PolyWorld: Polygonal Building Extraction with Graph Neural Networks in Satellite Images

Supplementary Material

In the following pages, we present additional qualitative examples of PolyWorld applied to the CrowdAI test dataset [3]. In particular, we show a larger comparison with the state-of-the-art Frame Field Learning (FFL) approach [1], additional results on challenging scenarios and on randomly sampled images from the CrowdAI test set, as well as failure cases of our approach. Moreover, we show an ablation study to evaluate the individual components of our method.

Qualitative results

A qualitative comparison between PolyWorld and the Frame Field Learning (FFL) method are shown in Figure 4. The images represent the results of the two different polygon extraction approaches on complex scenes selected from the CrowdAI test set. Overall, PolyWorld utilizes a lower amount of vertices compared to FFL, generating more regular contours. Results of our approach on challenging scenarios are shown in Figure 5. PolyWorld demonstrates to generalize well on complex and unusual building shapes, managing to detect and connect precisely all the building corners also in presence of severe occlusions. The vertex connections and the final polygon quality is noticeable even on buildings having curved walls as illustrated in the images in the bottom row of Figure 5. In Figure 6 additional PolyWorld polygonizations on randomly sampled images of the test set are shown. It is worth noting that some of the polygon predictions do not seem to be aligned with the building boundaries, especially on tall buildings. This is caused by the fact that many images of the CrowdAI dataset are off nadir, but the annotations are aligned to the base of the buildings. The vertex detection and selection procedure is shown on a sampled CrowdAI test image in Figure 1.

Failure cases

Even though PolyWorld experimentally proves to generate a reliable set of vertices and strong connections, it is interesting to show some failure cases caused by the optimal connection network. In Figure 2 we visualize three examples of wrong vertex matches, resulting from the linear sum assignment problem. Red points describe vertices assigned to the diagonal of the permutation matrix and therefore are filtered, while valid vertices and connections are coloured in green and cyan, respectively. On the left image two corners of the top-left building are assigned to the right building, generating an evident artifact. In the right image the network discards some false-negative corners and does not complete the building footprint. Another artifact is generated by wrongly connecting two building corners to a false-positive vertex as shown in the bottom image. These artifacts are very rare and therefore their impact in the segmentation performance is limited.

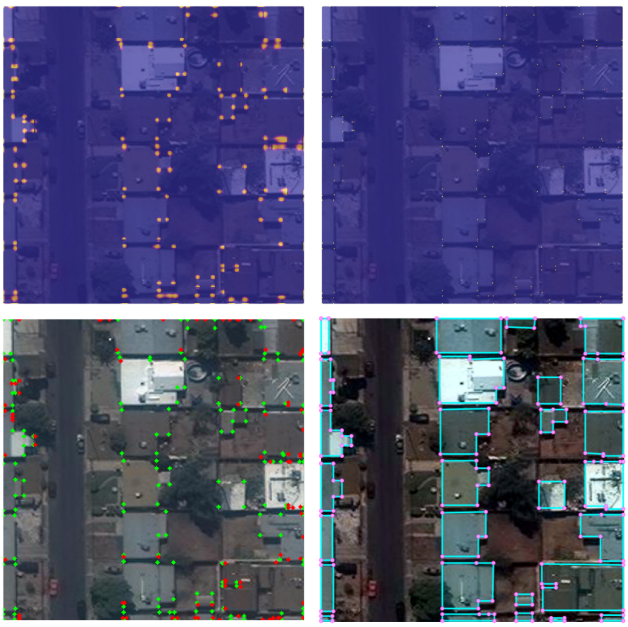


Figure 1. **Vertex detection heatmap.** Top-left: Probability map generated by the Vertex Detection Network. Top-right: probability map after Non Maximum Suppression (please zoom in for better view). Bottom-left: top-256 highest peaks. Green points indicate valid vertices, while red points indicate discarded vertices. Bottom-right: final result.

