Rotated Rectangles for Symbolized Building Footprint Extraction

Matt Dickenson & Lionel Gueguen   Uber Technologies

Overview

Building footprints provide useful visual context for users of digital maps. We demonstrate a method for extracting and symbolizing building footprints on the DeepGlobe Challenge data [1], using a convolutional neural network (CNN) with the following characteristics:

- First six (6) layers are the same as VGG-16
- One convolutional layer for detecting building presence
- One convolutional layer for rotated rectangle parameters

In this way, we are able to directly predict rotated rectangles from imagery. This approach works best in suburban areas of North America, such as Las Vegas.

Rotated Rectangles

Each best-fitting rotated rectangle is described by 5 parameters: its center x, y, its width and height h, w and its angle with respect to a horizontal line α. These parameters are relative to a grid cell characterized by a center x_g, y_g and dimensions h_g, w_g. The angle exists in the range $[-\pi/2, \pi/2]$, and we project it onto the unit circle using its cosine and sine representation. Thus, rotated rectangles are then represented by the following 6 parameters given some grid g:

$$\hat{x} = x - x_g \quad ; \quad \hat{y} = y - y_g, \tag{1}$$

$$\hat{h} = \log h/h_g \quad ; \quad \hat{w} = \log w/w_g, \tag{2}$$

$$\hat{\alpha}_c = \cos 2\alpha \quad ; \quad \hat{\alpha}_s = \sin 2\alpha. \tag{3}$$

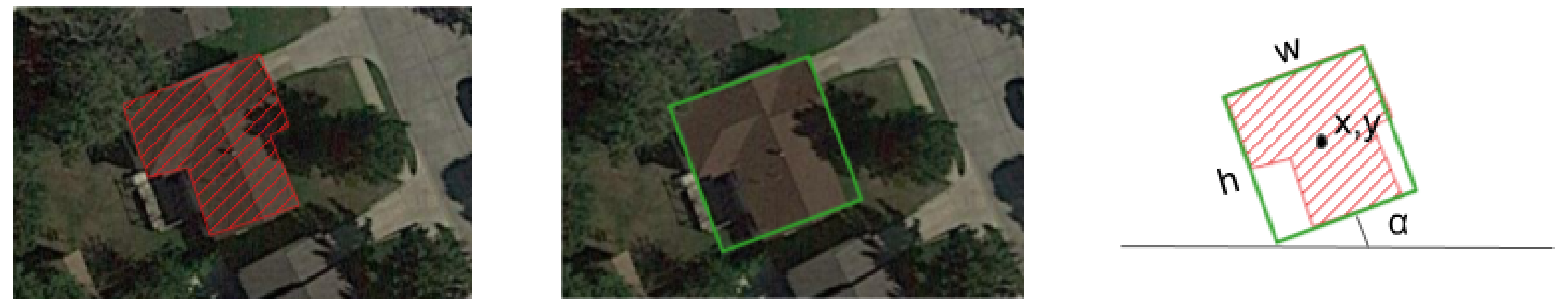

Figure 1: From buildings to rectangles

Architecture

Grid Selection

We design a grid such that at most one building can be predicted by a cell. The grid is characterized as follows so that each cell shall be smaller than the minimum building size. We resized the DeepGlobe Challenge imagery to an input dimension of 512×512. Our network then downscales these dimensions by a factor of $2^3 = 8$ to a grid of dimension of 64×64, so that the resulting grid cell size covers 4×4 m^2.

Non-Maximal Suppression

A non-maximal suppression stage is required to remove overlapping predictions. We use an R-Tree spatial index [2] in order to avoid trivial computations when two polygons do not intersect. Each polygon is associated to a score s_i produced by the network. The algorithm keeps the highest scored polygons, while removing rectangles with substantial overlap. Thanks to the R-Tree spatial index, each rotated rectangle is compared to a small subset of other polygons which potentially intersect it, greatly reducing the computational complexity.

Overall Architecture

We select the first three blocks (six layers) from VGG-16 [3] as the convolutional stack for our network, which leads to a downscaling of the input image dimensions by a factor of 8. One of the final layers predicts the presence of a building in a cell. The other output layer predicts the rotated rectangle parameters if present. We use mean-squared loss to optimize both layers.

