Pix2Poly: A Sequence Prediction Method for End-to-end Polygonal Building Footprint Extraction

Yeshwanth Kumar Adimoolam
CYENS CoE, Cyprus
y.adimoolam@cyens.org.cy

Charalambos Poullis
Concordia University
charalambos@poullis.org

Melinos Averkiou
CYENS CoE, Cyprus
m.averkiou@cyens.org.cy

Supplementary Material

In this supplementary material, we provide additional implementation, training, and inference details about our pipeline in Sec. 1. We include ablations for the vertex detection network of the Pix2Poly architecture in Sec. 2. We report quantitative comparisons on the AICrowd Mapping Challenge dataset [11] in Sec. 3. We also report quantitative comparisons on the Massachusetts Roads dataset [10], INRIA (170) dataset [9] and the AICrowd Mapping Challenge small validation set [11] in Sec. 4 using the evaluation script provided by the authors of TopDiG [17]. We also demonstrate the failure cases of Pix2Poly in Sec. 5. Finally, we report additional quantitative results and qualitative examples of polygonal building footprints and road networks predicted by Pix2Poly on all datasets in Sec. 6.

1. Miscellaneous training and inference details

1.1. Implementation Details

All images were resized to 224×224 before being passed to the network. We use the small variant of the standard vision transformer, ViT [2], with a patch size of 8 as the backbone in all of our experiments. The input image is divided into 8×8 patches and the latent dimension of each patch was set to 256. For the decoder, we employ a transformer with 6 decoder layers and 8 heads per layer. Also, all GT sequence tokens (start, end, pad, and vertex tokens) are embedded using a learnable linear embedding layer. For the optimal matching network, we employ two MLPs for predicting clockwise and counter-clockwise permutation matrices. During training, we compute the permutation matrix from the raw logits predicted by the optimal matching network using 100 Sinkhorn iterations. During inference, we compute the exact assignment matrix from the logit values using the Hungarian algorithm. Based on our analysis, the maximum number of building corners (N_v) in an image is set to 192 for both the INRIA [9] and Spacenet datasets [3]. N_v was set to 256 for the AICrowd dataset [11], 144 for the WHU Buildings [5] dataset, and 96 for the Massachusetts Roads dataset [10]. We employ the AdamW optimizer [8] with a learning rate of 4×10^{-4} and weight decay of 1×10^{-4}. We use weights $\lambda_s = 1.0$ and $\lambda_p = 10.0$ for the losses. In our experiments, a single forward pass on an NVIDIA RTX A5000 GPU and an AMD EPYC 7313 processor takes $\sim 18.2ms$ per image.

1.2. Training Details

Augmentations: Our training setup uses extensive geometric and radiometric augmentations to ensure high-quality polygon prediction. For radiometric augmentations, we use random brightness/contrast adjustments, color jittering, RGB shifts, grayscale conversions, and the addition of Gaussian noise. We also use extensive random rotations with a probability of 0.8. The combination of these augmentations provided the best results among which the rotation augmentations provided the most increase in evaluation performance.

End-to-end Gradient Flow: The Vertex Sequence Detector of Pix2Poly predicts a sequence of vertex tokens as class probabilities over the vocabulary defined by the tokenizer. Therefore, to ensure end-to-end gradient flow between the predicted vertex sequence and the subsequent optimal matching step, we directly pass the penultimate decoder vertex features (before applying softmax) to the Optimal Matching Network. These vertex features of shape $N_v \times d$ are self-repeated to construct a self-attention matrix of shape $N_v \times N_v \times d$, which in turn is passed to the Optimal Matching Network to predict the binary permutation matrix of shape $N_v \times N_v$.

Here, N_v is the maximum number of vertices per image and d is the feature dimension in the transformer decoder of the Vertex Sequence Detector.

Ordering of Predicted Vertices: In direct polygon prediction methods such as Polyworld [21] and TopDiG [17], the implicit ordering of the ground truth vertices is lost during the non-differentiable non-max suppression step. To overcome this, Polyworld [21] employs a vertex sorting step

to restore the ground truth vertex ordering so that the predicted array of vertices is in correspondence with the predicted permutation matrix. TopDiG [17] on the other hand, generates GT permutation matrics on-the-fly so that they are in correspondence with the predicted vertex array.