Ablation study

We conduct additional experiments to evaluate the performance contribution provided by different components of PolyWorld. In particular:

- The model is evaluated discarding the offsets that refine the position of the vertices.
- The model is evaluated only using S_{clock} or $S_{count}^{\top}$ as score matrix S, rather than the ensemble of the two.
- The model is retrained without using the GNN. Removing the GNN automatically means discarding the vertex offsets and the global aggregation of descrip-

PolyWorld	**offset**	**score matrix** S	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	AR_{50}	AR_{75}	AR_S	AR_M	AR_L
full method	off	$S_{clock} + S^{\top}_{count}$	58.7	86.9	64.5	31.8	80.1	85.9	71.7	92.6	79.9	47.4	85.7	94.0
full method	on	$S_{clock} + S^{\top}_{count}$	63.3	88.6	70.5	37.2	83.6	87.7	75.4	93.5	83.1	52.5	88.7	95.2
full method	on	S_{clock}	62.1	87.2	69.2	36.0	83.2	78.9	75.3	93.5	83.0	52.6	88.6	92.4
full method	on	$S^{\top}_{count}$	60.3	84.5	66.8	36.4	80.3	56.6	72.5	89.9	79.9	50.4	85.6	86.3
no GNN	off (n/a)	$S_{clock} + S^{\top}_{count}$	56.8	85.5	62.9	30.7	78.0	80.1	70.2	92.0	78.6	46.2	84.1	92.3
no $\mathcal{L}_{angle}$	on	$S_{clock} + S^{\top}_{count}$	63.6	88.5	70.6	37.7	83.9	88.1	75.9	93.7	83.6	53.2	89.1	95.6

Table 1. Ablation study. MS COCO [2] results on the CrowdAI test computed for different configurations of PolyWorld.

PolyWorld	**offset**	**score matrix** S	IoU	C-IoU	MTA
full method	off	$S_{clock} + S^{\top}_{count}$	89.9	86.9	35.0°
full method	on	$S_{clock} + S^{\top}_{count}$	91.3	88.2	32.9°
full method	on	S_{clock}	90.9	88.1	33.0°
full method	on	$S^{\top}_{count}$	88.4	84.7	33.0°
no GNN	off (n/a)	$S_{clock} + S^{\top}_{count}$	89.2	86.3	35.3°
no $\mathcal{L}_{angle}$	on	$S_{clock} + S^{\top}_{count}$	91.4	88.6	34.0°

Table 2. Ablation study. *Intersection over union (IoU)*, *mean tangent angle error (MTA)*, and *complexity aware IoU (C-IoU)* results on the test-set of the CrowdAI dataset [3] computed for different configurations of PolyWorld.

tors. In this case, the visual descriptors **d** are directly used for matching.

- The model is retrained discarding the angle loss $\mathcal{L}_{angle}$. Only the segmentation loss $\mathcal{L}_{seg}$ is kept to learn the offsets.

Quantitative results of the ablation study are reported in Table 1 and Table 2.

Discarding the refinement offsets results in a noticeable drop in detection and segmentation performance. Moreover, the polygons visually appear not as regular as the full PolyWorld results, as shown in Figure 3b.

Equivalent or even higher detection scores than the full PolyWorld method are achieved retraining the model without angle loss. Unfortunately, in this configuration PolyWorld is not encouraged to generate sharp building corners and the visual impact of the resulting building shapes is not as good as the polygons produced by the method trained with $\mathcal{L}_{angle}$, as shown in Figure 3c. This phenomenon is also explained by the fact that this configuration does not perform as well as the full method is terms of MTA (see Table 2).

Without GNN, PolyWorld still manages to make meaningful vertex connections even though it is not rare to encounter missing footprints or wrong matches, as shown in Figure 3d.

The quantitative results using $S = S_{clock}$ and $S = S^{\top}_{count}$ are reported in Table 1 and Table 2. As expected, only using S_{clock} or $S^{\top}_{count}$ leads to lower segmentation scores compared to the combination of the two: $S = S_{clock} + S^{\top}_{count}$.

