POLYGONIZER: AN AUTO-REGRESSIVE BUILDING DELINEATOR

Maxim Khomiakov[1,2], **Michael Riis Andersen**[1] **& Jes Frellsen**[1]
[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark
[2] Otovo AS
{maxk,miri,jefr}@dtu.dk

ABSTRACT

In geospatial planning, it is often essential to represent objects in a vectorized format, as this format easily translates to downstream tasks such as web development, graphics, or design. While these problems are frequently addressed using semantic segmentation, which requires additional post-processing to vectorize objects in a non-trivial way, we present an Image-to-Sequence model that allows for direct shape inference and is ready for vector-based workflows out of the box. We demonstrate the model's performance in various ways, including perturbations to the image input that correspond to variations or artifacts commonly encountered in remote sensing applications. Our model outperforms prior works when using ground truth bounding boxes (one object per image), achieving the lowest maximum tangent angle error.

1 INTRODUCTION

The application of deep learning in the surveying and analysis of objects has experienced considerable advancements. Alongside the progress of general computer vision methods in classification and object detection, rapid strides have been made in the task of building delineation, which involves accurately separating building objects from the background in remote sensing imagery. However, a persisting challenge concerns the learning of realistic geometric shapes of buildings. A prevalent and intuitive initial approach is to classify the pixels associated with the object of interest using semantic segmentation. The advantages of this method include the dense, fine-grained representation achieved through pixel-by-pixel classification, while its limitations encompass high computational cost and, more significantly, the necessity for non-trivial post-processing of the predicted object masks (Zorzi et al., 2020; Bittner et al., 2018). Post-processing is essential since semantic segmentation often displays the highest uncertainty around object edges. Addressing such pixels naively with a technique like convex-hull could result in distorted objects, as even minor softening of a right-angled corner may introduce substantial artifacts upon conversion. Consequently, our work is driven by the desire to solve the building footprint delineation problem using a direct learned approach, without any post-processing.

1.1 RELATED WORKS

A significant body of research has been published on this topic, with the majority of studies employing semantic segmentation as a fundamental component of their models. The existing literature can be classified into three categories: The first category comprises traditional computer vision approaches that utilize geometric priors and optimization algorithms to detect and polygonize buildings (Li et al., 2020; Bauchet & Lafarge, 2018). The second category encompasses studies that employ deep learning with semantic segmentation, combined with either heuristic or learned post-processing techniques (Lin et al., 2015; Alidoost et al., 2019; Girard et al., 2020; Yuan, 2016; Bittner et al., 2018; Zorzi et al., 2020; Zhao et al., 2020; Zorzi et al., 2022). For instance, Frame Field Learning (Girard et al., 2020) aims to learn effective regularization of object boundaries for precise edge definition, while PolyWorld (Zorzi et al., 2022) addresses the problem using a GNN model, parameterized through an adjacency matrix consisting of building corners. The third category involves research that seek to model polygons directly, thereby eliminating the necessity for post-processing

