Rethinking Language Models for Building Outline Extraction from Remote Sensing Imagery
* Supplementary Materials *

Kuanren Qian* Yang He* Mohamed Moustafa
Amazon Last Mile
Bellevue, Washington, USA
{kuanqian,yanhea,mmoustm}@amazon.com

Table 1. Impact of image encoder details *w.r.t* the PSP module and the spatial dimension.

Dataset	Pretrained Backbone	PSP Module	Spatial Dim.	IoU	N_{ratio}	MTA
INRIA	Swin-B		7×7	78.89	0.89	37.25
			14×14	**79.91**	**0.95**	**35.05**
		✓	14×14	79.90	**0.95**	35.08
INRIA	SAM2		7×7	79.01	0.94	36.33
			14×14	**80.47**	0.94	35.17
		✓	14×14	80.16	**1.01**	**33.28**
WHU	SAM2		7×7	86.89	0.97	35.60
			14×14	86.99	0.97	34.64
		✓	14×14	**88.69**	**0.98**	**33.83**

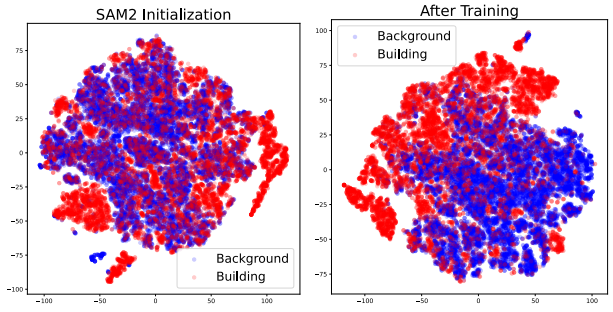


Figure 1. t-SNE comparison of the pixel-wise image embeddings between SAM2 initialization and our finetuned version.

1. Effects of vision backbones

We compare the impact of different network backbones and pretrained representations in Table 5, in the main submission. In this section, we show more experiments on image encoder details. In our main submission, we apply convolution and up-convolution layers to produce feature maps at 14×14 resolution for 224×224 image patches. Besides, there is a pyramid spatial pooling (PSP) module adopted to integrate contextual features. To help understand the impact of image encoder details, we show the performance of using different image feature maps in Table 1. Furthermore, we also remove the PSP module to test our model. From Table 1 highlights some metrics, which are mostly different from our observation. We can see increase the feature resolution is helpful in all different datasets and pretrained models, as the localization of vertices needs finer features. Therefore, we use 14×14 image resolution for all the experiments. In addition, we observe that PSP is not always very helpful for improved performance, for example, INRIA dataset shows similar performance no matter if the PSP module is applied. However, we observe clear improvement in WHU dataset, and then keep the PSP module, which is <CONTEXT> in the main submission. Finally, we keep the PSP model for all the datasets. Last, Figure 1 confirms our learned features form more concentrated clusters, demonstrating task-specific representation learning.

2. Additional Qualitative Results

Figure 2∼4 show additional qualitative results on INRIA, SpaceNet2, WHU datasets, respectively. Similar to our main submission, it is easily observed that the proposed LPM generates building polygons at much higher quality. First, our method can utilize fewer vertices (i.e., key points) to represent a building outlines. Second, our method produces much fewer unreasonable connections on the vertices from different buildings, while Pix2Poly suffers from complex situations and SAM2-UNet also merge two neighbor buildings into one building. Last, we observe that our approach is more robust to occlusions, the LLM encodes useful contextual dependency and predicts reasonable polygons even though the main part of a building is occluded.

Due to the space limit, we do not show qualitative results on CrowdAI dataset. To supplement, Figure 5 shows

*These authors contributed equally.

some comparison results on CrowdAI. Similar to the quantitative results, all the state-of-the-art methods perform comparable in detection performance (e.g., IoU score), our approach still achieves higher polygon quality. For example, we can use fewer vertices to represent buildings than Pix2Poly, while providing the similar IoU score. It is helpful to keep map data simplified, when we apply a building extraction model to construct a digital map.

Consequently, the additional qualitative results further demonstrate the robustness and the strong capability of LPM in producing building outlines in varying situations, compared with other methods.

