Supplementary material

Maxim Khomiakov
maxk@dtu.dk
Michael Riis Andersen
miri@dtu.dk
Jes Frellsen
jefr@dtu.dk

Department of Applied Mathematics
and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark

1 Model setup

The GeoFormer is composed of an image encoder and an auto-regressive decoder. The encoder uses a patch size of 4, a window size of 7, and SWINv2 encoder dimensions of 192. We apply a dropout rate of 0.2 in the SWINv2 layer paths, with depths of 2, 2, 18, and 2 for layers 1, 2, 3, and 4, respectively, each consisting of 6, 12, 12, and 48 attention heads. The encoder's hidden dimensions are 512. The decoder consists of 8 layers, each with 24 attention heads (8 dedicated to ALiBi attention), and dropout of 0.1 and 0.2 in the attention paths and feedforward layers, respectively. We use the Adam-W optimizer with a learning rate of 2×10^{-4}, $\beta = (0.9, 0.999)$, and a weight decay of 1×10^{-2}.

2 Ablation studies

We computed ablations on a combination of factors related to our model's contributions. These included applying the sorting of polygons in each image, incorporating an encoder with SWIN pyramidal feature maps [4], the value of added linear bias (ALiBi) [6], using relative rotational embeddings (RoPE) [7], and adding random token masking during training in our decoder. The total number of ablation combinations for all experiments was 32. In the main paper, we highlighted 8 experiments, while in Table 1, we present the full results.

A primary observation was the significant limitation of our model in fitting the data in any form when applying our pyramidal features without including either ALiBi or mask-based training. In fact, simply adding random masked decoding, along with our pyramidal feature maps, improved performance by 10 percentage points in mAP. Although not evident in this chart, we observed that the model immediately overfitted the data if only the pyramidal feature maps were present, indicating that most learning occurred in the Encoder, leading to poor generalisation when performing auto-regressive decoding.

Another notable observation was that the absence of positional embeddings, particularly in the form of ALiBi, substantially decreased performance. Additionally, when using only either ALiBi or RoPE embeddings, the model became less capable of predicting the correct number of points in the final prediction (N-ratio).

Table 1: Inference results from our ablation studies trained and validated on the small version of the Aicrowd dataset [5]

Sort polygons	Pyramid Features	ALiBi	RoPE	Mask	AP↑	AP_{50}↑	AP_{75}↑	AR↑	AR_{50}↑	AR_{75}↑	bAP↑	IoU↑	C-IoU↑	N-ratio
✓	✓	✓	✓		18.68	34.10	18.99	49.93	70.34	53.10	34.12	67.29	53.62	2.17
	✓	✓	✓		16.73	32.60	15.59	44.97	61.38	48.28	28.19	59.15	49.27	1.99
✓	✓	✓		✓	15.27	28.69	15.16	48.28	71.72	53.79	31.43	51.92	23.18	10.04
✓	✓			✓	15.15	27.96	15.12	52.07	69.66	56.55	32.44	61.33	30.74	8.51
✓	✓	✓	✓	✓	12.89	25.29	11.72	45.17	66.21	51.72	31.07	62.80	45.27	4.03
	✓	✓		✓	12.33	26.25	10.17	39.86	60.69	42.07	22.36	46.18	19.95	9.59
✓	✓	✓			12.10	21.68	12.61	58.07	76.55	62.07	36.73	67.59	48.35	4.30
✓	✓		✓	✓	10.82	22.05	9.28	36.41	54.48	40.00	25.62	53.50	21.49	10.20
✓			✓	✓	10.03	21.06	8.21	45.24	67.59	46.21	27.01	48.51	22.80	9.14
	✓			✓	9.95	20.86	8.50	50.90	71.72	57.24	22.43	40.11	13.68	11.90
	✓	✓			9.58	18.63	9.19	45.59	60.00	50.34	27.40	42.73	17.91	10.61
		✓	✓		9.29	22.42	5.71	39.24	59.31	41.38	20.68	49.70	39.23	2.81
✓		✓			9.01	16.72	8.91	48.55	66.90	51.03	30.67	60.09	24.64	10.19
✓				✓	8.39	16.26	7.85	40.14	56.55	44.83	26.85	53.63	20.17	9.99
✓		✓	✓		7.64	15.35	6.67	50.21	66.21	55.17	28.61	58.00	40.75	4.89
	✓	✓	✓	✓	6.77	16.57	4.10	36.14	56.55	37.93	18.00	41.00	25.18	6.07
				✓	6.23	13.73	4.92	41.38	60.00	45.52	20.46	43.36	23.22	5.51
✓		✓		✓	6.20	12.73	5.21	41.66	64.83	44.14	26.34	50.58	23.23	9.76
✓		✓	✓	✓	5.56	13.45	3.39	26.07	46.90	26.21	17.78	44.60	27.95	5.64
			✓	✓	5.56	13.45	3.39	26.07	46.90	26.21	17.78	44.60	27.95	5.64
	✓		✓	✓	4.54	12.05	2.32	23.72	42.76	24.14	13.67	31.08	11.49	10.88
			✓		4.05	12.08	1.46	22.07	41.38	18.62	15.13	40.94	32.87	2.43
		✓		✓	2.67	7.89	1.07	28.14	47.59	28.97	13.54	35.75	18.23	8.12
		✓	✓	✓	0.61	2.06	0.08	16.90	33.10	14.48	4.91	22.02	10.31	7.64
✓					0.16	0.20	0.20	5.38	17.93	2.07	0.13	5.35	1.71	11.92
		✓			0.13	0.59	0.02	7.52	19.31	7.59	3.05	21.15	13.89	4.20
	✓		✓		0.01	0.03	0.00	0.90	3.45	0.00	0.03	14.79	4.94	12.07
✓			✓		0.00	0.00	0.00	0.14	1.38	0.00	0.04	10.41	3.52	11.85
	✓				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
✓	✓		✓		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
✓	✓				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Overall, we observed that performance was generally higher when introducing the pyramidal features, as opposed to not using them. The introduction of both ALiBi and RoPE embeddings, along with the sorting of polygons, allowed for the best performance in our model. While masked decoding provided somewhat lower performance than without it, we observed faster convergence and better performance with masking when training on the full dataset.

