# HOTFormerLoc: Hierarchical Octree Transformer for Versatile Lidar Place Recognition Across Ground and Aerial Views

## Supplementary Material

In this document, we present supplementary results and analyses to complement the main paper. Sec. 7.1 provides a complexity analysis of HOTFormerLoc, Sec. 7.2 provides visualisations of our cylindrical octree attention, and Secs. 7.3 and 7.4 provide ablations of our pyramidal pooling and network size. Sec. 7.5 addresses the limitations and potential future work of our method. We include visualisations of our CS-Wild-Places dataset in Sec. 8. Qualitative examples highlighting components of HOTFormerLoc, and analysis of the learned attention patterns supported by visualisations are presented in Secs. 9 and 10.

## 7. HOTFormerLoc Additional Details

### 7.1. Complexity Analysis

Here, we provide a complexity analysis of the components introduced in Sec. 3.2 of the paper. The key to the efficiency of our approach is alleviating the $O(N^2C)$ complexity of full attention, which is intractable for point clouds with large values of $N$, *e.g.* $30K$. This number of points is essential to capture distinctive information in forest environments. Our H-OSA layer computes windowed attention between non-overlapping windows of size $k$ and their corresponding relay tokens, reducing the complexity to $O((k+1)^2\frac{N}{k}C)$. To facilitate global attention at reduced cost, we conduct RTSA on the relay tokens from $L$ levels of the feature pyramid, with complexity $O(L\frac{N^2}{k^2}C)$.

Our HOTFormer block thus has a total cost of $O(L(k+1)^2\frac{N}{k}C + L\frac{N^2}{k^2}C)$. This reduces the quadratic cost relative to $N$ by a factor of $k^2$, but this effect diminishes when $N \gg k$. For this reason, we opt to employ HOTFormer blocks after first processing and downsampling the $N$ input points into $N_d$ octants with the convolution embedding stem and a series of OSA transformer blocks (similar to H-OSA but with relay tokens disabled), where $N_d < N$. This approach allows us to efficiently initialise strong local features in early stages when semantic information is less developed, which can then be refined by HOTFormer blocks once the size of $N$ is less prohibitive. Another approach would be to consider a larger $k$ for HOTFormer blocks at the finest resolution where $N$ is largest, and smaller values of $k$ at coarser levels, but in this study we have elected to keep $k$ constant throughout the network.

### 7.2. Cylindrical Octree Attention

In Fig. 6 we visualise the relationship between the cylindrical octree hierarchy (albeit in 2D) up to depth 3, and corresponding attention windows with window size $k = 3$ (grouped by color) following $z$-ordering as described in Sec. 3.2. The HOTFormerLoc structure detailed in Fig. 2 can be used interchangeably with Cartesian or cylindrical octree attention windows.

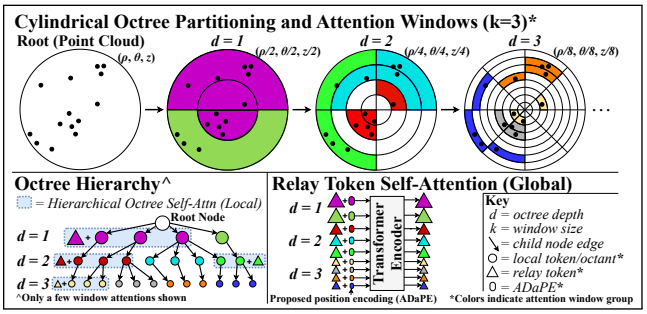



Figure 6. Cylindrical Octree Hierarchy and proposed attention mechanisms shown in 2D for simplicity (3D extends with $z$-axis, so technically the above is a quadtree). Cylindrical partitions and tree nodes are color-matched.

### 7.3. Pyramid Attentional Pooling

We provide an ablation of our pyramid attention pooling (proposed in Sec. 3.3) in Tab. 9, using different numbers of pooled tokens $q$ on Oxford [33], CS-Campus3D [13] and CS-Wild-Places. Overall, we find $q = [74, 36, 18]$ to produce the best results across most datasets, although $q = [148, 72, 36]$ performs marginally better on CS-Campus3D.