3. Failure Cases

Figure 6∼8 show more failure cases, that our model achieves relatively low IoU scores in each dataset. We also compare the extraction results from Pix2Poly [1] and SAM2-UNet [2]. From those figures, we can see that the performance decrease is mainly from the detection performance, instead of the building polygon quality. Despite some false positives and false negatives, the proposed LPM does not produce strange polygons as the detected buildings. Regarding the detection performance, this is common issues for all the models, related to the image representation capability and image quality. For example, a small building in the low-contrast areas are also easily missed by SAM2-UNet and Pix2Poly (i.e., last example in Figure 8).

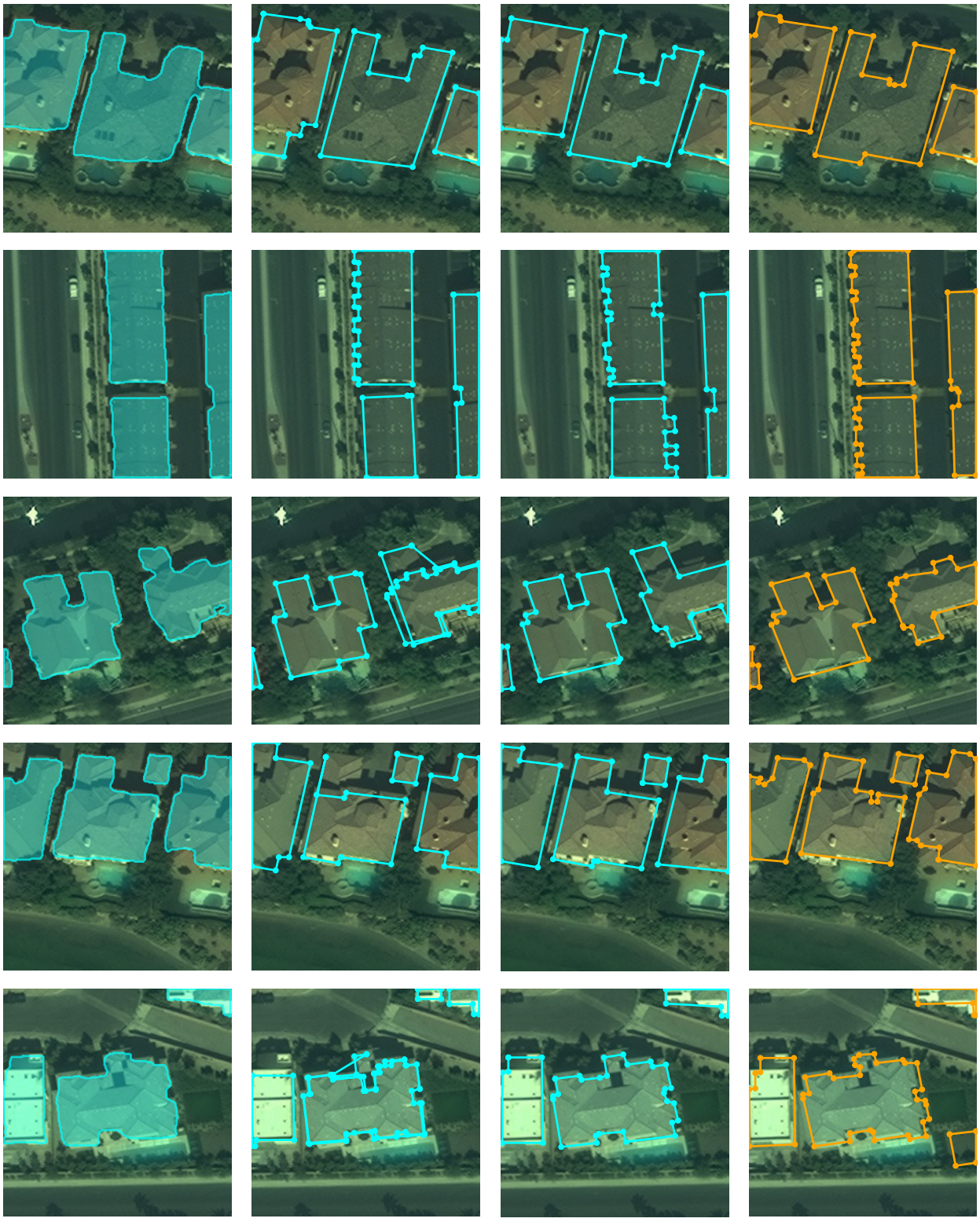
References

[1] Yeshwanth Kumar Adimoolam, Charalambos Poullis, and Melinos Averkiou. Pix2poly: A sequence prediction method for end-to-end polygonal building footprint extraction from remote sensing imagery. In *WACV*, 2024. 2

