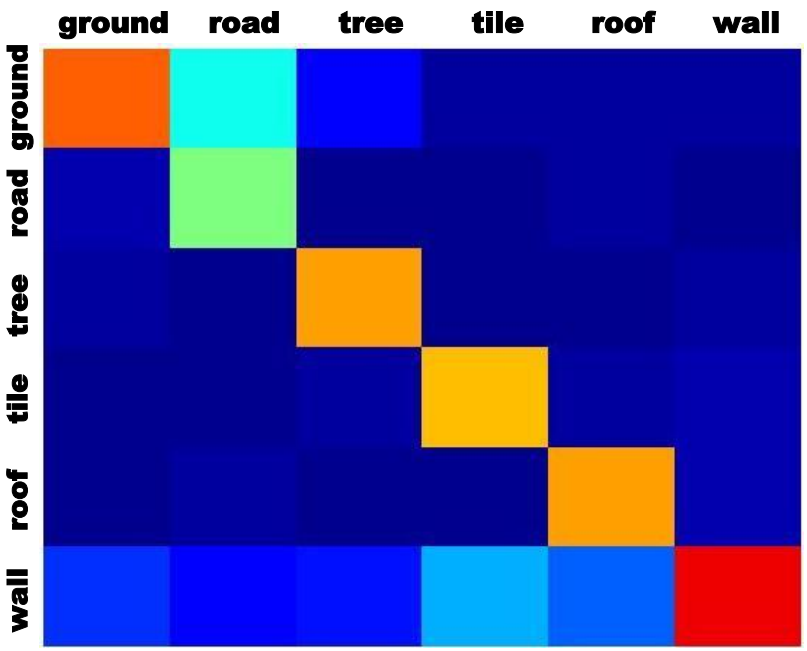


Exploiting 3D

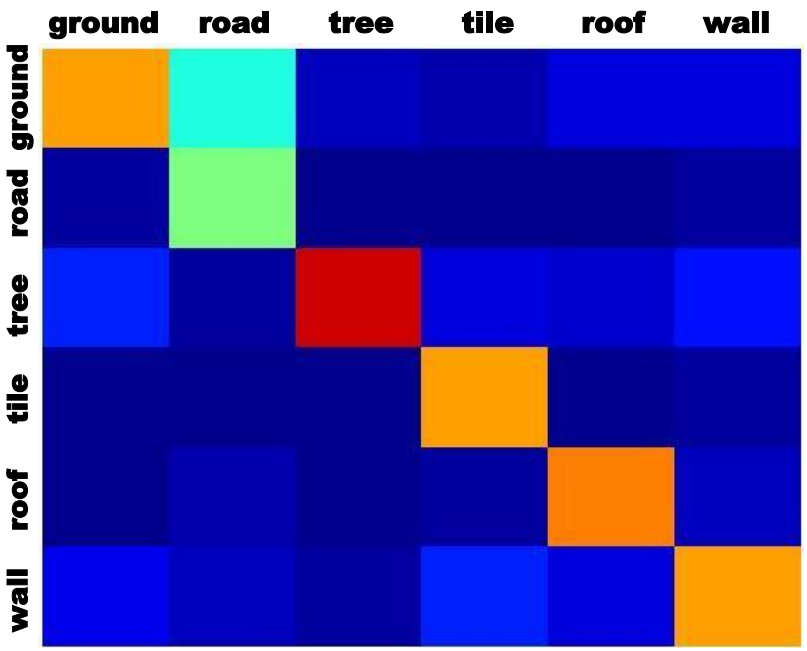


Results

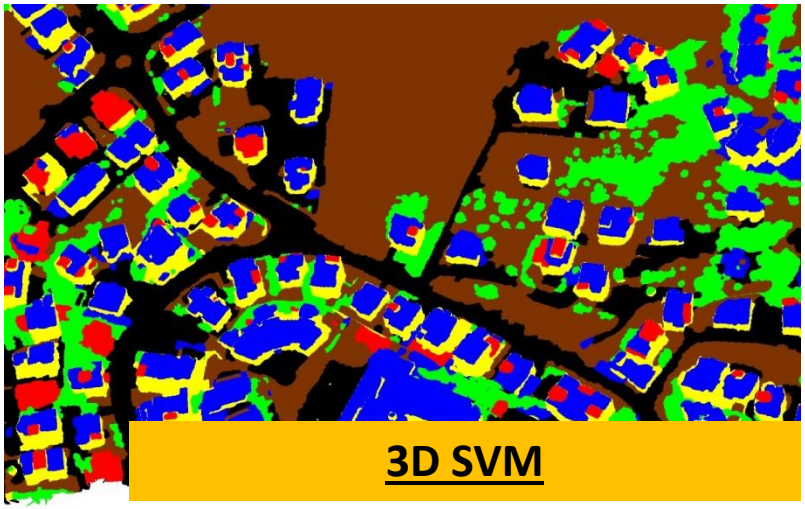