These multi-scale pooled tokens $\Omega_l$ are concatenated to form $\Omega'$ and processed by the token fuser [1], generating $q_{\text{total}} = 128$ tokens with $C = 256$ channels in our default configuration. In the MLP-Mixer [47], the channel-mixing and token-mixing MLPs project these tokens to $\bar{k} = 32$ and $\bar{C} = 8$, which are then flattened and $L_2$-normalised to produce the 256-dimensional global descriptor $d_{\mathcal{G}}$.

### 7.4. HOTFormerLoc Ablations

We provide ablations on the number of HOTFormer blocks and channel size in Tab. 10. HOTFormerLoc maintains SOTA performance with fewer parameters than the full-sized model, outperforming MinkLoc3Dv2 by 22.7% on

| Pooled Tokens | Oxford AR@1 (Mean) ↑ | CS-Campus3D AR@1 ↑ | CS-Wild-Places AR@1 (Mean) ↑ |
|---|---|---|---|
| 74, 36, 18 | **92.1** | 79.8 | **60.5** |
| 148, 72, 36 | 91.1 | **80.4** | 52.7 |
| 296, 144, 72 | 89.8 | 74.9 | 48.4 |

Table 9. Ablation study considering the number of pooled tokens used for pyramid attentional pooling on Oxford, CS-Campus3D and CS-Wild-Places.

| Channels | Blocks | Params | Runtime (Sparse / Dense) | Oxford AR@1 (Mean) | CS-Campus3D AR@1 | CS-Wild-Places AR@1(Mean) |
|---|---|---|---|---|---|---|
| C = 256 | M = 10 | 35.4 M | 62 / 270 ms | 92.1 (↑2.1) | 80.4 (↑9.7) | 60.5 (↑8.5) |
| C = 256 | M = 8 | 28.9 M | 50 / 250 ms | 91.8 (↑1.8) | 75.5 (↑4.8) | 58.9 (↑6.9) |
| C = 256 | M = 6 | 22.6 M | 41 / 228 ms | 91.5 (↑1.5) | 71.9 (↑1.2) | 57.6 (↑5.6) |
| C = 192 | M = 8 | 16.7 M | 40 / 192 ms | 90.8 (↑0.8) | 75.2 (↑4.5) | 58.1 (↑6.1) |

Table 10. Ablation on number of HOTFormer blocks and channel size. (↑X.X) indicates improvement in AR@1 over SOTA method per-dataset.

CS-Campus3D and 6.1% on CS-Wild-Places with just 16.7M params. This parameter count is similar to existing transformer-based LPR methods [21, 56], whilst outperforming them by 32.2% on CS-Campus3D. We also report the runtime on dense point clouds from CS-Wild-Places, and the sparse point clouds from CS-Campus3D, with HOTFormerLoc achieving $40-62\ ms$ inference time when limited to 4096 points.

### 7.5. Limitations and Future Work

While HOTFormerLoc has demonstrated impressive performance across a diverse suite of LPR benchmarks, it has some limitations. The processing of multi-grained feature maps in parallel is a core design of HOTFormerLoc, and while effective, it causes some redundancy. For example, there is likely a high correlation between features representing the same region in different levels of the feature pyramid. Currently, these redundant features can be filtered by the pyramid attentional pooling layer, but this does not address the wasted computation earlier in the network within HOTFormer blocks. In future work, token pruning approaches can be adopted to adaptively remove redundant tokens, particularly at the finest resolution where RTSA is most expensive to compute.

Another source of redundancy is related to the number of parameters in our network. A large portion of these are attributed to the many transformer blocks, as each pyramid level has its own set of H-OSA layers with channel size $C = 256$. In the future, the parameter count can be reduced by utilising different channel sizes in each level of the feature pyramid, with linear projections to align the dimensions of relay tokens during RTSA.

As mentioned in Sec. 5.1, the runtime of HOTFormerLoc can be improved through parallelisation. While our design is best suited for parallel implementation, currently, the H-OSA layers for each pyramid level are computed in serial. To unlock the full potential of our network design for optimal runtime, these layers can be combined into a single operation. Furthermore, the octree implementation used in HOTFormerLoc can be parallelised to enable more efficient octree construction.

## 8. CS-Wild-Places Dataset Visualisations

We provide additional visualisations of our CS-Wild-Places dataset, highlighting its unique characteristics. In Fig. 7, we compare a section of the ground and aerial global maps from Karawatha. One notable feature of our dataset is the large-scale aerial coverage, creating a challenging retrieval task where ground queries must be matched against potentially tens of thousands of candidates.

