Lifting GIS Maps into Strong Geometric Context for Scene Understanding

UCIRVINE

Raúl Díaz, Minhaeng Lee, Jochen Schubert, Charless C. Fowlkes
School of Information and Computer Sciences
The Henry Samueli School of Engineering
Sponsored by The Balsells Fellowship Program

1. Introduction

Contextual information can have a substantial impact on the performance of Computer Vision. Geographic Information Systems (GIS) offers a rich source of contextual information that has been largely untapped by researchers. We propose to leverage such information for scene understanding by combining GIS with large sets of unorganized photographs.

We present a pipeline to generate strong 3D priors from 2D GIS data with minimal user input. Given a test image, we generate robust predictions of depth, semantic labels, and surface normals. We demonstrate the utility of these constraints for re-scoring pedestrian detection and improving semantic segmentation.

2. Model Construction

- Generate a 3D static point cloud from internet photo collections (Structure from Motion).
- Register it with 2D GIS data with minimal user input using procrustes alignment.
- Lift 3D geosemantic model by using custom Sketchup plugin.

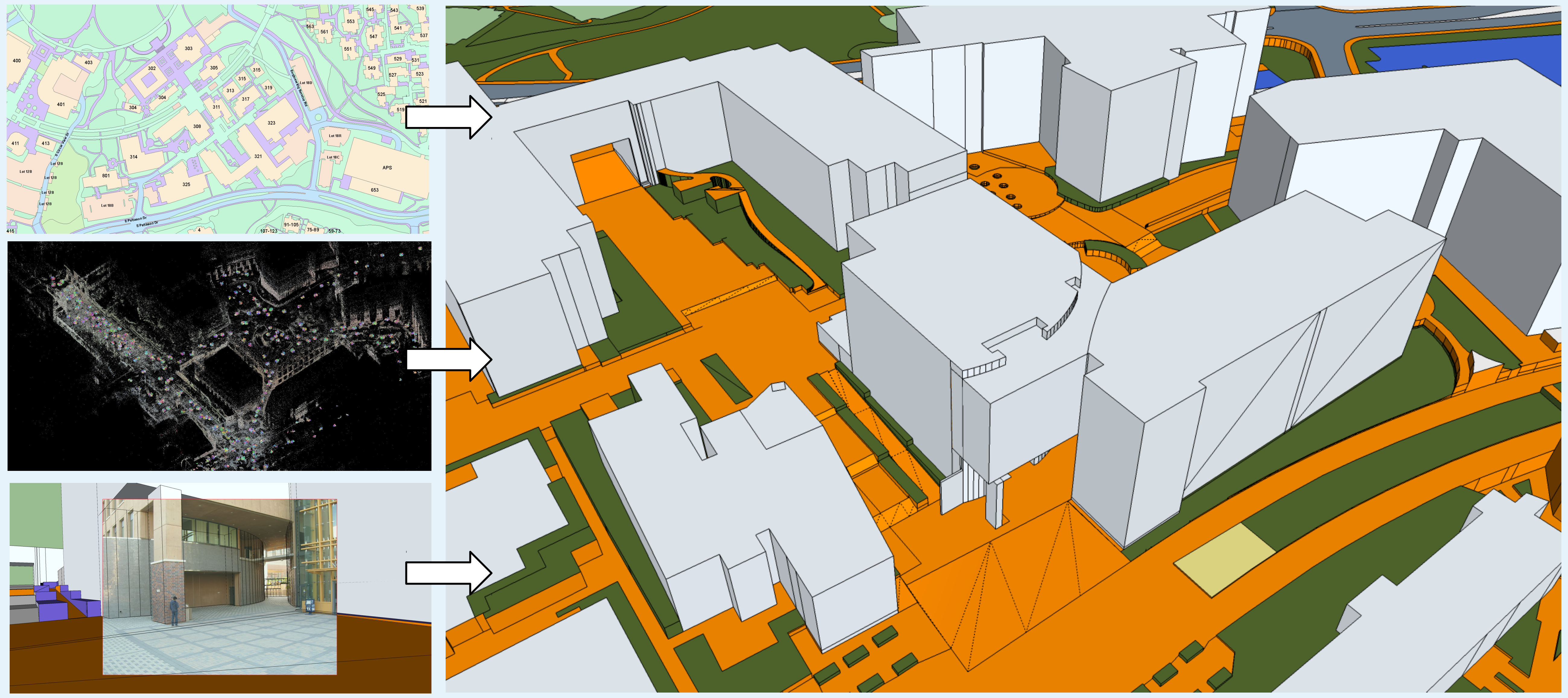
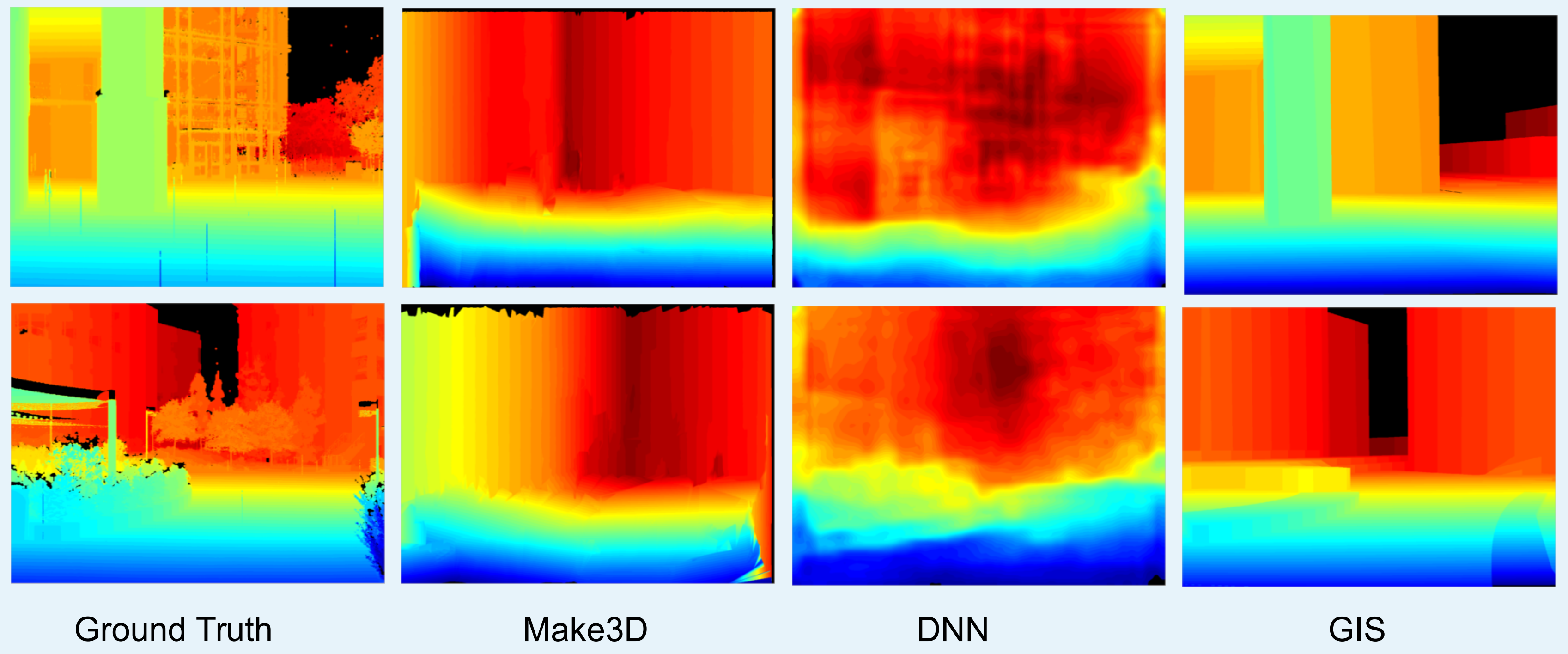


Figure 1: 2D GIS data and SfM reconstruction are shown in top left and center left respectively. Bottom left shows the Sketchup user interface to extrude GIS data into rich 3D models by a natural image overlay. Right shows the resulting 3D GIS model used to evaluate our scene understanding framework.

3. Depth estimation by image resection.

- By simply resectioning a test image in the model, we can estimate depth with much more precision without training complex models or collecting and labelling big amounts of data!