3D SVM: F1 score 0.71



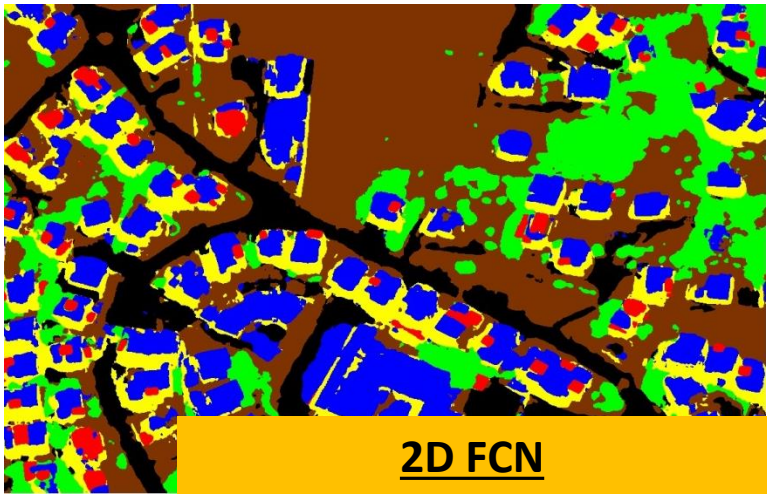
2D FCN: F1 score 0.7



Results



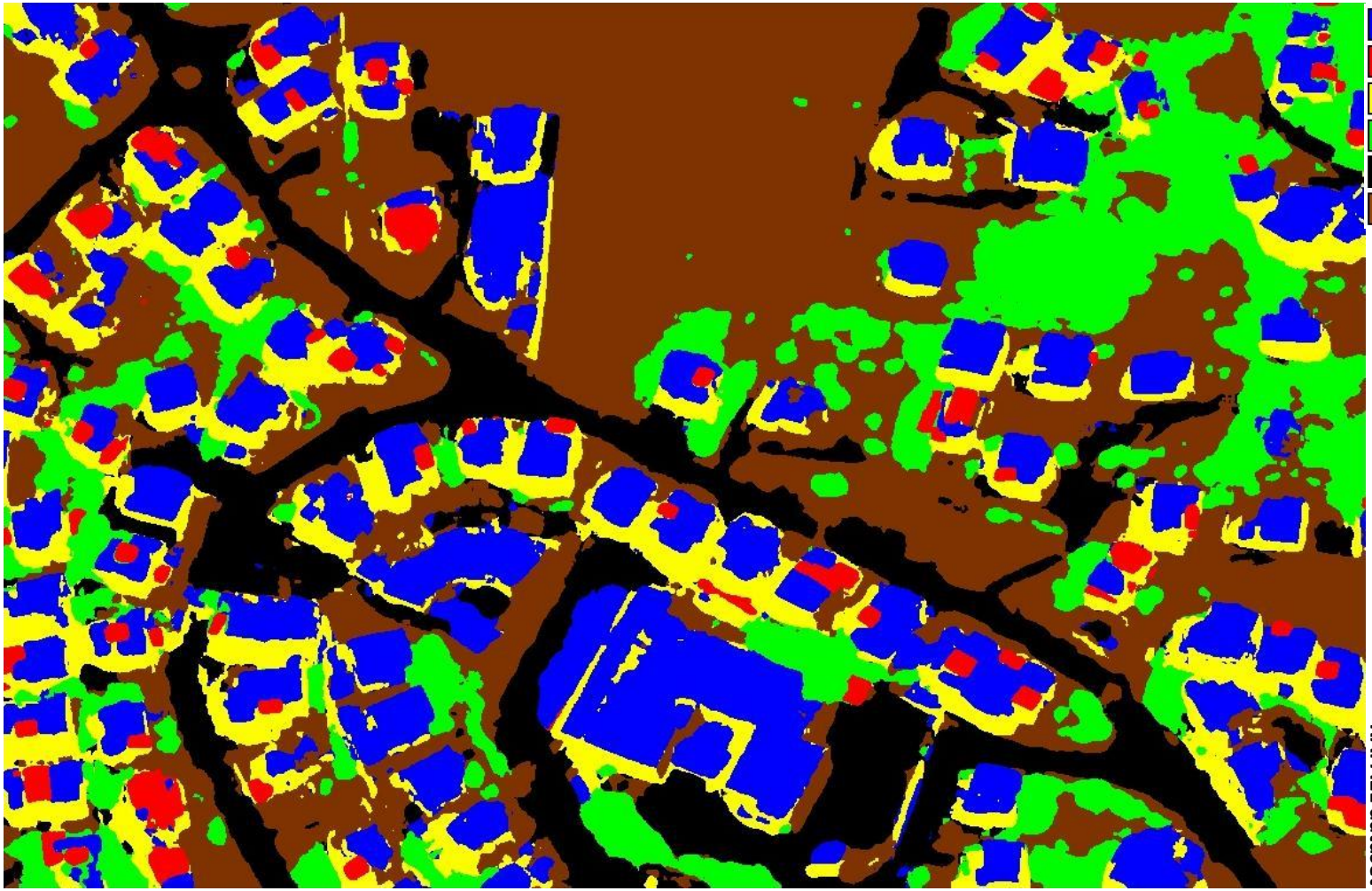