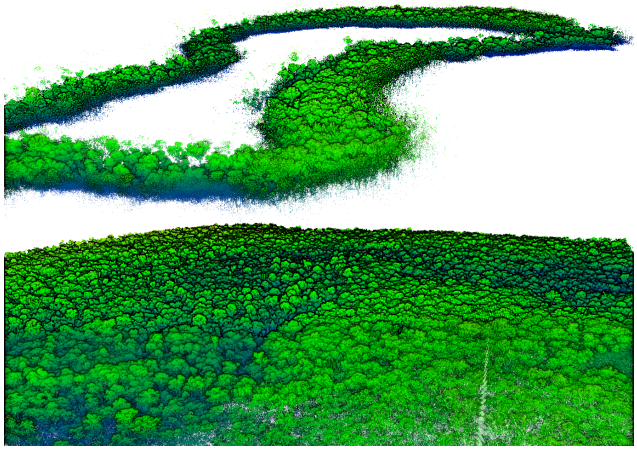
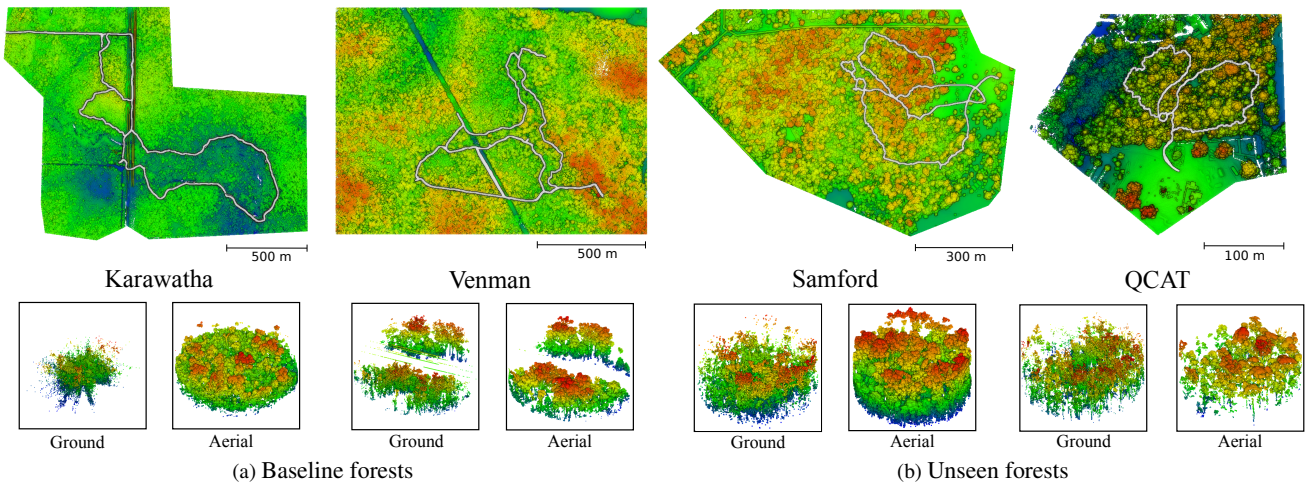


Figure 7. Matched portions of the ground (top) and aerial (bottom) global maps from Karawatha forest in CS-Wild-Places. The aerial maps cover a significantly larger area than the ground traversals, increasing the likelihood of false positive retrievals. Maps are shifted along z for visualisation purposes.

In Fig. 8, we exhibit the scale and point distribution of all four forest environments in the CS-Wild-Places dataset. The Baseline forests have a combined aerial coverage of 3.1 $km^2$, while the Unseen forests add a further 0.6 $km^2$ of aerial coverage. Submaps visualised from each forest showcase the distinct distributional differences between environments. Additionally, the limited overlap between ground and aerial perspectives clearly demonstrate why ground-to-aerial LPR in forested areas is challenging. Notably, our dataset is the first to provide high-resolution aligned aerial and ground lidar scans of this scale in forested environments, offering a valuable benchmark for training and evaluating place recognition approaches.

## 9. Attention Map Visualisations

We provide visualisations of the local and global attention patterns learned by HOTFormerLoc in Figs. 9 to 11. In Fig. 9, we analyse the attention patterns learnt by RTSA for a submap from the Oxford dataset [33] to verify the intuition behind relay tokens. Here, we visualise the attention scores of the multi-scale relay tokens within the octree representation for each level of the feature pyramid (where points represent the centroid of each octant, for ease of visualisation). We select a query token (highlighted in red), and colourise other tokens in all pyramid levels by how strongly the query attends to each (yellow for strong activation, purple for weak activation). We compare the attention patterns of this query token from the first, middle, and last RTSA layer in the network.