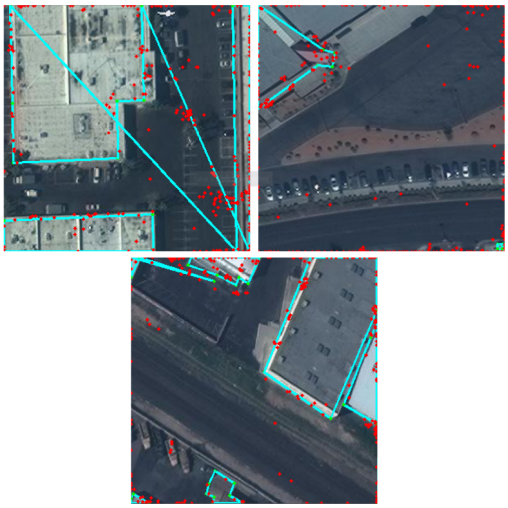


Figure 2. **Examples of wrong connections.** Green points indicate valid vertices. Red points indicate discarded vertices. Generated connections are shown in cyan.

Runtimes

The Frame Field Learning [1] paper reports a computation time of 0.04s on CrowdAI using a GTX 1080Ti. PolyWorld achieves a comparable computation time, taking 0.047s per image with the same configuration (or 0.024s on a GTX 3090).

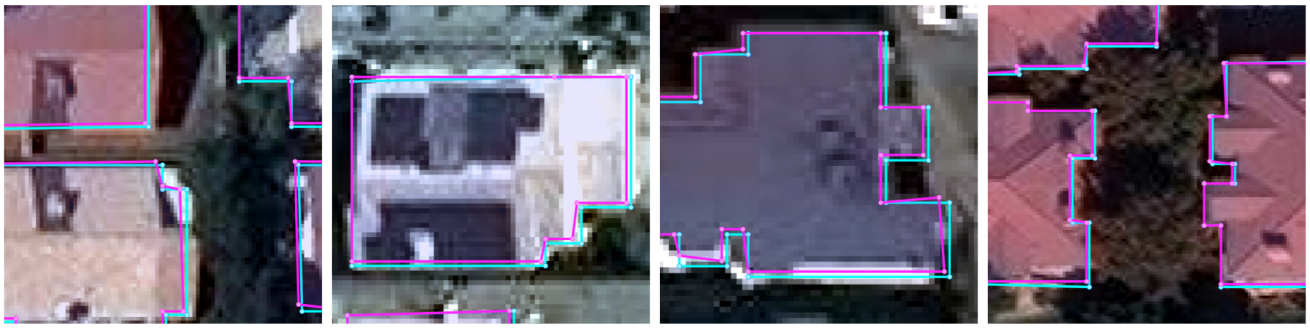
References

[1] Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka. Polygonal building extraction by frame field learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2021. 1, 2, 4

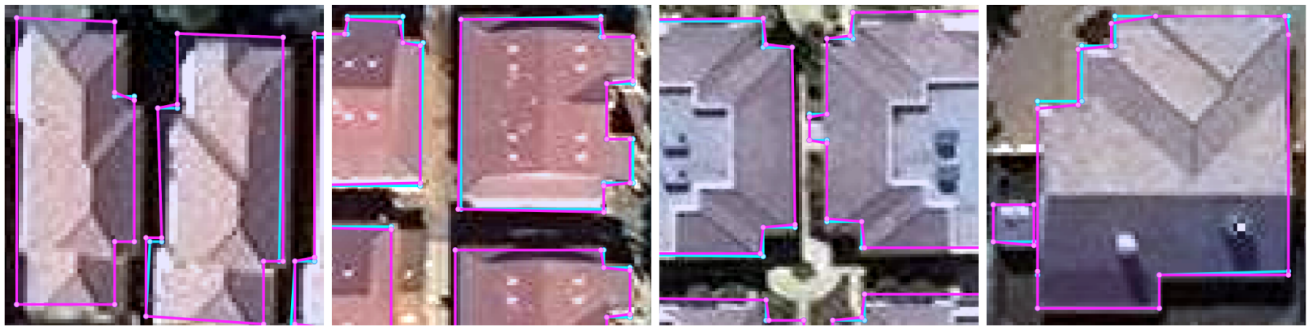
[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[3] S. P. Mohanty. *CrowdAI mapping challenge 2018 dataset*, 2019 (accessed November 10, 2019). https://www.crowdai.org/challenges/mapping-challenge. 1, 2

(a) **Magenta**: Ground truth polygons. **Cyan**: Results of the full PolyWorld method.



(b) **Magenta**: PolyWorld results obtained discarding the refinement offsets. **Cyan**: Results of the full PolyWorld method.