[2] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*, 2024. 2

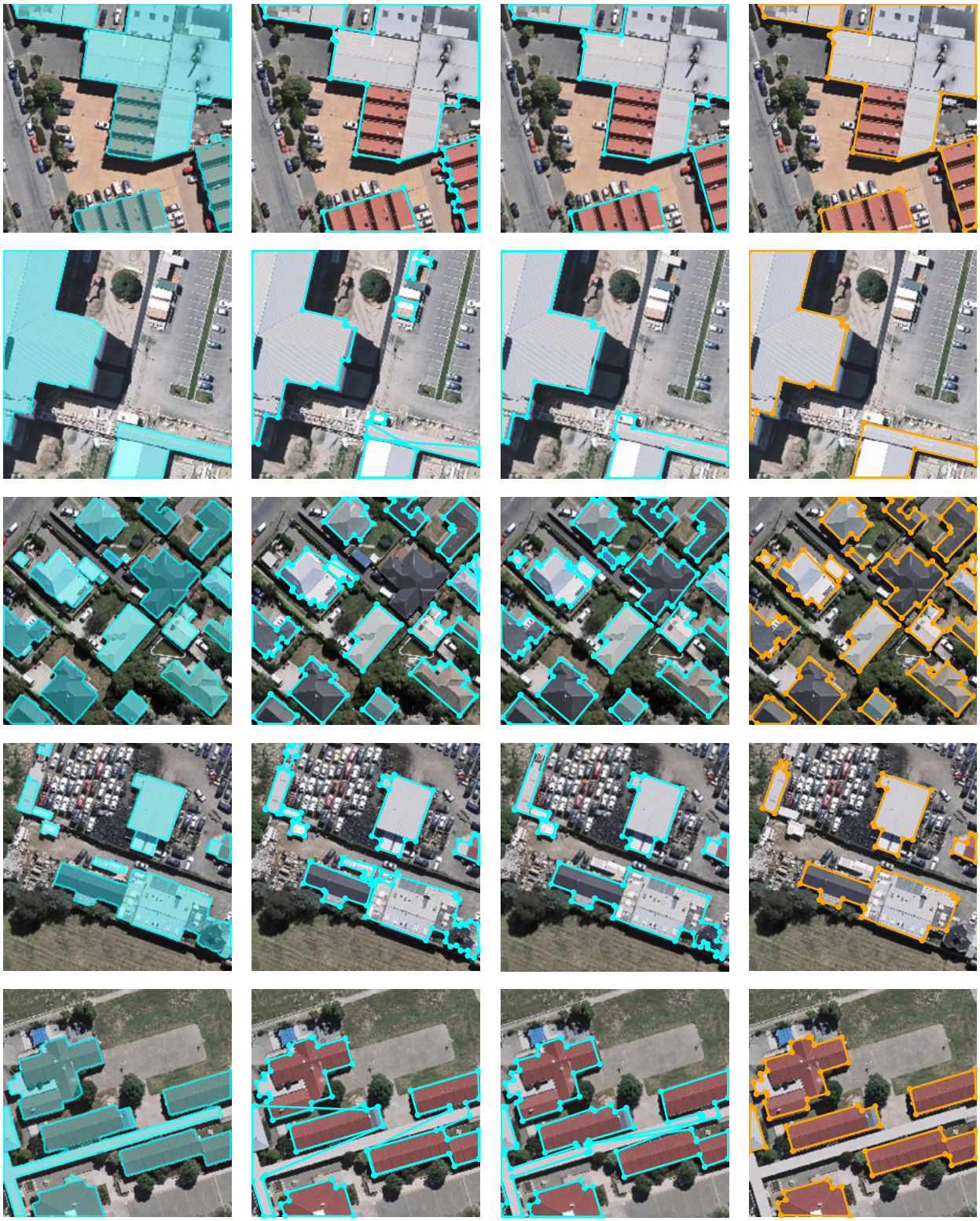
SAM2-UNet (Seg.) | Pix2Poly (Key point) | **LPM (Ours)** | Ground Truth

Figure 2. Qualitative comparison on INRIA *val* set. Cyan outlines are predictions, orange outlines are ground truth. LPM produces cleaner polygons with fewer vertices while maintaining accurate building coverage.