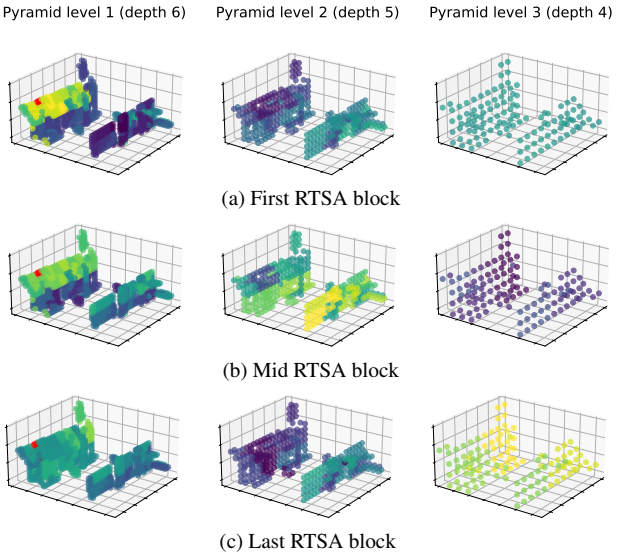



(a) Baseline forests (b) Unseen forests

Figure 8. (Top row) bird's eye view of aerial maps from all forests of CS-Wild-Places. (Bottom row) ground and aerial submap from each. Our dataset features high-resolution ground and aerial lidar scans from four diverse forests, with major occlusions between viewpoints.

We see that RTSA learns a local-to-global attention pattern as it progresses through the network. In the first RTSA layer, the query token primarily attends to other neighbour tokens of the same granularity. In the middle RTSA layer, the local neighbourhood is still highly attended to, but we see higher attention to distant regions in level 2 of the feature pyramid with coarser granularity. In the final layer, the query token primarily attends to tokens in the coarsest level of the pyramid, taking greater advantage of global context. We provide further visualisations of the attention matrices from RTSA in Fig. 10, which highlights the multi-granular attention patterns learnt by different attention heads as tokens propagate through the HOTFormer blocks.

In Fig. 11, we visualise the attention patterns of H-OSA layers, comparing the patterns learnt for different local attention windows as tokens pass through each HOTFormer block. In particular, the presence of strong local dependencies is indicated by square regions with high activations. Interestingly, the relay token (top- and left-most element of each matrix) is uniformly attended to by the local tokens in each window, but with gradually higher attention values in later HOTFormer blocks, indicating the shift towards learning global context in later stages of the network.




(a) First RTSA block

(b) Mid RTSA block

(c) Last RTSA block

Figure 9. Relay token multi-scale attention visualised on the octree feature pyramid at different layers in the network, colourised by attention weight relative to the red query token (brighter colours indicate higher weighting). The network learns a local-to-global attention pattern from the first to last layer.

## 10. Octree Attention Window Visualisations

In Fig. 12 and Fig. 13 we visualise Cartesian and cylindrical octree attention windows generated on real submaps from Oxford [33] and Wild-Places [26]. On the Oxford dataset, which features highly structured urban scenes with flat geometries (such as walls), Cartesian octree windows are a better representation of the underlying scene. Point clouds in Oxford are generated by aggregating 2D lidar scans, as opposed to a single scan from a spinning lidar, producing a uniform point distribution. Furthermore, at coarser levels, the cylindrical octree distorts the flat wall on the left side of the scene to appear as though it is curved. For these reasons, we find that Cartesian octree attention windows perform best on this data.

In contrast, we see the advantage of cylindrical octree attention windows on a submap from Wild-Places in Fig. 13. In the red circled region, it is clear that the coarsest level of the cylindrical octree better represents the shape and distribution of circular lidar scans than the Cartesian octree. Further, the size of each cylindrical attention window reflects the density of points, with smaller, concentrated windows near the centre, and larger, sparse windows towards the edges of the scene. In contrast, the Cartesian attention windows all cover a similar sized region.