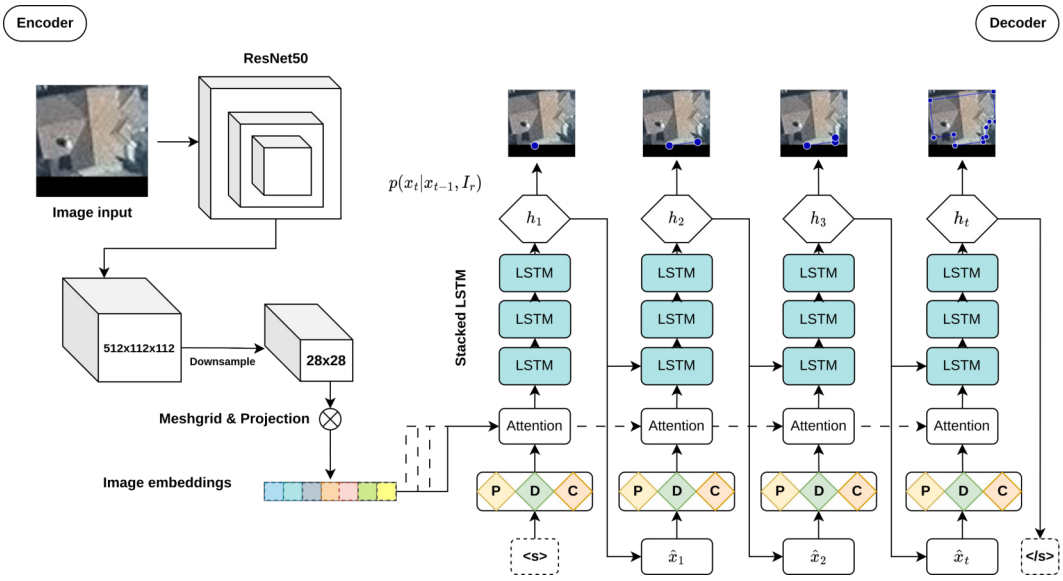

Figure 1: The encoder consists of a modified ResNet50, where the final layers are skipped, upsampled, and concatenated. Positional encoding is added to the resulting feature map, which is then fed into the decoder at each time step. The decoder processes the token at time step t, incorporating position, dimension, and coordinate embeddings along with special tokens. These special tokens include a starting token, denoted as <s>, and a stopping token, denoted as </s>.

(Acuna et al., 2018; Li et al., 2019; Zhao et al., 2021). Although each approach has its limitations, we argue that the benefits of minimal to no assumptions in post-processing render it the most promising direction for future research.

2 METHODS

We present our proposed model in Figure 1 above. Polygonizer consists of an Encoder and a Decoder, with the Encoder being a modified ResNet50, designed similarly to Acuna et al. (2018); Hu et al. (2022), but incorporating learned embeddings for the coordinate token, coordinate dimension, and coordinate position. In addition, we introduce a small fixed value to every feature map in the 512x28x28 output generated by the encoder, which assists the network in learning spatial dependencies, resulting in our encoded feature representation I_r. The encoded image is then supplied as initial input, alongside a special starting token <s>, to a stacked LSTM Decoder, which learns to output parameters to a categorical distribution spanning the dimensions of the image $p(x_t|x_{t-1}, I_r)$. At each timestep, the LSTM employs Bahdanau attention Bahdanau et al. (2014) between the encoder representation I_r and the prior hidden state of the Decoder c_{t-1}, aiming to align the LSTM representation with the image representation at every time step. The Decoder is trained with teacher forcing and performs inference until the special </s> token is encountered. While bearing similarities to the model by Li et al. (2019); Acuna et al. (2018), our model relies solely on the preceding token and an image embedding to learn the sequence, and utilizes an LSTM instead of the ConvLSTM employed by Li et al. (2019); Acuna et al. (2018).

2.1 EXPERIMENTAL SETUP

Our model is trained on the Aicrowd mapping challenge dataset (Mohanty et al., 2020) using ground truth bounding boxes. All cropped objects were padded on the smallest edge to maintain an identical aspect ratio. Our embeddings, as well as the encoder and decoder hidden dimensions, are set to 512, while the decoder is a 3-layer stacked LSTM. We optimize our model using the Adam optimizer with a negative log likelihood loss and a learning rate of $2 \cdot 10^{-4}$. The results shown in this paper were all obtained using greedy inference and an identical seed.

3 RESULTS

Upon completing our research, we became aware of concurrent work by Hu et al. (2022). As we lacked access to their model weights at the time of writing, we compared our approach with two other recent state-of-the-art methods: PolyWorld and Frame-Field Learning (Zorzi et al., 2022; Girard et al., 2020). We present the COCO-evaluation metrics for an aggregate performance overview in Table 2. Although our model performs reasonably well, it was trained on ground truth bounding boxes, simplifying the learning task. However, metrics like the maximum tangent angle error, one of our model's key advantages, should be less affected by this. The results in Figure 2 and Table 2 show general alignment between the methods, except for the image examples in the fourth and fifth columns. Notably, our method generally outperforms FFL and PolyWorld when using ground truth bounding boxes. However, considering the study by Hu et al. (2022), Polygonizer ranks second or third in most metrics while demonstrating a substantial improvement over the alternative methods in having the lowest maximum tangent angle error.