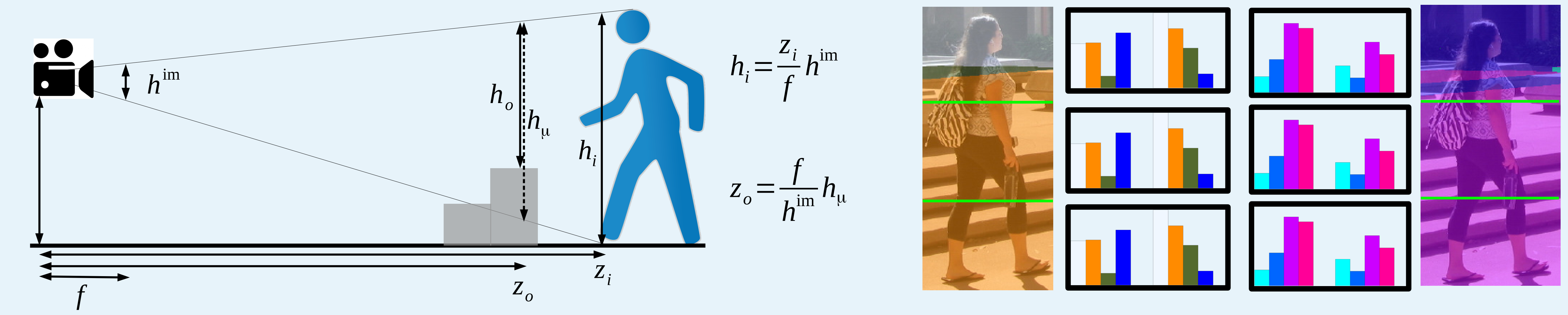
Threshold	Make3D	DNN	GIS
$\delta < 1.25$	11.03%	28.21%	67.04%
$\delta < 1.25^2$	30.71%	45.09%	82.37%
$\delta < 1.25^3$	49.47%	56.04%	88.25%

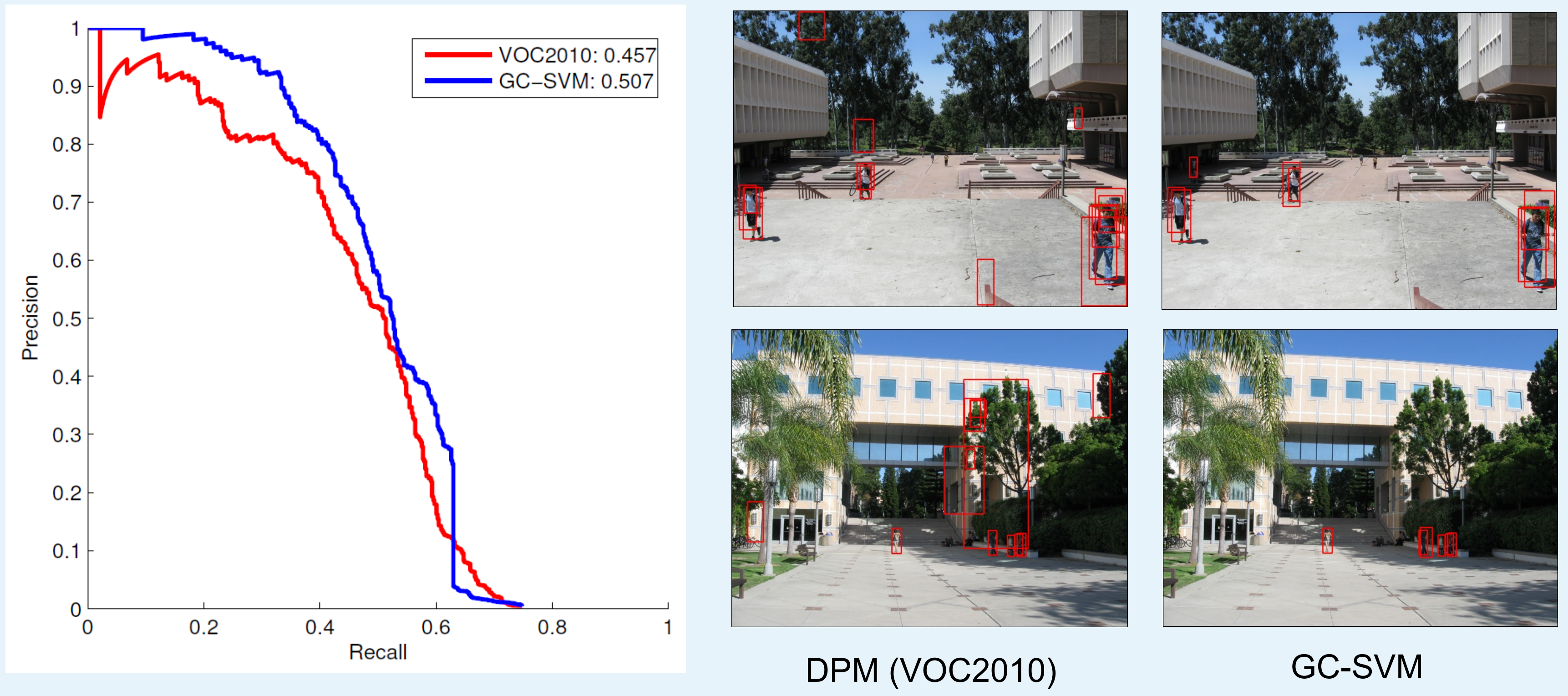
Figure 2: Qualitative and quantitative comparison of depth estimation methods. While our GIS model lacks fine detail such as foliage and pedestrians, simply backprojecting it provides more accurate estimations than existing monocular approaches. The relative error is defined as:

$$\delta = max\left(\frac{d_{gt}}{d_{est}}, \frac{d_{est}}{d_{gt}}\right)$$

4. Geometric reasoning for Object Detection

- We hypothesize the object's supporting geometry by intersection with the ground plane and by depth at average human height.
- We encode histograms of the distribution of labels and normals and link them with their corresponding DPM mixture.
- We train an SVM classifier with the collected features alongsise their DPM score (GC-SVM).





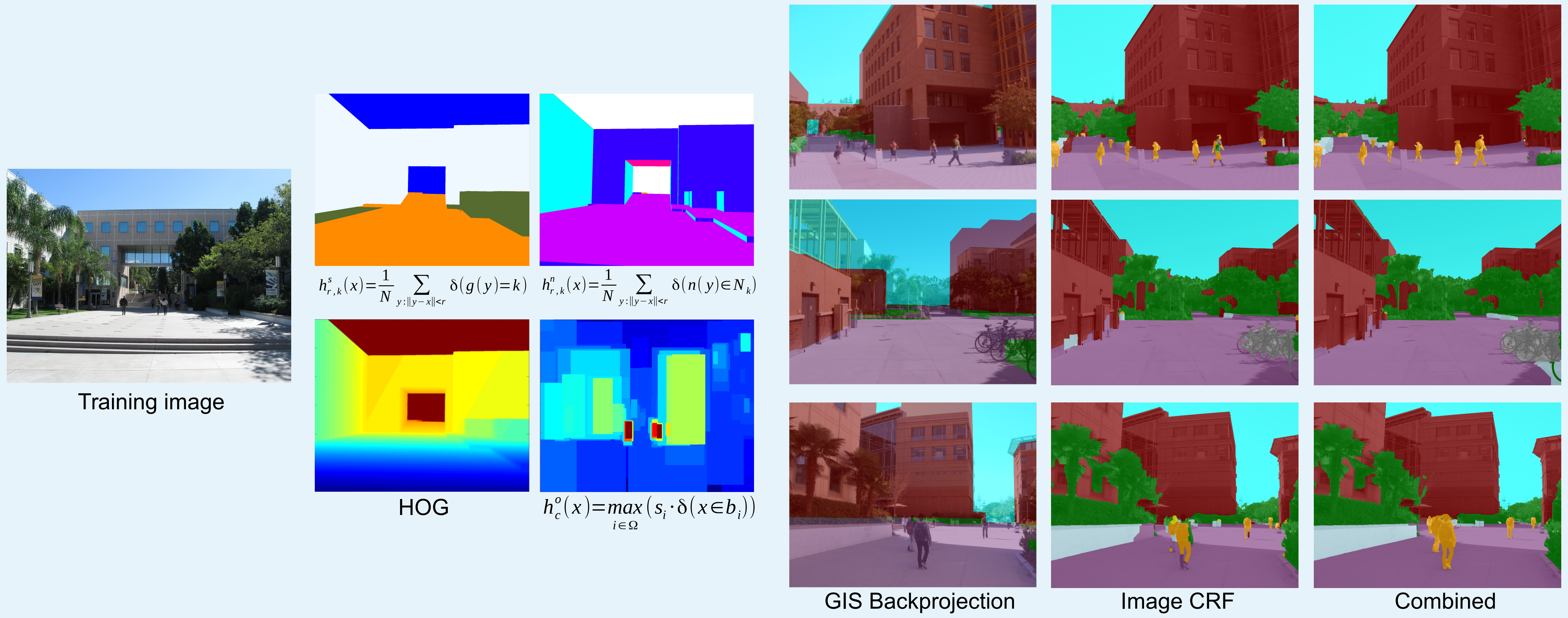
DPM (VOC2010)