(c) **Magenta**: PolyWorld results obtained discarding the angle loss $\mathcal{L}_{angle}$. **Cyan**: Results of the full PolyWorld method.

(d) **Magenta**: PolyWorld results obtained discarding the Graph Neural Network. **Cyan**: Results of the full PolyWorld method.

Figure 3. Ablation study: qualitative results obtained with different configurations of PolyWorld.

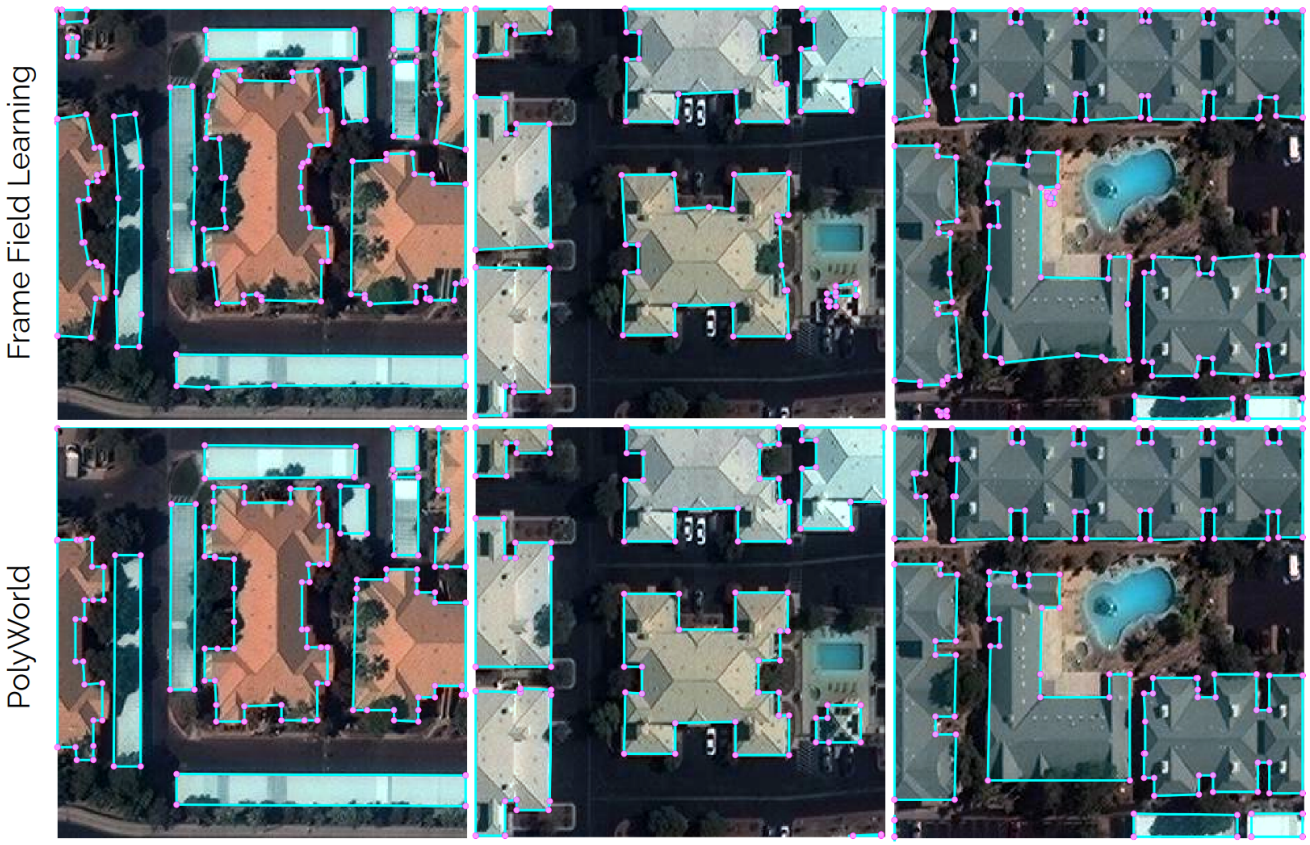


Figure 4. Examples of building extraction and polygonization on CrowdAI test dataset. Top row: Frame Field Learning (FFL) approach [1] with Res101-UNet as backbone and ACM polygonization. Bottom row: PolyWorld results.