Figure 2: Qualitative comparison of results. From the top to the bottom: Ground truth, Frame Field Learning, PolyWorld and Polygonizer (ours).

Method	**AP** ↑	**AP** 50 ↑	**AP** 75 ↑	**AR** ↑	**AR** 50 ↑	**AR** 75 ↑	**IoU** ↑	**MTA** ↓	**N ratio**	**C-IoU** ↑
PolyMapper (Li et al., 2019)	55.7	86	65.1	62.1	88.6	71.4	-	-	-	-
FFL (Girard et al., 2020)	60.9	87.4	70.4	64.5	89.2	73.4	84.4	33.5	1.13	74
PolyWorld (Zorzi et al., 2022)	63.3	88.6	70.5	75.4	93.5	83.1	91.3	32.9	0.93	88.2
W. Li et al. (Li et al., 2021)	73.8	92	81.9	72.6	90.5	80.7	-	-	-	-
PolyBuilding (Hu et al., 2022)	**78.7**	**96.3**	**89.2**	**84.2**	**97.3**	**92.9**	**94.0**	32.4	0.99	88.6
Polygonizer (ours)	71.9	94.2	81.4	82.3	92.9	90.9	89.8	**10.50**	1	87.9

Table 1: Our model employs ground truth bounding boxes, which introduces an element of disparity in the comparison.

3.1 A STUDY OF ROBUSTNESS

Although our arguments supporting the benefits of our approach may not be definitive, we aim to investigate the extent to which comparable models can yield satisfactory results under adversarial conditions. We have devised three perturbations of the input image, which we consider to be realistic sources of error in the remote sensing domain. We evaluate the models under conditions of masking pixels (Erased dropout Zhong et al. 2020), downsampling the input image to simulate lower ground sampling distance, and rotating the input image in 15-degree increments. The results from these experiments are presented in Table 2 and in Appendix B. Our method demonstrates increased robustness towards downsampling and, to a lesser extent, dropout phenomena, although its performance, as expected, declines as the downsampling becomes more severe. Interestingly, PolyWorld overestimates the number of points relatively early. Both methods exhibit a significant drop in performance as the downsampling of the input image increases.

	AP ↑	AR ↑	IoU ↑	N ratio	C-IoU ↑	PF
PolyWorld	0.539	0.671	0.783	1.680	0.715	2
	0.106	0.256	0.498	1.143	0.344	4
	0.003	0.012	0.081	0.311	0.013	8
	AP ↑	**AR ↑**	**IoU ↑**	**N ratio**	**C-IoU ↑**	**PF**
FFL	0.531	0.640	0.781	0.890	0.713	2
	0.186	0.327	0.599	0.952	0.489	4
	0.003	0.030	0.288	0.552	0.137	8
	AP ↑	**AR ↑**	**IoU ↑**	**N ratio**	**C-IoU ↑**	**PF**
LSTM	0.683	0.797	0.880	0.963	0.860	2
	0.650	0.774	0.864	0.930	0.841	4
	0.523	0.686	0.801	0.877	0.761	8

	AP ↑	AR ↑	IoU ↑	N ratio	C-IoU ↑	PF
PolyWorld	0.448	0.592	0.747	1.657	0.667	4
	0.384	0.532	0.719	1.609	0.627	6
	0.291	0.444	0.664	1.517	0.553	8
	AP ↑	**AR ↑**	**IoU ↑**	**N ratio**	**C-IoU ↑**	**PF**
FFL	0.443	0.565	0.739	0.966	0.678	4
	0.373	0.501	0.701	0.983	0.631	6
	0.295	0.432	0.665	0.990	0.584	8
	AP ↑	**AR ↑**	**IoU ↑**	**N ratio**	**C-IoU ↑**	**PF**
LSTM	0.579	0.725	0.844	1.086	0.816	4
	0.552	0.706	0.831	1.086	0.801	6
	0.532	0.688	0.823	1.095	0.792	8

Table 2: Performance relative to input pertubations. Left: Pertubation where a value of 2 equates a downsampling of 2x. Right: Erased dropout performance where the fraction of pixels dropped corresponds to PF$\cdot 0.02\%$