In contrast, Pix2Poly does not need any such intermediate step since there is no loss in the implicit ordering of the vertices. Due to the end-to-end gradient flow in Pix2Poly, the vertex sequence detector learns to predict the vertices in the right order as imposed by the ground truth sequence of vertices. This helps in reducing the overhead introduced by any intermediate sorting steps.

Handling Pad Tokens: Pad tokens are treated as self-connected vertices during the optimal matching step. Therefore, following [21], rows corresponding to pad vertices are assigned '1' in the binary GT permutation matrix diagonal during training and discarded as self-connected vertices during inference.

1.3. Inference Details

Patched Inference: Since Pix2Poly is trained with backbones with a fixed input size, we adopt a patched inference strategy for aerial image tiles spanning a much larger area on the ground. We patch large aerial images (eg. 5000×5000 tiles of the test split of INRIA(155) [9] dataset) into 224×224 patches with a 10% overlap with adjacent patches. These patches are passed as inputs to Pix2Poly and the resulting building footprint polygons are translated to their corresponding locations in the 5000×5000 tile. The redundant polygons in the overlapping regions are simply merged with a unary union operation. In the case of buildings with inner yards (i.e., polygons with holes), we treat overlapping polygons that are entirely contained within a larger polygon as an inner hole of that polygon. This strategy is followed by competing methods as well [14, 17, 18, 21, 22]. Besides this patching strategy, we do not perform any test-time augmentations such as rotations, flip, crops, etc. for the patch predictions that are commonly adopted by competing methods. All polygons in a patch are obtained in a single pass autoregressively. Due to Pix2Poly's accurate predictions, we can observe strong consistency of polygon predictions in the overlap regions resulting in high-quality building polygons for large aerial image tiles as shown in Figs. 3 to 7.

2. Vertex Sequence Detector Ablations

Since the sequence detection approach for vertex detection is our primary contribution, we ablate the proposed Vertex Sequence Detector to demonstrate its effectiveness in generating highly accurate building polygons without the need for the computationally expensive regularization losses, differentiable rasterizer, topology concentrated node detectors in competing methods [17, 21]. To effectively demonstrate this, we design a baseline that is identical to Pix2Poly except for the vertex detection step. For vertex detection, we replace the sequence decoder of Pix2Poly with the mask decoder approach of PolyWorld [21], TopDiG [17] and UniVecMapper [18]. We predict a vertex heatmap and use a non-differentiable non-max suppression layer to extract the vertex coordinates. We also use the vertex sorting step described in [21] to ensure correspondence with the ground truth permutation matrix. We report the quantitative comparison of this baseline with the proposed Pix2Poly with the vertex sequence detector in Tab. 1, from which it is evident that Pix2Poly can outperform the baseline and generate high-quality building polygon predictions without any complex regularization modules. We further demonstrate this via qualitative comparisons of building polygons between the baseline and Pix2Poly in Fig. 1.

In addition to the ablation for the vertex sequence detector, we observed that the patch size of the backbone vision transformer encoder also had a significant impact on the performance of the Vertex Sequence Detector. Using a smaller patch size in the backbone encoder resulted in significant improvement in performance as shown in Tab. 2. Therefore, we decided to use the ViT Small variant with a patch size of 8 as the encoder backbone.

3. Quantitative Comparison - AICrowd Mapping Challenge Dataset

In this section, we report the quantitative comparisons on the official validation split of the AICrowd Mapping Challenge dataset [11]. Although this dataset is a popular choice for benchmarking building footprint extraction methods [4, 6, 14, 17, 21, 22] we wish to reiterate the numerous issues of data leakage and excessive duplication recently discovered in this dataset [1] and hence decided against including comparisons on this dataset in the main paper. We still report our performance on this dataset in Tabs. 3 and 4 for the sake of complete comparisons.