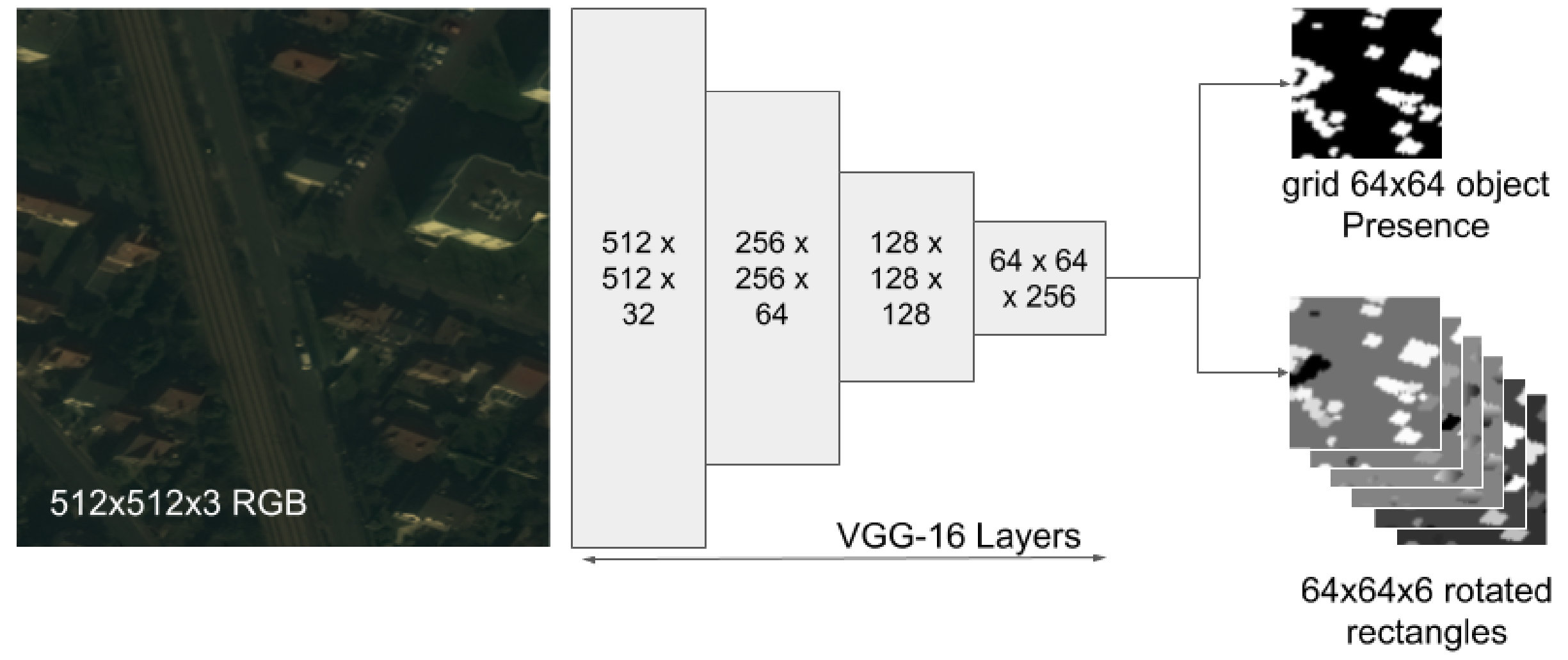

Figure 2: CNN architecture

Results 1

We employed our approach in an experiment on the DeepGlobe building detection challenge [1], encompassing four AOIs covering the cities of Las Vegas in the U.S., Paris in France, Khartoum in Sudan and Shanghai in China. We trained city-specific weights by fine-tuning from a model trained on five U.S. cities.

@IOU 0.5	Precision	Recall	F1
Las Vegas	0.753	0.593	0.664
Paris	0.333	0.258	0.291
Khartoum	0.243	0.161	0.194
Shanghai	0.125	0.082	0.099
Total	0.498	0.365	0.431

Table 1: DeepGlobe test set results

The proposed approach provides good approximations of small and well-separated buildings which are dominant in U.S. cities. However, when buildings are close to each other, of larger size, or when they have non-rectangular shapes the proposed approach does not allow us to capture them with a reasonable IOU, explaining lower F1 scores in Khartoum and Shanghai.