2.1 Robustness studies

To better understand the situations where GeoFormer outperforms previous methods, we conducted a series of robustness studies. These studies were designed to simulate artefacts typically encountered in remote sensing imagery, including variations in spatial image resolution, rotational changes, and missing values. A visual example of the perturbations performed is illustrated in Figure 1. The results are presented in Table 2 and Figures 2. As mentioned in the main paper, we used the smaller Aicrowd dataset [5] for computing metrics, while employing the best model checkpoints trained on the full dataset. In Table 2, we present metrics similar to those in the main paper, with additional columns for Perturbation and Perturbation Factor (PF). Initially, we computed a baseline for each model, representing the performance across the chosen metrics without any perturbations. We followed the same approach as in the main paper, where COCO-metrics are computed as defined in the original MS-COCO benchmark [3], while C-IoU and boundary average precision (bAP)[1] are calculated between the predicted polygon that overlaps the most with the ground truth (requiring a minimum of 0.5 overlap).

Our results show interesting trends across various perturbations. FFL [2] seems to rapidly deteriorate in performance regardless of the perturbation type. Meanwhile, both PolyWorld

Table 2: Robustness results on the small version of the Aicrowd dataset [5]

	Model	AP↑	AP$_{50}$↑	AP$_{75}$↑	AR↑	AR$_{50}$↑	AR$_{75}$↑	bAP↑	C-IoU↑	IoU↑	N-ratio	PF
Baseline	GeoFormer	90.8	95.75	91.94	99.38	100.0	99.31	97.03	97.32	98.02	1.0	0
	HiSup	70.33	88.43	76.53	96.7	99.82	99.52	67.41	89.71	94.19	1.0	0
	PolyWorld	52.65	80.07	58.16	76.34	83.52	80.32	55.01	79.57	83.87	0.91	0
	FFL	37.44	61.6	40.93	82.94	97.57	92.18	49.86	27.85	81.19	5.91	0
Downsample	HiSup	62.1	83.95	69.67	92.73	99.61	98.95	61.75	86.54	91.48	1.0	2
	HiSup	56.65	80.13	63.62	88.91	99.16	97.88	57.22	83.11	88.81	1.0	3
	HiSup	46.13	70.1	51.98	79.83	96.71	91.15	48.2	75.12	82.56	0.98	4
	PolyWorld	45.34	71.83	50.44	64.03	75.29	68.19	48.23	71.43	78.39	0.84	2
	PolyWorld	33.42	57.93	36.33	45.63	58.81	50.34	37.47	56.64	66.65	0.74	3
	FFL	32.84	56.44	35.33	78.41	95.42	88.68	45.94	25.82	78.22	6.29	2
	HiSup	32.59	55.31	35.29	65.68	87.38	76.64	37.6	63.51	72.55	1.01	5
	GeoFormer	31.4	56.03	31.93	59.86	74.48	64.83	64.48	57.82	67.4	0.78	2
	FFL	24.3	45.58	24.13	73.1	94.07	81.13	40.38	21.72	73.2	7.32	3
	PolyWorld	20.71	39.25	20.81	27.67	37.53	30.66	25.37	38.1	49.73	0.57	4
	GeoFormer	17.0	35.25	14.83	48.62	63.45	51.03	57.6	39.43	49.84	0.68	3