4. Quantitative Comparisons with TopDiG [17] and UniVecMapper [18]

To compare the performance of the proposed Pix2Poly with TopDiG [17] and UniVecMapper [18], we used the evaluation script provided by the authors of TopDiG. However, we realized that the authors were computing a multi-class confusion matrix and averaging across both the buildings(or roads) and background classes for the mask and topology metrics used in their paper. This deviates from the standard convention of reporting only on building class IoU followed by previous methods [4, 14, 21, 22] and by us in the main paper. Therefore, we removed these results from the main paper and moved them to the supplementary

Method	Desc	IoU ↑	C-IoU ↑	N-Ratio = 1	MTA ↓	PoLiS ↓	IoU^{topo} ↑	$F1^{topo}$	PA^{topo} ↑
Pix2Poly (baseline)	Vertex Segmentation + NMS	80.52	72.89	0.919	34.11°	1.751	58.13	72.39	93.49
Pix2Poly (ours)	Vertex Sequence Detection	**81.81**	**75.05**	**1.041**	**33.40°**	**1.717**	**60.31**	**74.20**	**93.80**

Table 1. **Ablation results for the Vertex Sequence Detector: Polygonal Footprint Quality metrics.** IoU & additional metrics assessing the quality of building footprints extracted from the *Spacenet Vegas dataset's val split*. **Bold** indicates the best scores.

Figure 1. **Qualitative comparisons.** Examples of predicted building polygons from the INRIA test set. We compare between **Pix2Poly (baseline) in the top row** and **Pix2Poly (ours) in the bottom row**. The sequence prediction approach for vertex detection enables Pix2Poly to predict accurate and high-quality building polygons without the use of complex regularization losses, a differentiable rasterizer, and a topology feature learning module employed in competing methods. Zoom in for a better view.

Method	Backbone Patch Size	IoU ↑	C-IoU ↑	MTA ↓	PoLiS ↓
Pix2Poly	16 x 16	71.06	62.79	35.62°	2.695
	8 x 8	**75.06**	**67.27**	**35.24°**	**2.261**

Table 2. **Polygonal Footprint Quality results.** Comparison of IoU and other polygon quality metrics from the ablation experiments for the ViT backbone patch size, performed on the *INRIA dataset's validation split*. **Bold** indicates the better-performing configuration.

in Tab. 5 to avoid ambiguity. We also report the mask and topology scores computed on the building/road class as per the standard convention in *green italics* in Tab. 5.

5. Failure Cases

In Fig. 2, we illustrate some examples of failure cases of Pix2Poly from the Spacenet Vegas dataset's validation split. It can be seen that the following are the most common causes of failure:

- Partially or fully missing buildings in the predictions.
- Incorrect vertex connections learned by the permutation matrix result in polygons with topological errors.
- Misalignment between the ground truth and predicted polygons.

6. Additional Results

In this section, we demonstrate additional quantitative results and qualitative examples of predictions made by Pix2Poly from the various datasets described in the main paper in Tab. 6 and Figs. 3 to 7.

While we compare Pix2Poly with competing methods by training and testing on 224×224 patches of the INRIA(155) dataset, it should be noted that some methods provide their pre-trained checkpoints. In particular, FFL [4] and HiSup [14] provide pre-trained weights for their models after training on 512×512 images of the INRIA(155) dataset. HiT [19], while not providing any code or pre-trained weights, also reports metrics on 512×512 of the INRIA(155) dataset. Therefore, for the sake of complete comparisons, we also evaluate Pix2Poly on 512×512 patches of the INRIA(155) dataset using the patched inference strategy described in Sec. 1.3. The results of these comparisons are reported in Tab. 6.

Method	$AP \uparrow$	$AP_{50} \uparrow$	$AP_{75} \uparrow$	$AP_S \uparrow$	$AP_M \uparrow$	$AP_L \uparrow$	$AR \uparrow$	$AR_S \uparrow$	$AR_M \uparrow$	$AR_L \uparrow$
PolyMapper [6]	55.7	86.0	65.1	30.7	68.5	58.4	62.1	39.4	75.6	75.4
FFL (ACM poly) [4]	61.3	87.4	70.6	33.9	75.1	83.1	64.9	41.2	78.7	85.9
PolyWorld [21]	63.3	88.6	70.5	37.2	83.6	87.7	75.4	52.5	88.7	95.2
BuildMapper [13]	63.9	90.1	75.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Re:PolyWorld [22]	67.2	89.8	75.8	42.9	85.3	89.4	78.6	58.3	90.3	96.2
HiSup [14]	79.4	**92.7**	**85.3**	55.4	**92.0**	**96.5**	81.5	60.1	94.1	**97.8**
Pix2Poly (ours)	**79.6**	91.6	85.2	**61.4**	91.9	91.7	**87.7**	**73.6**	**96.0**	97.5

Table 3. **Quantitative results.** The MS-COCO AP/AR metrics from experiments on the *AICrowd dataset's official validation split containing 60,317 images.* **Bold** and underlined scores indicate best and second-best scores respectively. Pix2Poly matches HiSup [14] on average precision scores and outperforms on average recall scores. From the AP_S and AR_S scores, it is evident that Pix2Poly is significantly better at detecting smaller building objects in the dataset.