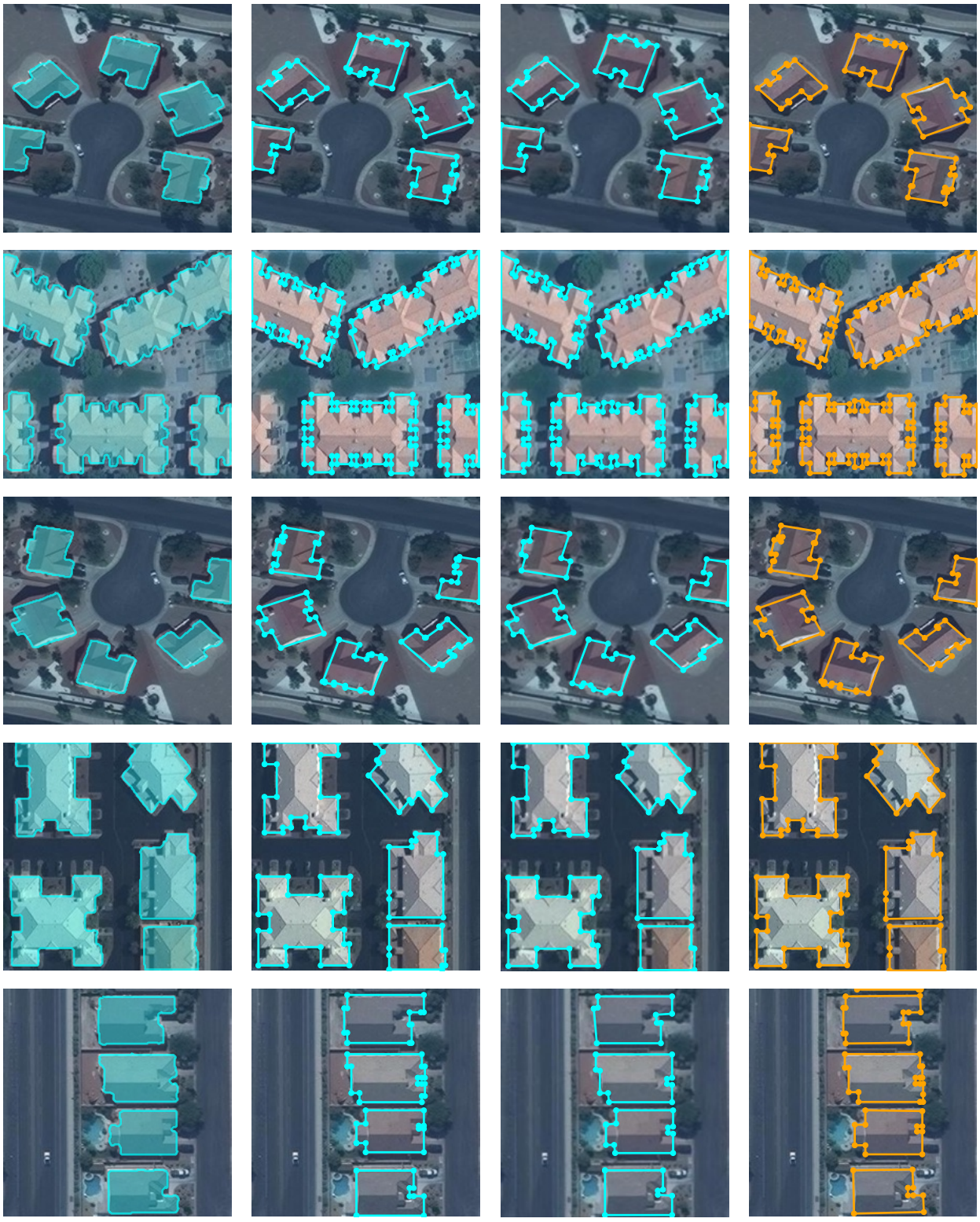
SAM2-UNet (Seg.) | Pix2Poly (Key point) | **LPM (Ours)** | Ground Truth

Figure 3. Qualitative comparison on SpaceNet2. Cyan outlines are predictions, orange outlines are ground truth. LPM produces cleaner polygons with fewer vertices while maintaining accurate building coverage.