	FFL	12.13	27.23	9.51	54.99	77.63	60.38	26.93	16.07	60.27	9.29	4
	PolyWorld	10.54	22.12	9.12	17.99	25.86	18.99	15.36	21.82	32.88	0.42	5
	GeoFormer	5.25	12.23	3.91	35.66	51.72	37.24	53.48	16.83	26.27	0.38	4
	FFL	3.68	9.62	2.2	43.02	67.92	46.09	14.43	11.63	47.0	11.29	5
	GeoFormer	1.51	3.83	1.11	24.69	36.55	24.14	52.69	8.92	15.52	0.29	5
Dropout	GeoFormer	77.39	89.77	79.52	87.59	94.48	89.66	91.85	90.73	92.91	1.0	1
	GeoFormer	66.04	82.09	69.5	79.52	93.79	81.38	87.26	84.27	87.73	0.98	2
	GeoFormer	57.37	77.47	60.13	66.76	82.76	68.97	83.27	77.96	82.13	0.97	3
	HiSup	56.97	75.73	61.13	83.65	93.18	85.29	60.91	77.05	82.62	1.11	1
	HiSup	51.92	69.11	56.7	74.2	83.1	76.23	57.06	69.45	74.92	1.15	2
	GeoFormer	50.49	72.34	53.17	59.66	83.45	58.62	80.52	74.38	79.31	0.93	4
	HiSup	47.55	63.6	52.11	66.24	73.0	68.3	53.76	63.38	68.55	1.16	3
	PolyWorld	46.95	74.81	50.46	69.36	81.46	73.68	51.0	73.77	79.33	0.88	1
	HiSup	44.61	60.33	48.57	60.31	66.54	62.2	50.87	58.91	64.15	1.18	4
	PolyWorld	44.2	71.32	47.4	61.14	75.97	64.07	48.94	70.08	76.43	0.85	2
	PolyWorld	41.54	67.65	44.2	56.5	70.71	58.35	47.06	66.88	73.92	0.83	3
	PolyWorld	39.59	65.33	41.98	47.8	63.39	49.43	45.1	63.08	70.73	0.8	4
	FFL	33.12	56.57	35.0	77.63	95.69	87.87	47.78	26.41	78.4	6.32	1
	FFL	30.39	53.21	31.68	71.37	94.61	79.25	46.0	25.5	75.79	6.56	2
	FFL	27.89	49.64	28.7	64.31	89.49	66.85	44.33	24.79	72.87	6.85	3
	FFL	26.27	47.08	26.65	58.01	84.64	60.65	43.09	24.3	70.58	6.9	4
Rotation	GeoFormer	70.52	93.32	80.72	98.48	99.31	99.31	83.62	91.98	92.64	1.0	3
	HiSup	69.9	87.43	76.37	96.56	99.88	99.46	67.02	89.44	94.11	1.01	3
	PolyWorld	51.33	78.99	57.49	74.16	81.84	79.54	54.09	78.1	82.94	0.9	3
	GeoFormer	16.54	35.32	14.16	33.22	56.52	33.04	55.84	47.71	59.26	1.34	2
	GeoFormer	15.82	33.47	13.36	35.26	56.14	42.11	55.82	46.91	58.28	1.36	1
	GeoFormer	14.87	32.88	11.48	34.74	57.02	37.72	54.15	46.04	57.28	1.19	4
	PolyWorld	14.54	29.47	13.61	19.84	33.24	21.8	18.12	30.37	45.86	0.42	1
	HiSup	14.33	26.05	15.24	59.14	85.66	68.57	26.79	41.65	48.25	1.53	1
	PolyWorld	14.3	29.18	13.35	19.84	31.88	22.34	18.18	30.65	46.17	0.44	4
	HiSup	14.08	25.85	14.8	58.62	84.95	67.49	26.62	41.64	48.24	1.45	4
	HiSup	14.05	25.92	14.8	58.51	84.82	67.93	26.6	41.89	48.78	1.53	2
	PolyWorld	13.13	27.26	11.78	24.36	40.0	26.67	17.5	30.33	46.42	0.43	2
	FFL	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.54	0.94	1.21	1
	FFL	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.55	0.95	1.2	2
	FFL	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.56	1.04	1.19	3
	FFL	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.54	0.92	1.19	4