Method	IoU ↑	C-IoU ↑	N-Ratio = 1	MTA ↓	PoLiS ↓	$\text{IoU}^{topo} \uparrow$	F1^{topo}	$\text{PA}^{topo} \uparrow$
FFL (ACM poly) [4]	84.10	73.70	n/a	33.5°	3.454	n/a	n/a	n/a
PolyWorld [21]	91.24	88.39	0.945	32.9°	0.962	76.75	86.61	97.04
Re:PolyWorld [22]	92.20	89.70	n/a	31.9°	n/a	n/a	n/a	n/a
HiSup [14]	94.27	89.67	**1.016**	31.9°	0.726	84.08	91.14	98.05
Pix2Poly (ours)	**95.03**	**89.85**	1.111	**23.1°**	**0.479**	**89.05**	**93.75**	**98.62**

Table 4. **Polygonal Footprint Quality metrics.** IoU & additional metrics assessing the quality of building footprints extracted from the *AICrowd dataset's val split of 60,317 images*. **Bold** & underlined scores indicate best & 2^{nd}-best scores respectively.

Dataset	Method	Class	$PA^{mask} \uparrow$	$F1^{mask} \uparrow$	$IoU^{mask} \uparrow$	$PA^{topo} \uparrow$	$F1^{topo} \uparrow$	$IoU^{topo} \uparrow$
Inria (170) [9]	Curve-GCN [7]	Building & background	87.00	84.00	75.00	93.00	62.00	55.00
	DeepSnake [12]		93.00	86.00	79.00	93.00	73.00	64.00
	E2EC [20]		88.46	70.85	63.64	92.69	65.83	58.61
	FFL [4]		92.00	85.00	77.00	92.00	68.00	59.00
	PolyWorld [21]		90.82	83.54	73.41	92.92	73.60	63.47
	BuildMapper [13]		n/a	n/a	63.64	n/a	n/a	58.61
	TopDiG [17]		94.70	91.32	84.56	93.88	78.47	68.39
	UniVecMapper [18]		n/a	n/a	85.15	n/a	n/a	69.77
	Pix2Poly (ours)		**95.78**	**92.39**	**87.33**	**94.35**	**86.51**	**78.58**
	Pix2Poly (ours)	Building only	*95.78*	*87.80*	*80.40*	*94.35*	*76.46*	*63.67*
AICrowd [11]	E2EC [20]	Building & background	95.62	92.11	86.72	93.70	78.67	69.13
	PolyWorld [21]		93.67	90.29	82.89	93.21	77.71	67.43
	TopDiG [17]		96.45	94.77	90.23	94.51	82.20	72.51
	Pix2Poly (ours)		**98.87**	**98.05**	**96.54**	**98.54**	**96.23**	**93.41**
	Pix2Poly (ours)	Building only	*98.87*	*96.92*	*94.65*	*98.54*	*93.30*	*88.49*
Massachusetts Roads [10]	Enhanced-iCurb [16]	Roads & background	-	-	-	89.00	68.00	58.00
	RNGDet++ [15]		-	-	-	n/a	n/a	50.54
	PolyWorld [21]		-	-	-	94.28	76.56	66.59
	TopDiG [17]		-	-	-	95.16	80.33	70.66
	UniVecMapper [18]		-	-	-	n/a	n/a	75.87
	Pix2Poly (ours)		-	-	-	**97.51**	**85.74**	**77.52**
	Pix2Poly (ours)	Roads only	-	-	-	*97.51*	*72.80*	*57.64*

Table 5. **Quantitative results.** Mask and Topology quality metrics reported on the *INRIA (170), AICrowd (small val set), and Massachusetts Roads datasets.* Pix2Poly consistently outperforms SOTA methods on the quality of building and road graphs. **Bold** and underlined scores indicate best and second-best scores respectively. *Green italicized* scores indicate metrics computed on the building/road class using the standard convention.