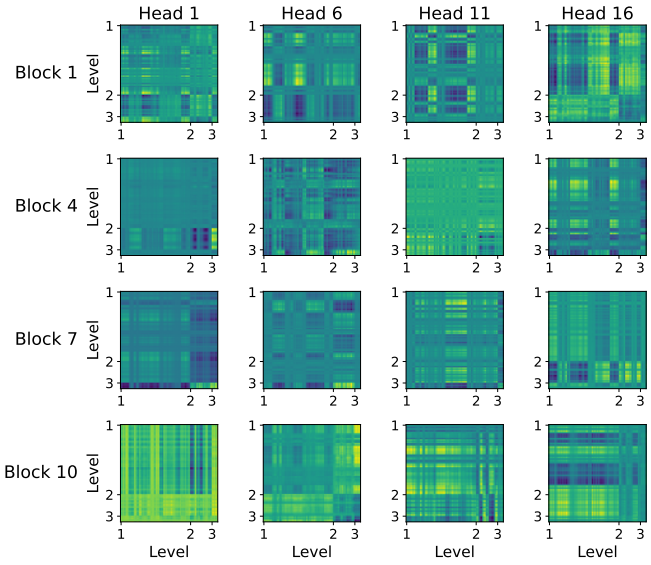



Figure 10. Multi-scale relay token attention matrices from different RTSA heads and blocks for a submap from Oxford. Attention heads learn to focus on different feature granularities (axis ticks indicate pyramid level of corresponding relay tokens).

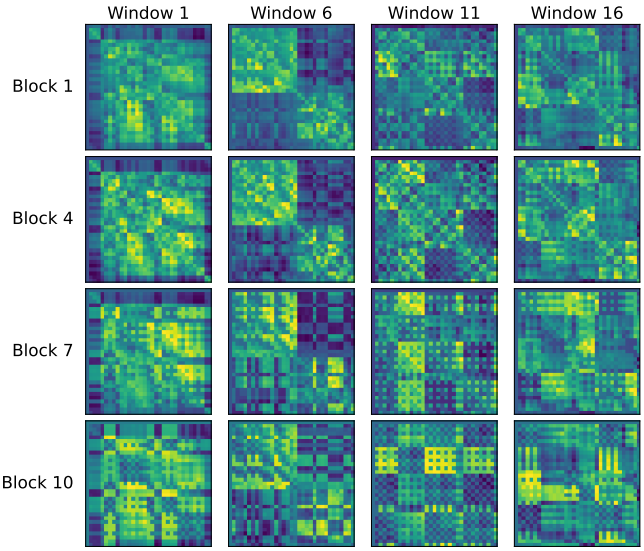
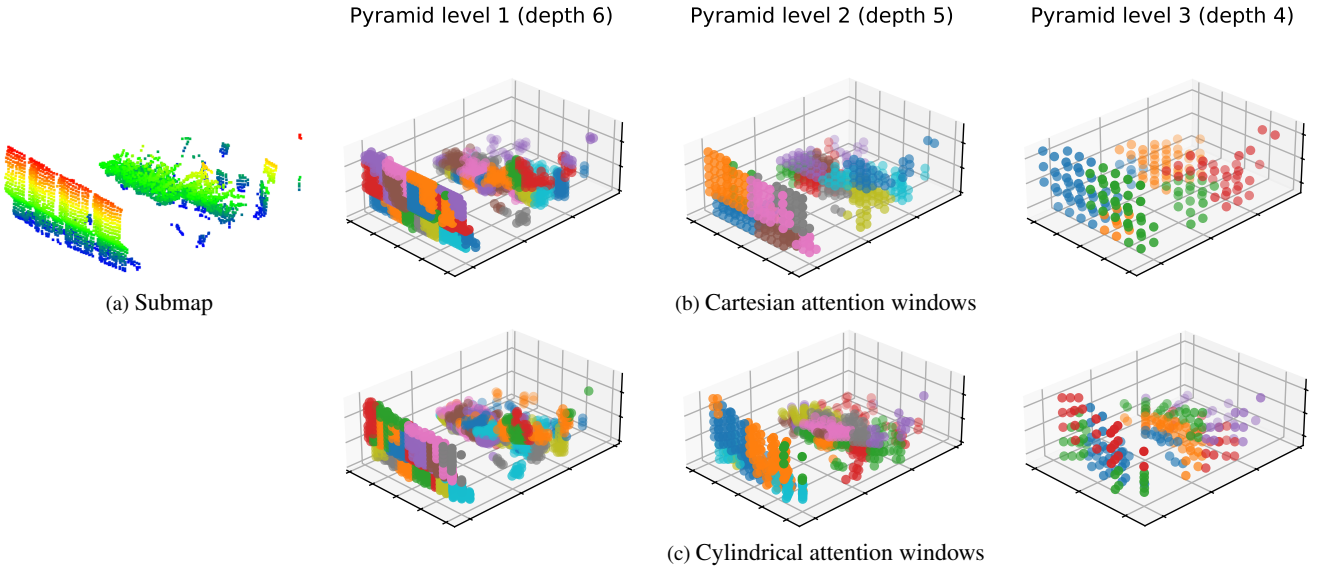