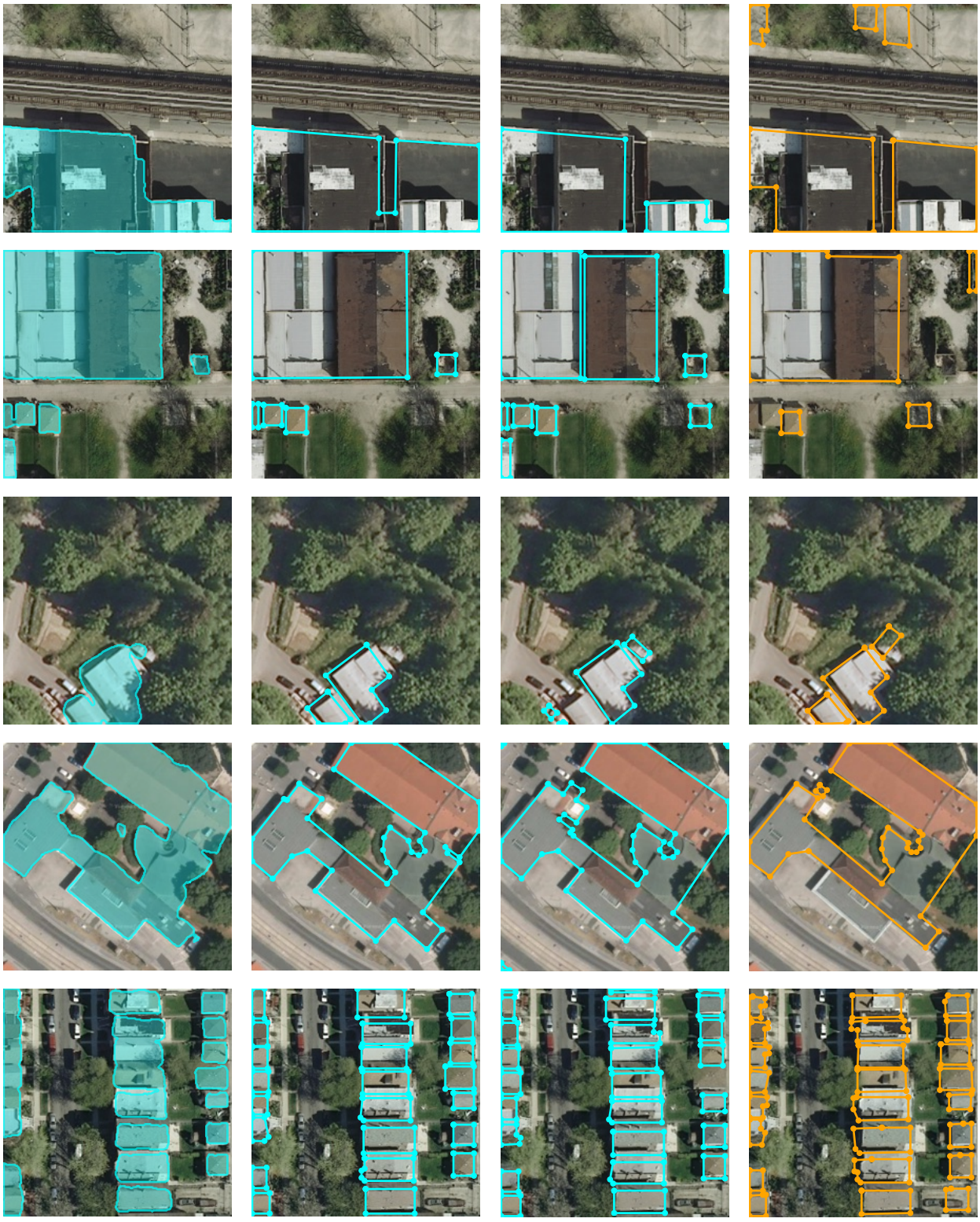
SAM2-UNet (Seg.) | Pix2Poly (Key point) | **LPM (Ours)** | Ground Truth

Figure 4. Qualitative comparison on WHU *test* set. Cyan outlines are predictions, orange outlines are ground truth. LPM produces cleaner polygons with fewer vertices while maintaining accurate building coverage.