3D SVM



2D FCN

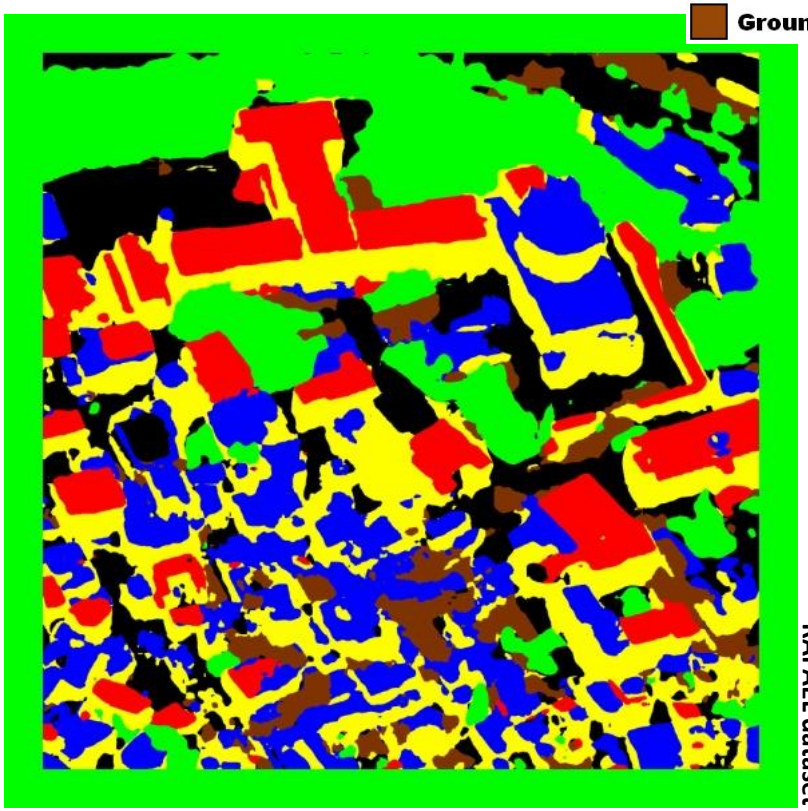
Results

Results



Church of Annunciation Classification