Figure 11. Local attention matrices from different attention windows within H-OSA blocks (averaged over attention heads) for a submap from Oxford. The relay token is represented by the top-left element of each map.

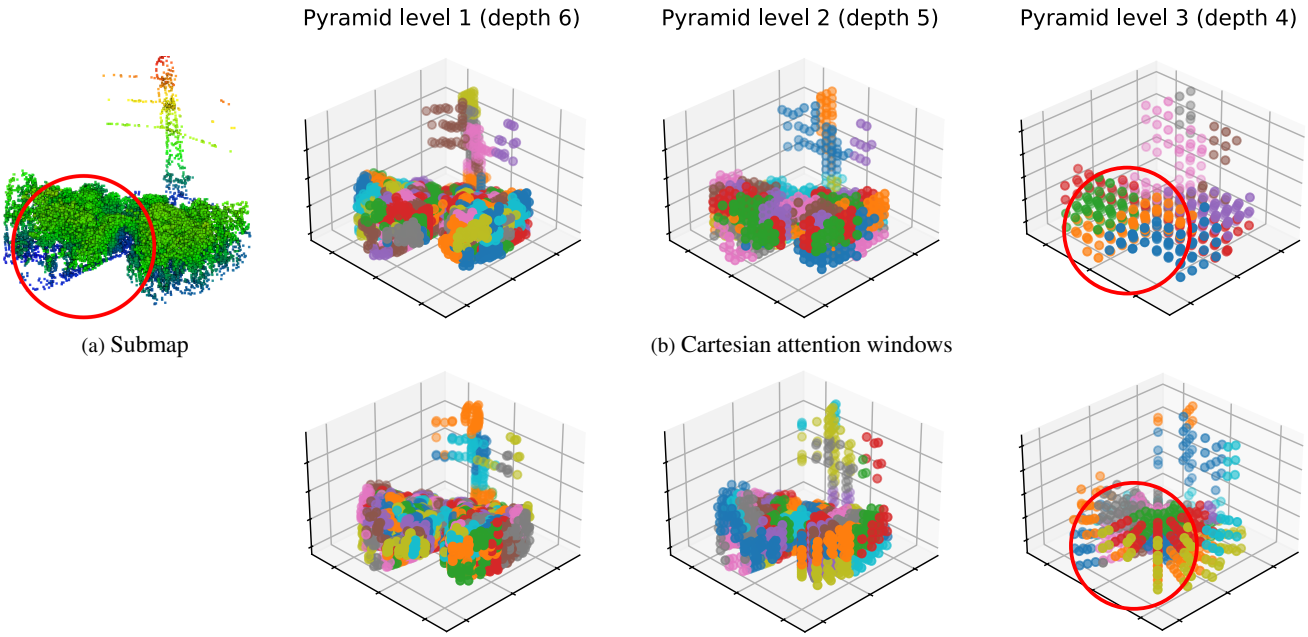



(a) Submap

(b) Cartesian attention windows

(c) Cylindrical attention windows

Figure 12. Comparison of Cartesian *vs*. cylindrical octree attention windows on submaps from Oxford Robotcar [33], where nearby points are colourised by which local attention window they belong to. The uniform nature of aggregated 2D lidar scans and highly-structured scene geometry make Cartesian attention windows a better representation for Oxford.




(a) Submap

(b) Cartesian attention windows

(c) Cylindrical attention windows

Figure 13. Comparison of Cartesian *vs*. cylindrical octree attention windows on submaps from Wild-Places [26]. The variable density of spinning lidar is better captured by cylindrical attention windows in coarser levels, and tree trunks are better represented. We highlight a region where the effect is most noticeable.