References

[1] Yeshwanth Kumar Adimoolam, Bodhiswatta Chatterjee, Charalambos (Charis) Poullis, and Melinos Averkiou. Efficient deduplication and leakage detection in large scale image datasets with a focus on the crowdai mapping challenge dataset. *ArXiv*, abs/2304.02296, 2023. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-

Figure 2. **Failure Cases.** Examples of some failure cases of Pix2Poly from the Spacenet Vegas dataset's validation split. The most common causes of failure are partially or fully missing buildings in an image. Incorrect connections between vertices and overlap errors are also occasionally occurring failure cases.

vain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[3] Adam Van Etten, David Lindenbaum, and Todd M. Bacastow. Spacenet: A remote sensing dataset and challenge series. *ArXiv*, abs/1807.01232, 2018. 1

[4] 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 (CVPR)*, pages 5891–5900, June 2021. 2, 3, 4, 6

[5] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open

Dataset	Type	Method	IoU ↑	C-IoU ↑	NR = 1	MTA ↓	PoLiS ↓	IoU^{topo} ↑	F1^{topo}	PA^{topo} ↑
INRIA(155) dataset *val* [9]	Indirect	FFL [4]	75.6	66.0	1.32	35.25°	2.976	42.19	58.76	94.76
		HiSup [14]	74.6	**67.2**	**1.04**	43.86°	3.079	48.87	64.32	93.74
	Direct	HiT [19]	-	64.5	0.8	**33.20°**	-	-	-	-
		Pix2Poly	**77.71**	66.1	1.33	34.81°	**2.296**	**55.45**	**69.85**	**95.00**

Table 6. **Polygonal Footprint Quality metrics.** IoU & additional metrics assessing quality of building footprints predicted by Pix2Poly on the INRIA(155) dataset with 512px × 512px images. FFL [4] and HiSup [14] were evaluated with the corresponding provided pre-trained checkpoints. Pix2Poly was evaluated after training on INRIA(155) 224 × 224 images using the patched inference approach. **Bold** & underlined scores indicate best & 2^{nd}-best scores respectively.

aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019. 1

[6] Zuoyue Li, Jan Dirk Wegner, and Aurelien Lucchi. Topological map extraction from overhead images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1715–1724, Oct 2019. 2, 4

[7] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *CVPR*, 2019. 4

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. 1

[9] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017. 1, 2, 4, 6

[10] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. 1, 4

[11] 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. 1, 2, 4

[12] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *CVPR*, 2020. 4

[13] Shiqing Wei, Tao Zhang, Shunping Ji, Muying Luo, and Jianya Gong. Buildmapper: A fully learnable framework for vectorized building contour extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:87–104, 2023. 4

[14] Bowen Xu, Jiakun Xu, Nan Xue, and Guisong Xia. Accurate polygonal mapping of buildings in satellite imagery. *ArXiv*, abs/2208.00609, 2022. 2, 3, 4, 6

[15] Zhenhua Xu, Yuxuan Liu, Yuxiang Sun, Ming Liu, and Lujia Wang. Rngdet++: Road network graph detection by transformer with instance segmentation and multi-scale features enhancement. *IEEE Robotics and Automation Letters*, 8(5):2991–2998, 2023. 4

[16] Zhenhua Xu, Yuxiang Sun, and Ming Liu. Topo-boundary: A benchmark dataset on topological road-boundary detection using aerial images for autonomous driving. *IEEE Robotics and Automation Letters*, 6(4):7248–7255, 2021. 4

[17] Bingnan Yang, Mi Zhang, Zhan Zhang, Zhili Zhang, and Xiangyun Hu. Topdig: Class-agnostic topological directional graph extraction from remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1265–1274, June 2023. 1, 2, 4

[18] Bingnan Yang, Mi Zhang, Zhili Zhang, Yuanxin Zhao, and Jianya Gong. Univecmapper: A universal model for thematic and multi-class vector graph extraction. *International Journal of Applied Earth Observation and Geoinformation*, 130:103915, 2024. 2, 4

[19] Mingming Zhang, Qingjie Liu, and Yunhong Wang. Hit: Building mapping with hierarchical transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2023. 3, 6

[20] Tao Zhang, Shiqing Wei, and Shunping Ji. E2ec: An end-to-end contour-based method for high-quality high-speed instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4443–4452, June 2022. 4

[21] 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*, pages 1848–1857, 2022. 1, 2, 4