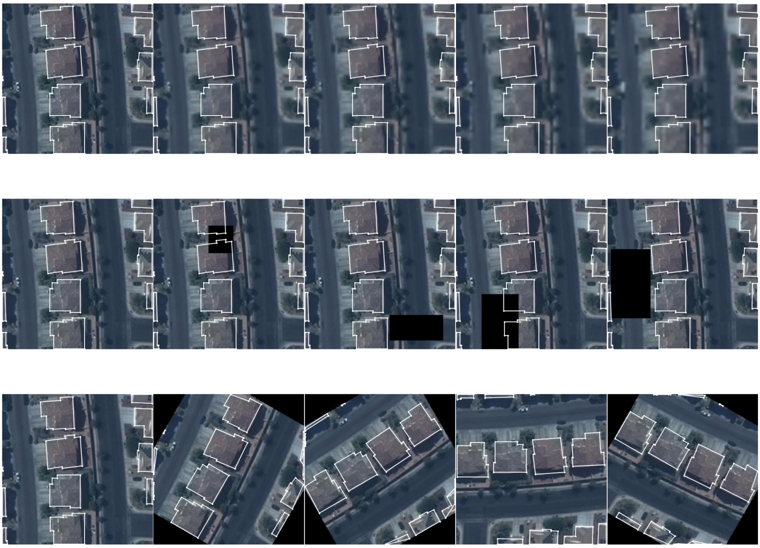

Figure 1: Visual examples of perturbations performed to input images in the robustness studies. From top row: downsampling, erased dropout, and rotations.

[9] and GeoFormer (our model) exhibit significant deterioration upon image downsampling, as seen in Table 2 and Figure 2. In contrast, HiSup [8] also shows degradation in the downsampling scenario, but to a lesser extent. For rotation and dropout perturbations, GeoFormer demonstrates stronger robustness compared to competing methods HiSup, PolyWorld, and FFL. We also observe that GeoFormer is quite robust in terms of boundary average precision [1] in all scenarios except image downsampling, as illustrated in Figure 2 and Table 2. Overall, GeoFormer is on par with or slightly better than competing methods in handling image rotations, while it falls behind in scenarios involving image downsampling. However, it excels relative to other methods in dealing with missing values. The latter is likely due to masked training, while the robustness to rotations is attributed to rotation augmentations during training. The reason for its underperformance in downsampling scenarios is likely explained by the small feature map of 36x36 it needs to decode tokens from, but warrants further investigation.

2.2 Visualising attention maps

We can visualise how the attention mechanism weights the image for each token s_t. This visualisation involves projecting the normalised attention scores onto the input image, following a bi-linear upsampling to match the image's dimensions. We calculate the normalised scores from $\hat{z}$ by averaging across all M attention heads, as follows:

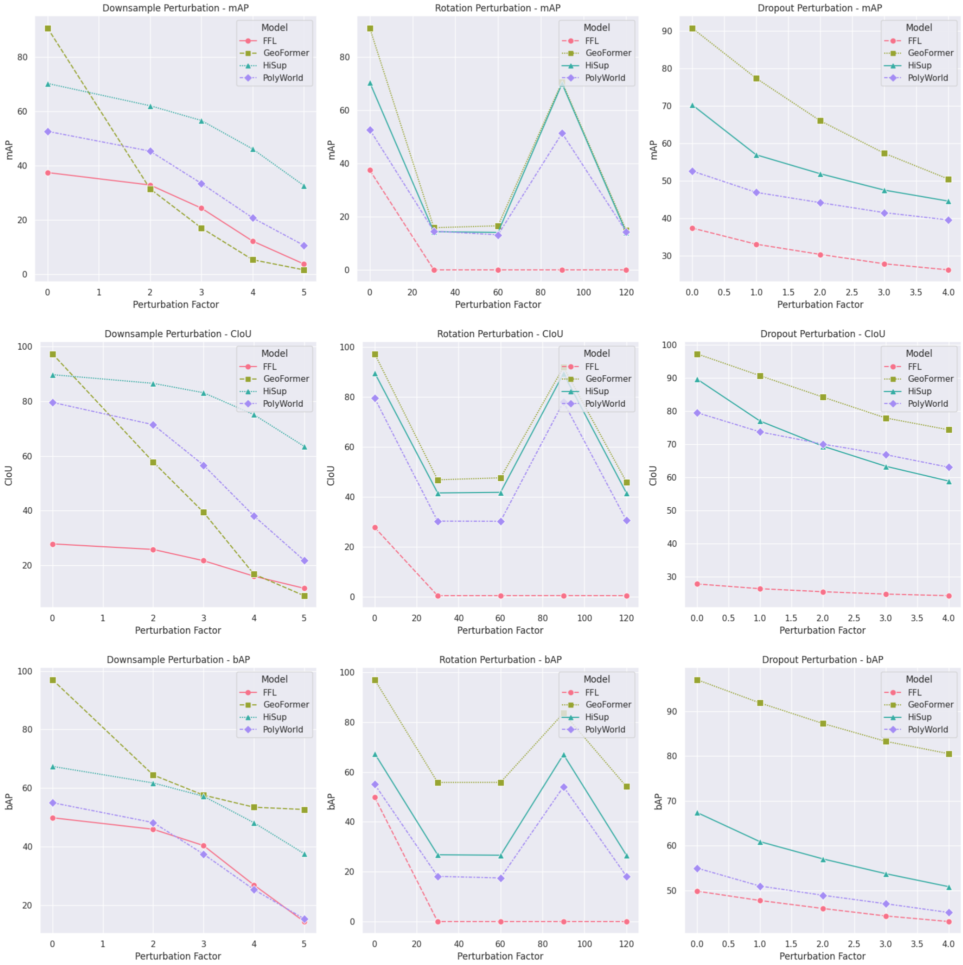

Figure 2: Performance relative to perturbations performed on the Aicrowd small dataset. We perform downsampling, rotations and random dropout. For downsampling a perturbation factor of 2 would equate to a 2x lower spatial resolution, while for dropout, each perturbation factor corresponds to 3%×perturbation factor of pixels that are erased, while for the rotations the perturbation factor is the angles by which the input is rotated.

$$\mathbf{Q} = xW^q, \mathbf{K} = xW^k \tag{1}$$

$$\hat{z}^L = \frac{1}{M} \sum_{m=1}^{M} \sigma \left(\frac{\mathbf{Q}^{(m)} \mathbf{K}^{(m)^T}}{\sqrt{d}} \right) \tag{2}$$

Here, $\mathbf{Q}$ is derived from the decoder tokens, and the keys $\mathbf{K}$ are represented as $I_F \in \mathbb{R}^{36 \times 36 \times C}$, which is the image feature map comprising 36×36 image patches with C hidden dimensions. $\hat{z}^L$ represents the averaged attention scores from the final layer of the decoder.

The visualisation is shown in Figure 3, where we display the attention maps for inferred samples in token pairs. This means averaging over two consecutive token pairs, resulting in each image being composed of an x, y-pair. The images include fully predicted polygons in white, the input RGB image overlaid with attention scores, and a red star indicating the predicted coordinate at the paired timestep $s_{t:t+1}$. We observe how the model shifts its attention towards the area of the building object it aims to predict and how the attention shifts upon the completion of the object, denoted by the special separator $||$-token.