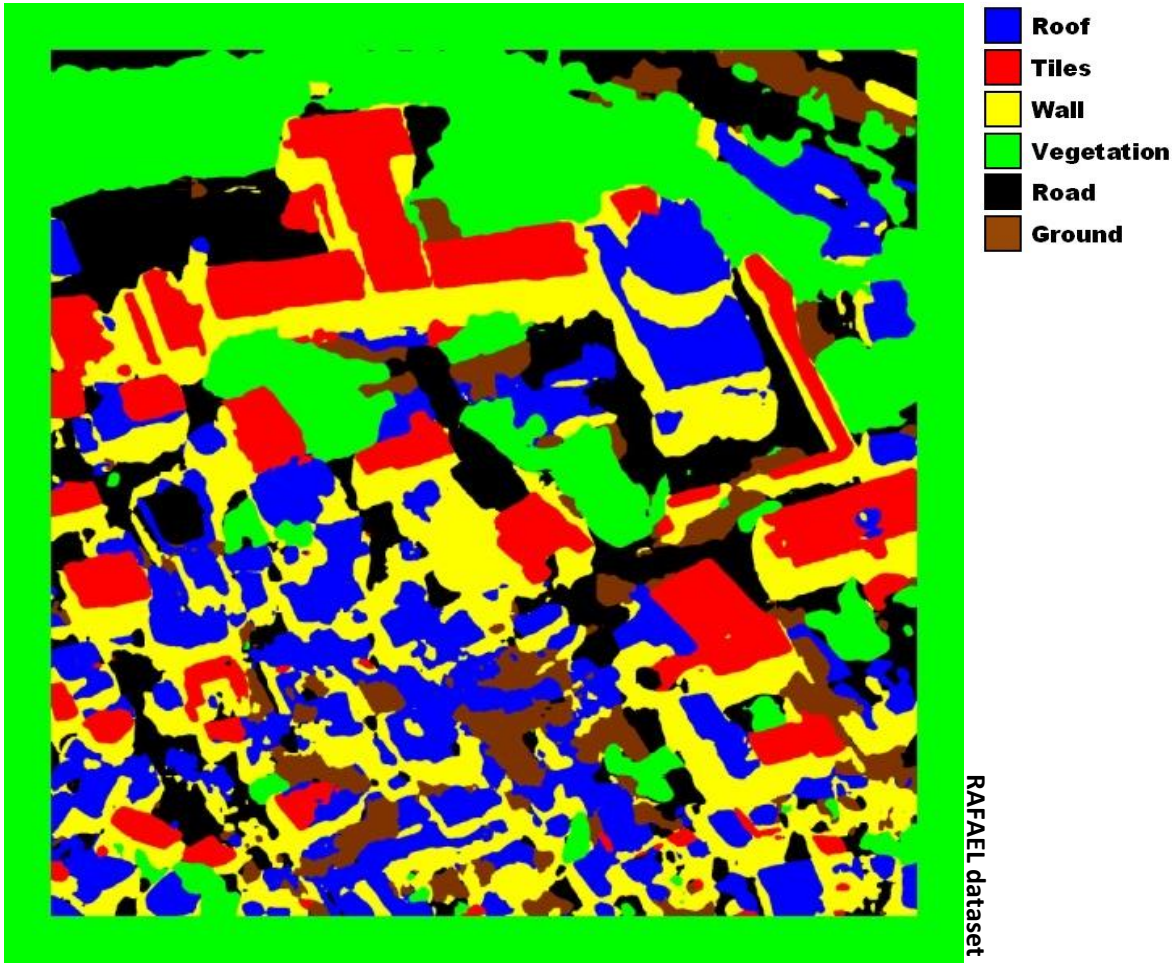
RAFAEL dataset



RAFAEL dataset

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

Questions