4 DISCUSSION AND FUTURE WORKS

In this paper, we introduce a new auto-regressive method for building delineation that retains the performance of similar works (Acuna et al., 2018) with reduced complexity. Our method does not necessitate a separate model to predict the first vertex and relies solely on the prior token to predict the subsequent one, unlike previous work (Zhao et al., 2021; Li et al., 2019). However, our model has its limitations. Primarily, it depends on having just one object per scene and is limited in its ability to learn long sequences. Nonetheless, the model is quite adaptable to learning right-angled geometry, which is a crucial property for remote sensing applications. Future research will focus on expanding our model to automatically detect objects in the image and introducing inductive biases that may further enhance the model's robustness.

5 CONCLUDING REMARKS

We have proposed a novel and simplified method for polygonizing buildings from remote sensing imagery. This approach proves to be remarkably effective at learning the angles between vertices surrounding the building while also being accurate at completing the sequence. We believe this is partially explained by our fixed and learned embeddings, which, in conjunction with Bahdanau attention (Bahdanau et al., 2014) at every timestep, facilitate learning the sequential structure of the data.

REFERENCES

David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 859–868, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00096.

F. Alidoost, H. Arefi, and F. Tombari. Building outline extraction from aerial images using convolutional neural networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(4/W18):57–61, 2019. ISSN 16821750. doi: 10.5194/isprs-archives-XLII-4-W18-57-2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Jean-Philippe Bauchet and Florent Lafarge. KIPPI: KInetic Polygonal Partitioning of Images. Technical report, 2018. URL https://hal.inria.fr/hal-01740958v2.

Ksenia Bittner, Fathalrahman Adam, Shiyong Cui, Marco Körner, and Peter Reinartz. Building Footprint Extraction From VHR Remote Sensing Images Combined With Normalized DSMs Using Fused Fully Convolutional Networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2615–2629, 2018. ISSN 21511535. doi: 10.1109/JSTARS.2018.2849363.

Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka. Polygonal Building Segmentation by Frame Field Learning. pp. 1–30, 2020. URL http://arxiv.org/abs/2004.14875.

Yuan Hu, Zhibin Wang, Zhou Huang, and Yu Liu. Polybuilding: Polygon transformer for end-to-end building extraction. *arXiv preprint arXiv:2211.01589*, 2022.

Muxingzi Li, Florent Lafarge, and Renaud Marlet. Approximating shapes in images with low-complexity polygons. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8630–8638, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00866.

Weijia Li, Wenqian Zhao, Huaping Zhong, Conghui He, and Dahua Lin. Joint semantic-geometric learning for polygonal building segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1958–1965, 2021.

Zuoyue Li, Jan DIrk Wegner, and Aurelien Lucchi. Topological map extraction from overhead images. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob: 1715–1724, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00180.

Jingbo Lin, Weipeng Jing, Houbing Song, and Guangsheng Chen. ESFNet : Efficient Network for Building Extraction from High-Resolution Aerial Images. 14(8):1–9, 2015.

Sharada Prasanna Mohanty, Jakub Czakon, Kamil A Kaczmarek, Andrzej Pyskir, Piotr Tarasiewicz, Saket Kunwar, Janick Rohrbach, Dave Luo, Manjunath Prasad, Sascha Fleer, et al. Deep learning for understanding satellite imagery: An experimental survey. *Frontiers in Artificial Intelligence*, 3, 2020.

Jiangye Yuan. Automatic Building Extraction in Aerial Scenes Using Convolutional Networks. 2016. URL http://arxiv.org/abs/1602.06564.

W. Zhao, I. Ivanov, C. Persello, and A. Stein. BUILDING OUTLINE DELINEATION: From VERY HIGH RESOLUTION REMOTE SENSING IMAGERY to POLYGONS with AN IMPROVED END-TO-END LEARNING FRAMEWORK. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 43(B2):731–735, 2020. ISSN 16821750. doi: 10.5194/isprs-archives-XLIII-B2-2020-731-2020.