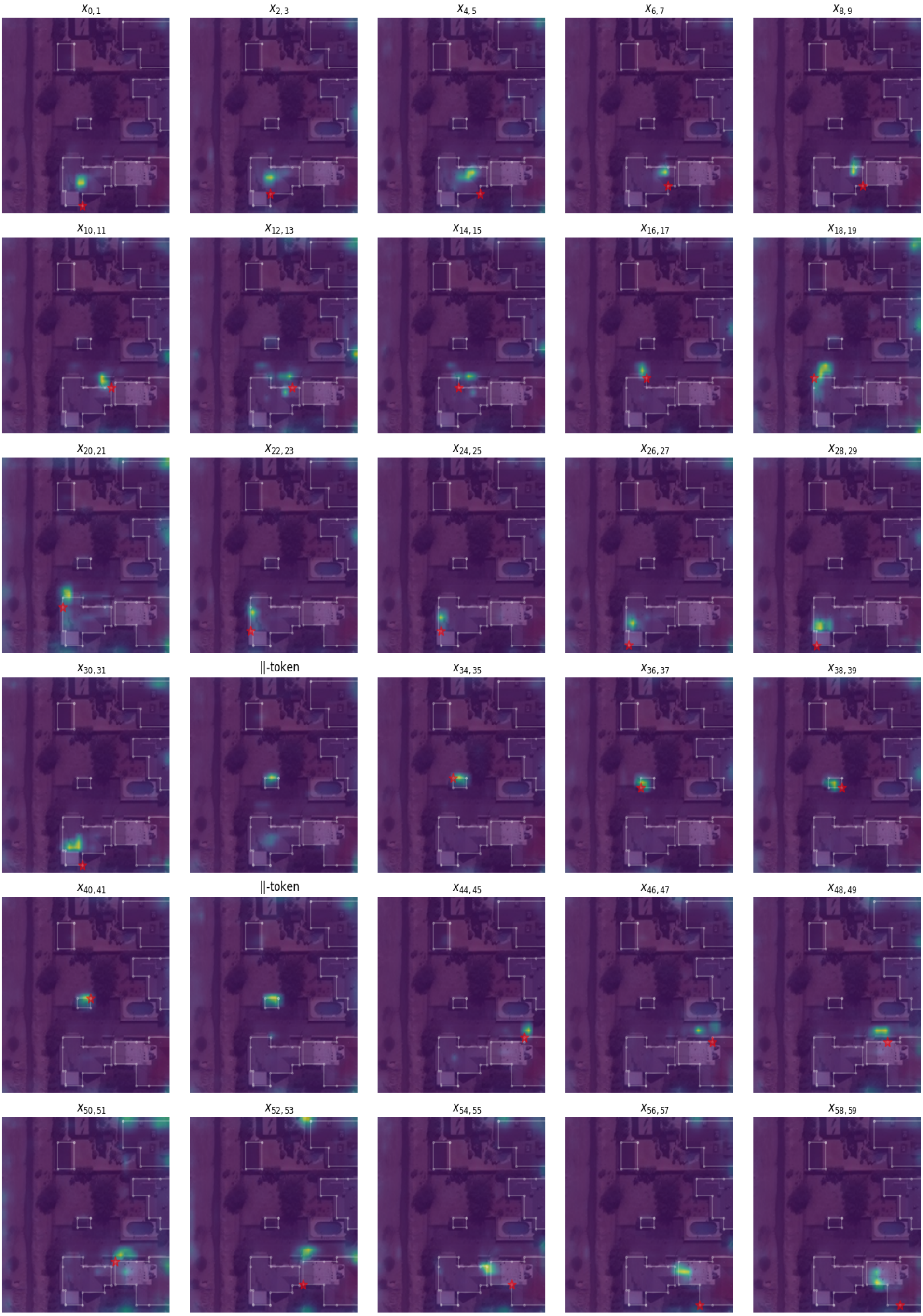

Figure 3: Visualisation of the attention maps on top of the input image and predicted polygons for pairs of tokens $s_{t:t+1}$ from the final layer of the GeoFormer decoder.

References

[1] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15334–15342, 2021.

[2] 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.

[3] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[4] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[5] 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.

[6] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

[7] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[8] Bowen Xu, Jiakun Xu, Nan Xue, and Gui-Song Xia. Hisup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198:284–296, 2023.

[9] 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.