[22] Stefano Zorzi and Friedrich Fraundorfer. Re:polyworld - a graph neural network for polygonal scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16762–16771, October 2023. 2, 4

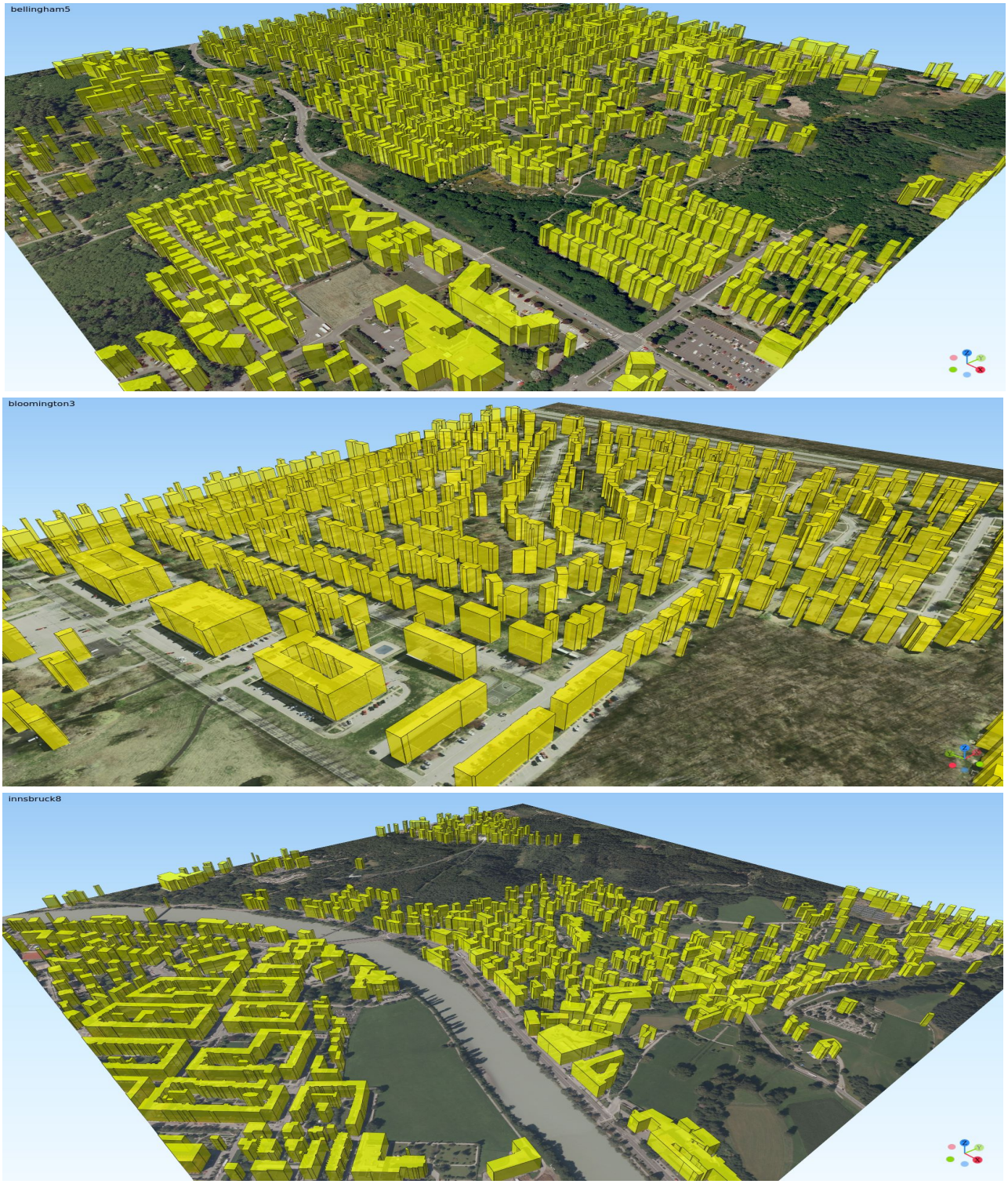

Figure 3. **Qualitative results.** Qualitative examples of extruded building polygons from the INRIA (150) dataset's official test split. Pix2Poly can predict high-quality building footprints that are immediately usable for 3D reconstruction.