Wufan Zhao, Claudio Persello, and Alfred Stein. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175(September 2020):119–131, 2021. ISSN 09242716. doi: 10.1016/j.isprsjprs.2021.02.014. URL https://doi.org/10.1016/j.isprsjprs.2021.02.014.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Stefano Zorzi, Ksenia Bittner, and Friedrich Fraundorfer. Machine-learned Regularization and Polygonization of Building Segmentation Masks. 2020. URL http://arxiv.org/abs/2007.12587.

Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. Polyworld: Polygonal building extraction with graph neural networks in satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1848–1857, 2022.

A Appendix

B Experimental pertubations

Additional tables with experiments relating to rotation.

AP ↑	AR ↑	IoU ↑	N ratio	C-IoU ↑	Rotation
0.000	0.000	0.121	1.795	0.113	90
0.000	0.000	0.085	0.861	0.052	60
0.000	0.000	0.072	0.865	0.043	120
0.000	0.001	0.095	0.821	0.056	45
0.000	0.008	0.246	1.047	0.172	15

Table 3: PolyWorld (Zorzi et al., 2022) performance as a function of rotation.

AP ↑	AR ↑	IoU ↑	N ratio	C-IoU ↑	Rotation
0.003	0.034	0.336	0.831	0.311	15
0.000	0.005	0.136	0.829	0.122	60
0.000	0.002	0.115	0.825	0.103	120
0.000	0.003	0.132	0.893	0.123	90
0.000	0.006	0.151	0.820	0.134	45

Table 4: Frame Field Learning Girard et al. (2020) performance as a function of rotation.

AP ↑	AR ↑	IoU ↑	N ratio	C-IoU ↑	Rotation
0.371	0.510	0.762	1.252	0.726	15
0.193	0.357	0.656	1.247	0.597	45
0.167	0.325	0.637	1.260	0.578	60
0.164	0.324	0.632	1.253	0.580	90
0.158	0.319	0.625	1.259	0.570	120

Table 5: Polygonizer (ours) performance as a function of rotation.

B.1 ADDITIONAL RESULTS

Figure 3: Qualitative comparison of results. From the top to the bottom: Ground truth, Frame Field Learning, PolyWorld and Polygonizer (ours).