GC-SVM

Figure 3: Clockwise from top-left: toy example of our geometric estimation of object's height, either by multiple intersections with the model to avoid blocking objects, or by depth at average human height; semantic label and normal features extracted from each body part, associated to every mixture of our DPM model; examples showing improved detection results at 0.6 recall; precision-recall curves resulting in a 5% boost over baseline at almost any level of recall.

5. Geometric Context for Semantic Segmentation

- We augment the pool of features for semantic segmentation by collecting the distribution of GIS labels and normals per pixel at several angular ratios around each pixel. We add HOG features extracted from the depth maps obtained by model backprojection. We aldo incorporate DPM *heat-maps* to enforce pedestrian and bike segmentation.
- These new feaures expand the pool a standard CRF model, improving baseline results in all categories.


$h^s_{r,k}(x) = \frac{1}{N} \sum_{y:\|y-x\|<r} \delta(g(y)=k)$
$h^n_{r,k}(x) = \frac{1}{N} \sum_{y:\|y-x\|<r} \delta(n(y) \in N_k)$
$h^o_c(x) = \max_{i \in \Omega}(s_i \cdot \delta(x \in b_i))$

Training image

HOG

GIS Backprojection

Image CRF

Combined

Figure 4: Clockwise from top-left: training image with its corresponding feature maps extracted from model backprojection; qualititive segmentation results from model backprojection, baseline CRF, and augmented CRF with GIS features respectively; table of segmentation results with an ablation analysis of the influence of different features. Accuracy is measured using intersection-over-union (IOU) protocol.

Model	Overall	building	plants	pavement	sky	ped.	ped. sit	bicycle	bench	wall
GIS Label Backprojection	0.242	0.688	0.099	0.810	0.581	0.000	0.000	0.000	0.000	0.000
GIS CRF	**0.290**	**0.730**	**0.316**	**0.847**	**0.705**	0.000	0.000	0.000	0.000	**0.014**
Image CRF (ENGQ)	0.561	0.917	0.886	0.925	0.949	0.400	0.010	0.370	0.241	0.348
+GIS	0.584	0.937	0.894	0.936	0.963	0.394	0.060	0.385	0.208	**0.481**
Depth	0.569	0.935	**0.895**	0.937	0.957	0.390	0.011	0.358	0.179	0.455
Labels	0.575	**0.938**	0.892	0.933	**0.966**	0.374	0.064	0.366	0.221	0.419
Normals	0.568	0.935	0.893	0.933	0.961	0.389	0.007	0.385	0.192	0.418
+DPM	0.590	0.920	0.892	0.929	0.947	**0.583**	0.013	0.482	0.184	0.360
+DPM+GIS	**0.627**	0.936	0.894	**0.938**	0.961	0.568	**0.108**	**0.520**	**0.245**	0.472

References

[1] P. Moulon *et al.* Adaptive structure from motion with a contrario model estimation. ACCV, 2012.
[2] S. Gould *et al.* Decomposing a scene into geometric and semantically consistent regions. ICCV, 2009.
[3] Felzenszwalb *et al.* Object detection with discriminatively trained part-based models. TPAMI, 2010

Acknowledgements

This work was supported by NSF IIS-1253538 and a Balsells Fellowship to Raúl Díaz.