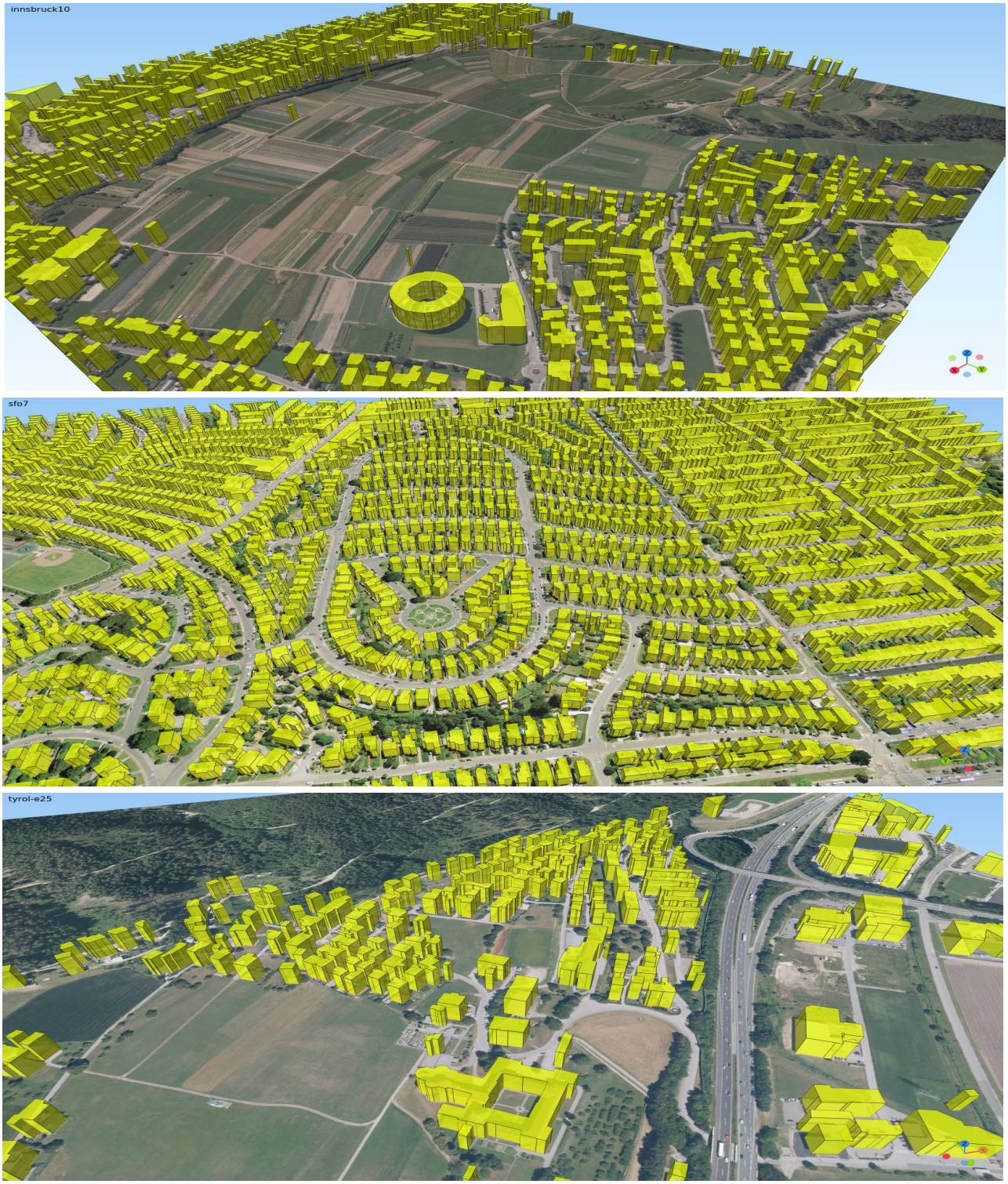

Figure 4. **Qualitative results.** Qualitative examples of extruded building polygons from the INRIA (150) dataset's official test split. Pix2Poly can predict high-quality building footprints that are immediately usable for 3D reconstruction.

Figure 5. **Qualitative results.** Additional qualitative examples of building predictions from the Spacenet Vegas dataset's validation split.

Figure 6. **Qualitative results.** Additional qualitative examples of building predictions from the INRIA (150) dataset's validation split.

Figure 7. **Qualitative results.** Additional qualitative examples of road network predictions from the Massachusetts Roads dataset's test split.