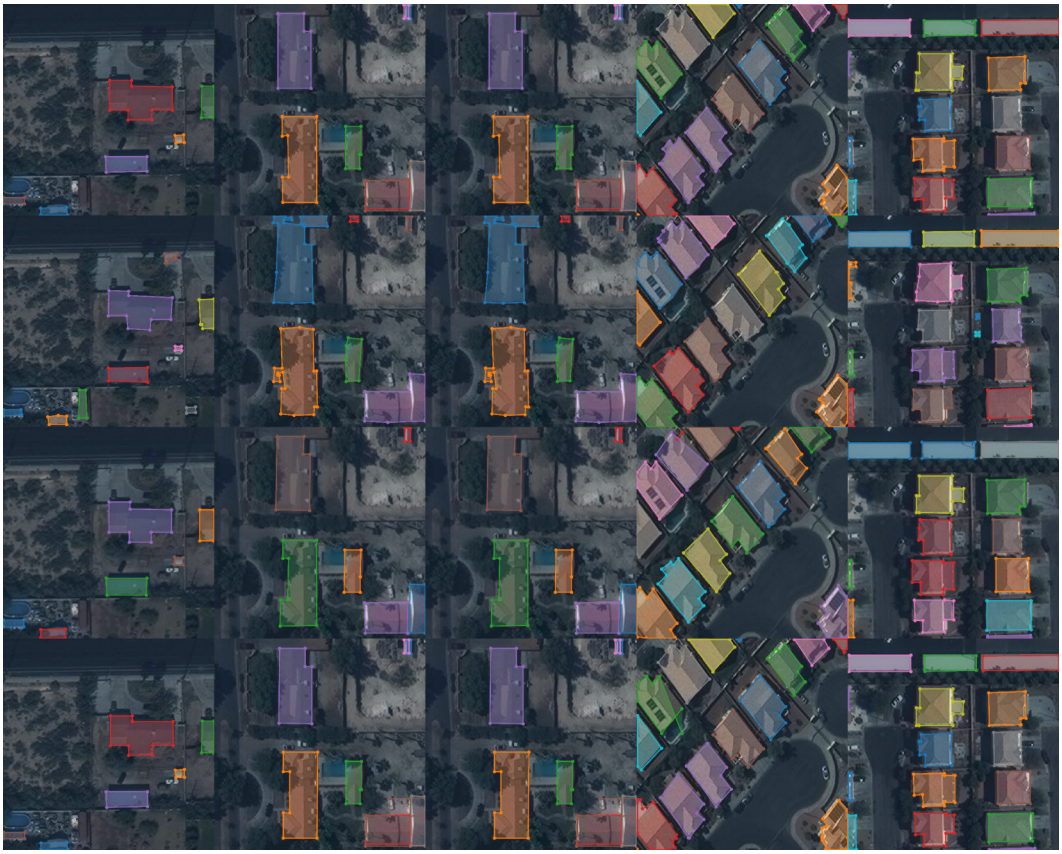

Figure 4: Qualitative comparison of results. From the top to the bottom: Ground truth, Frame Field Learning, PolyWorld and Polygonizer (ours).

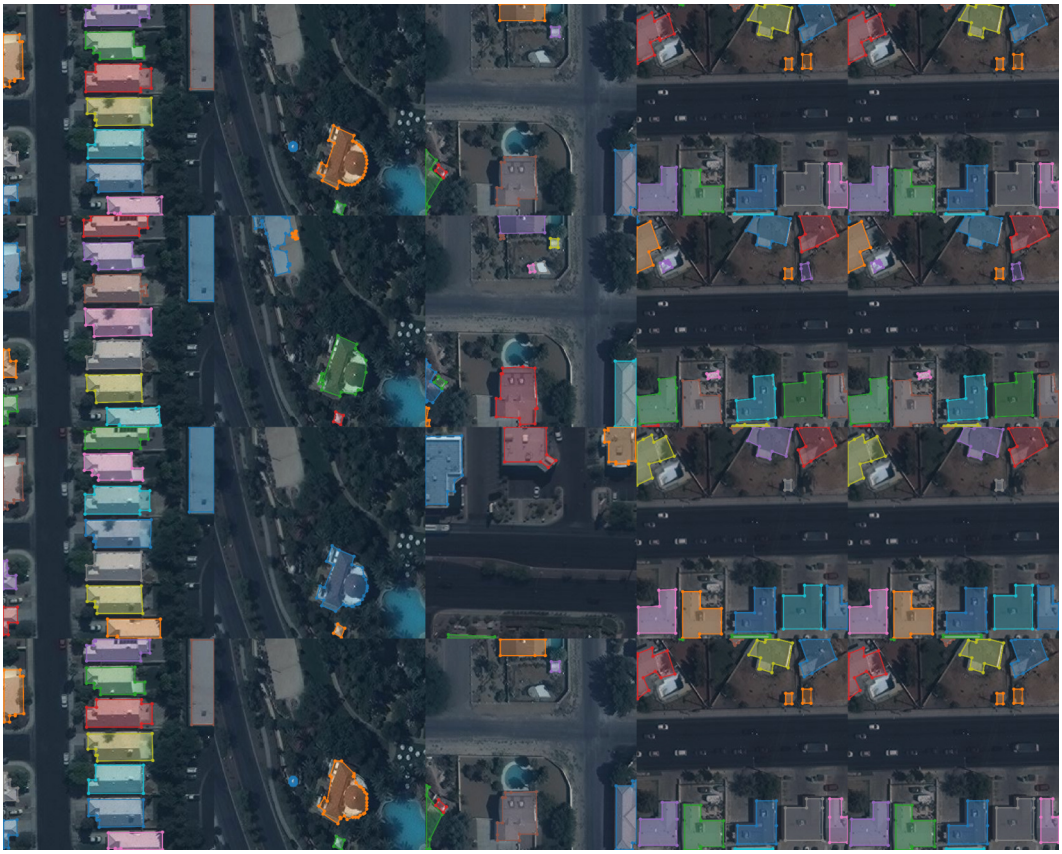

Figure 5: Qualitative comparison of results. From the top to the bottom: Ground truth, Frame Field Learning, PolyWorld and Polygonizer (ours).