@IOU 0.5	Precision	Recall	F1
Las Vegas	0.760	0.601	0.671
Paris	0.323	0.257	0.286
Khartoum	0.253	0.167	0.201
Shanghai	0.132	0.084	0.103
Total	0.500	0.364	0.432

Table 2: DeepGlobe validation set results

Results 2

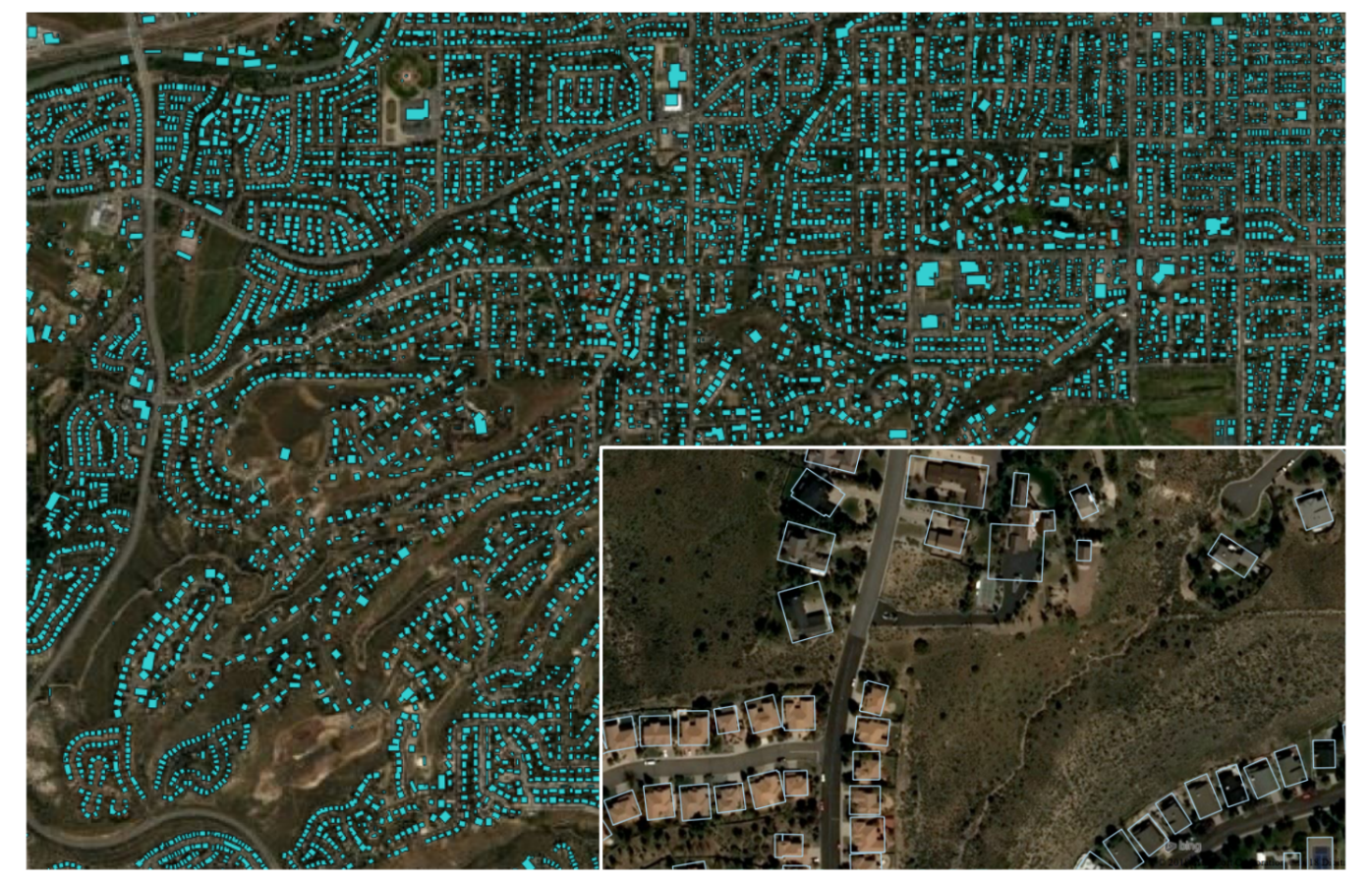

Figure 3: Illustration of the proposed extraction and symbolization method in a suburban area of Reno, NV, U.S.A.

Results 3

In addition to the quantitative results in Tables 1 and 2, Figure 4 shows that our method performs best on somewhat large, well-separated buildings such as individual houses in a suburban area.