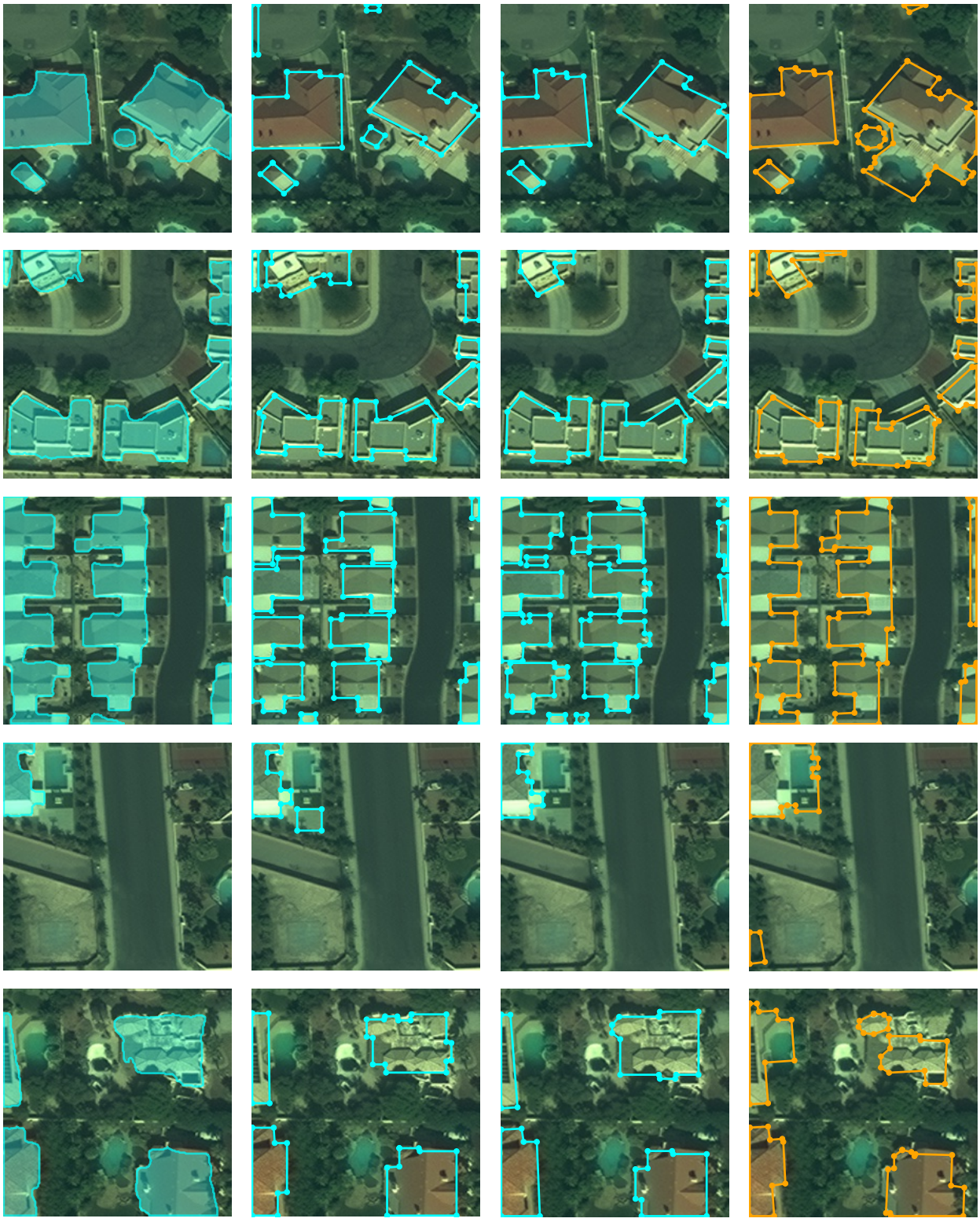
SAM2-UNet (Seg.) | Pix2Poly (Key point) | **LPM (Ours)** | Ground Truth

Figure 5. Qualitative comparison on CrowdAI. Cyan outlines are predictions, orange outlines are ground truth. LPM produces cleaner polygons with fewer vertices while maintaining accurate building coverage.



SAM2-UNet (Seg.) | Pix2Poly (Key point) | **LPM (Ours)** | Ground Truth

Figure 6. Failure cases on INRIA *val* set. Cyan outlines are predictions, orange outlines are ground truth. LPM produces cleaner polygons with fewer vertices while maintaining accurate building coverage.



SAM2-UNet (Seg.) | Pix2Poly (Key point) | **LPM (Ours)** | Ground Truth

Figure 7. Failure cases on SpaceNet2. Cyan outlines are predictions, orange outlines are ground truth. LPM produces cleaner polygons with fewer vertices while maintaining accurate building coverage.

SAM2-UNet (Seg.) | Pix2Poly (Key point) | **LPM (Ours)** | Ground Truth

Figure 8. Failure cases on WHU *test* set. Cyan outlines are predictions, orange outlines are ground truth. LPM produces cleaner polygons with fewer vertices while maintaining accurate building coverage.