Figure 5. Results of PolyWorld on **challenging images** from the CrowdAI test dataset. First row: unusual and complex buildings. Second row: buildings with corners occluded by vegetation. Third row: buildings with curved walls.

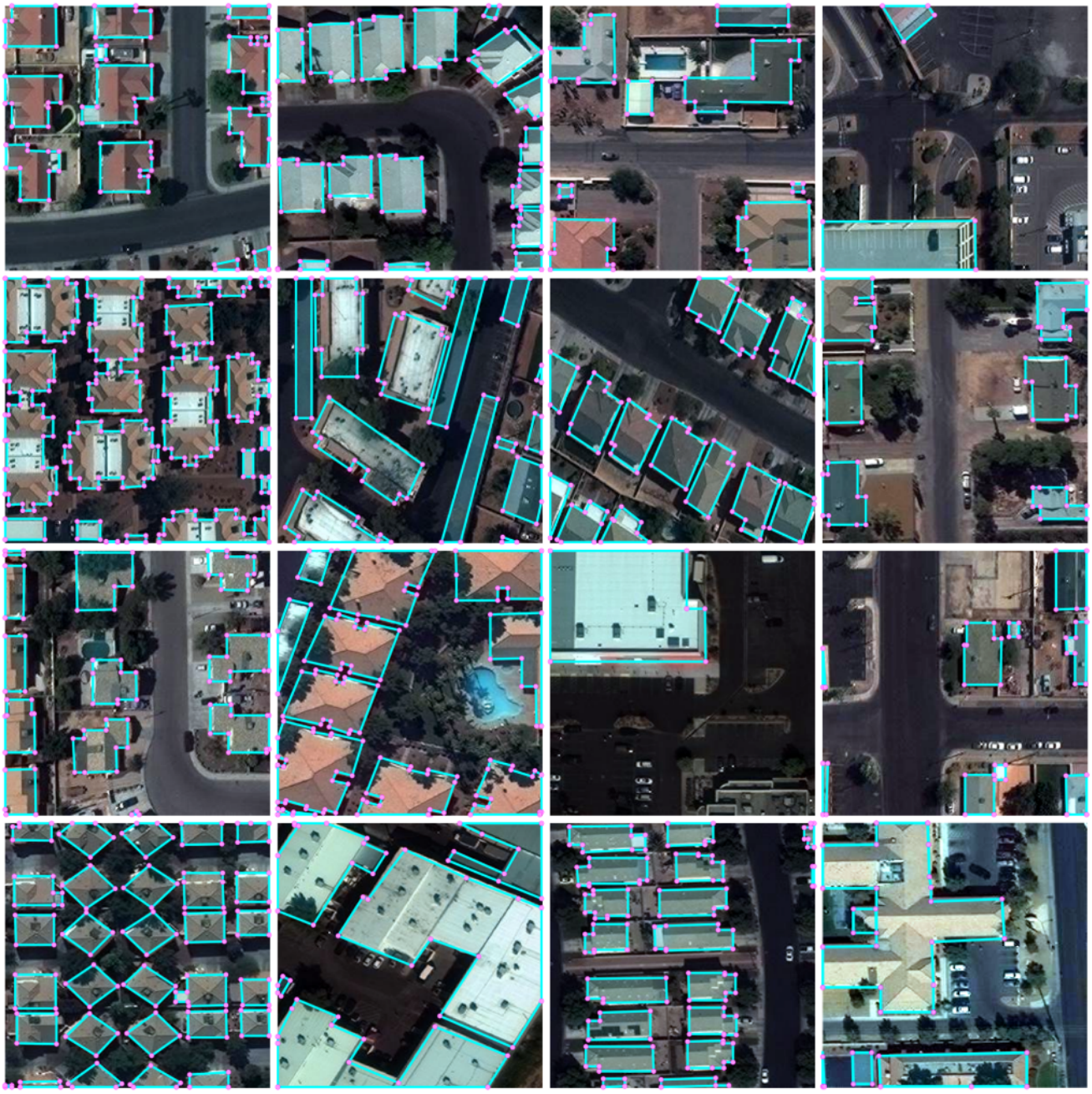


Figure 6. Results of PolyWorld on **randomly sampled images** from the CrowdAI test dataset.