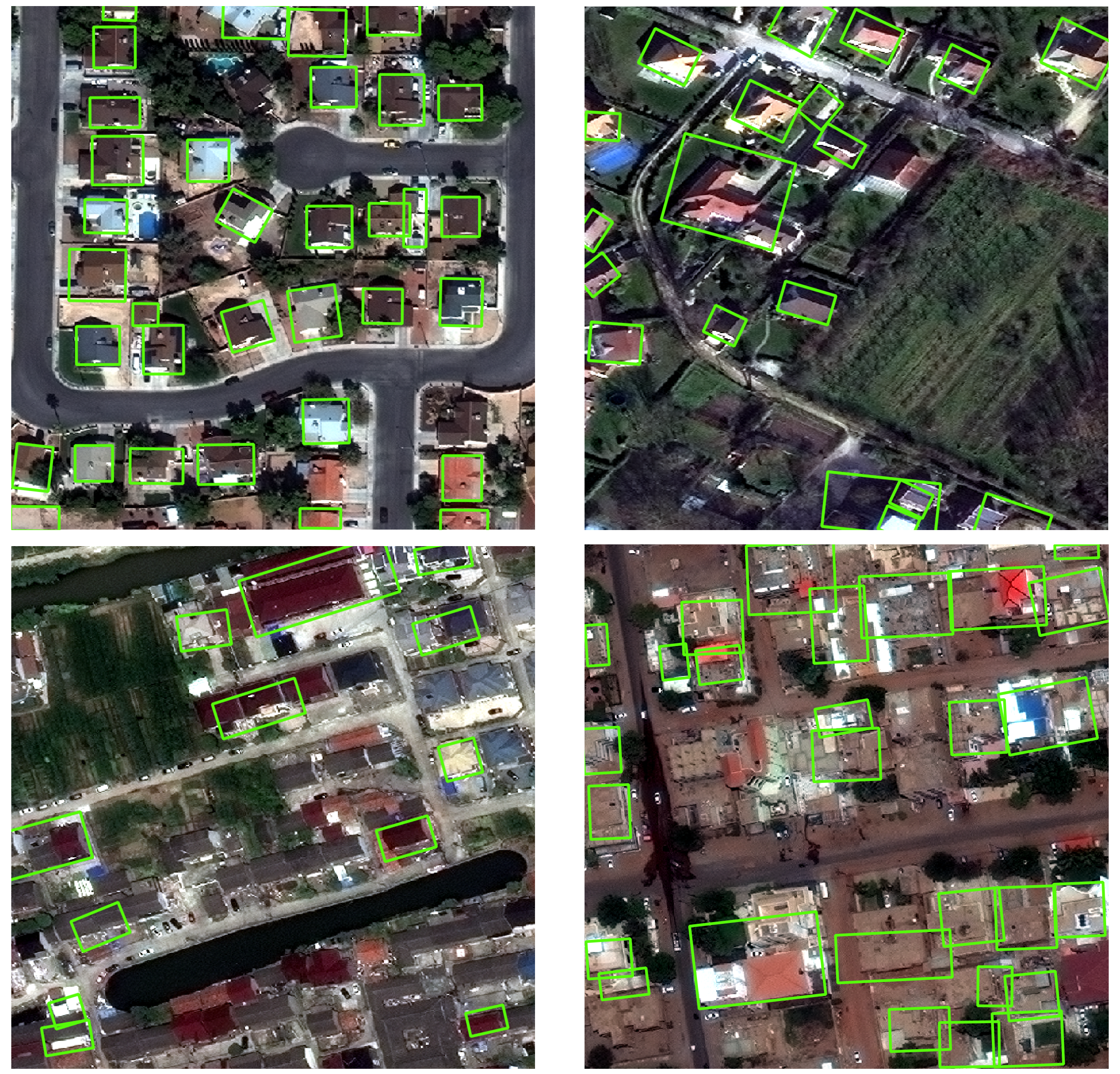

Figure 4: Results from the DeepGlobe test dataset. Top row: Las Vegas and Paris. Bottom row: Shanghai and Khartoum.

Conclusion

- Our CNN architecture outputs rotated rectangles providing a symbolized approximation for small buildings.
- Experimental results show that this method performs best on suburbs consisting of individual houses.
- Large buildings or buildings without clear delineation produce weaker results in terms of precision and recall.

Future Research

Given the difficulties that our method encountered with symbolizing large buildings, we propose that in future work our approach could be combined with segmentation-based architectures known to achieve better F1 scores [4].

References

[1] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018.

[2] Yannis Manolopoulos, Alexandros Nanopoulos, Apostolos N Papadopoulos, and Yannis Theodoridis. *R-trees: Theory and Applications*. Springer, 2010.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *CVPR*, 2017.