# HEAT: Holistic Edge Attention Transformer for Structured Reconstruction Supplementary Document

Jiacheng Chen[1] Yiming Qian[2] Yasutaka Furukawa[1]

[1]Simon Fraser University [2]University of Manitoba

The supplementary material is organized as follows:

- Sect.1: More experimental results including
  - Analysis on the masked learning strategy, the geometry-only edge decoder, and the iterative inference;
  - Ablation studies on the choice of positional encoding for representing edge coordinates;
  - Qualitative evaluation for more testing samples on the two structured reconstruction benchmarks (See outdoor_qualitative.pdf and indoor_qualitative.pdf).
- Sect 2: Details of our corner detection module adapted from the HEAT edge classification architecture.
- Sect.3: Additional implementation details such as the choice of hyper-parameters, training setups for different experiments, training data preparation, and how we reproduced the competing methods.

## 1. Additional Experimental Results

**Masked learning, geometry-only decoder, and iterative inference:** Table 1 complements the Table 4 of the main paper by providing more details about the masked learning strategy the geometry-only decoder, and the iterative inference.. The masked learning strategy alone marginally improves the region scores while adding the geometry-only decoder clearly boosts the region-level performance. These results suggest that the image-aware decoder might overfit to the image features and neglect the geometric patterns revealed by the coordinate features. The geometry-only decoder could effectively alleviate the above issue as a regularization by sharing weights with the image-aware decoder and conducting geometry-only inference.

**Choice of positional encoding:** Table 2 provides an ablation study for the choice of positional encoding for edge coordinates. The first row shows that a proper positional encoding is vital for region-level performance. Besides, the learnable embedding is worse than the trigonometric encoding, potentially because the trigonometric positional encoding preserves useful ordinal priors (*e.g.*, relative distance) that are hard to be learned automatically from data.

Table 1. Detailed ablation study for masked learning strategy, iterative inference, and the geometry-only (geom-only) decoder. "Iter" denotes the number of inference iterations. The pre-trained Faster-RCNN from ConvMPN [6] is used for corner detection.

| Eval Type → | | | Edge | | | Region | | |
|---|---|---|---|---|---|---|---|---|
| Mask | Dec$^{geom}$ | Iter | Prec | Recall | F-1 | Prec | Recall | F-1 |
| - | - | 1 | 75.7 | 60.5 | 67.3 | 74.1 | 50.7 | 60.2 |
| ✓ | - | 1 | 77.4 | 61.2 | 68.3 | 76.4 | 49.7 | 60.2 |
| ✓ | - | 2 | 77.8 | 61.0 | 68.4 | 75.3 | 49.4 | 59.7 |
| ✓ | - | 3 | 77.9 | 61.2 | 68.5 | 76.0 | 50.6 | 60.7 |
| ✓ | ✓ | 1 | 76.7 | 60.4 | 67.6 | 73.7 | 52.2 | 61.1 |
| ✓ | ✓ | 2 | 77.5 | 60.8 | 68.1 | 75.0 | 53.6 | 62.5 |
| ✓ | ✓ | 3 | 77.5 | 60.9 | 68.2 | 74.7 | 53.8 | 62.5 |

Table 2. Ablation study on the choice of coordinate encoding. "Learn" means using a learnable embedding for each discrete value. "Sin/Cos" is the trigonometric positional encoding used by HEAT.

| Eval Type → | Edge | | | Region | | |
|---|---|---|---|---|---|---|
| Coord Enc. | Prec | Recall | F-1 | Prec | Recall | F-1 |
| ∅ | 67.2 | 58.9 | 62.7 | 29.7 | 39.6 | 33.9 |
| Learn | 76.2 | 61.2 | 67.9 | 75.3 | 49.7 | 60.7 |
| Sin/Cos | 77.5 | 60.9 | 68.2 | 74.7 | 53.8 | 62.5 |

**More qualitative results** The files outdoor_qualitative.pdf and indoor_qualitative.pdf provide additional high-resolution qualitative results for outdoor architecture reconstruction and floorplan reconstruction, respectively. The presentation formats are the same as the qualitative figures in the main paper. Due to the size limit, we randomly pick 100 testing samples for each task. Please enlarge the figures to assess the details.

## 2. HEAT Corner Detection

This section explains the details of the corner detector, which is an adaptation of our HEAT edge classification architecture. The HEAT-based corner detection and edge classification modules are trained end-to-end.

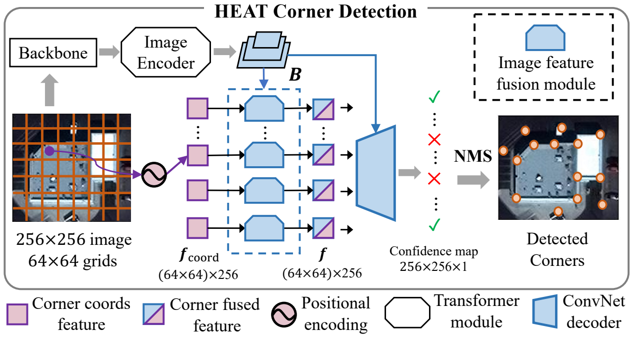

Figure 1. **The corner detection model adapted from HEAT architecture**, illustrated with image size 256×256. We take the HEAT architecture (Fig.2a of the main paper) until the edge-filtering module as the corner detector. Each node is a 4×4 super-pixel. A ConvNet decoder takes the node features and produces a 256×256 confidence map, Non-maximum suppression (NMS) is applied to the confidence map to produce the final corner detection results.

Figure 1 illustrates our corner detection model adapted from HEAT architecture. Pixels are the corner candidates and thus become the nodes. Instead of making every pixel in the 256×256 image space as a candidate, we make each 4×4 super-pixel as a node to reduce the memory cost. An extra MLP takes the concatenated coordinate features of the 16 pixels inside a super-pixel and produces the $\boldsymbol{f}_{\text{coord}}$ for each node. After the image feature fusion, a ConvNet decoder converts the 64×64×256 feature maps into the final 256×256 confidence map. The ConvNet decoder consists of stacks of convolution layers and up-sampling layers, as well as a final linear layer for producing the confidence. Non-maximum suppression is applied to the confidence map to produce the final corner detection results.

We train the above corner detection model jointly with HEAT edge classification. The corner and edge models share the same ResNet backbone. The training data for the edge model are generated on the fly based on the corner detection results. See Sect. 3 for details about training data preparation.

## 3. Additional Implementation Details

**Hyper-parameter and training settings.** As mentioned in the main paper, there are four binary cross-entropy (BCE) losses in the full HEAT framework: one for corner and three for edge. We use a weight of 3.0 (resp. 10.0) for positive samples to balance the positive and negative samples for the edge (resp. corner) BCE loss.

We apply non-maximum suppression (NMS) to the HEAT corner prediction results to clean up the detected corners. Non-maximum predictions inside a local 5×5 window are suppressed. When using a pre-trained corner detection model (*i.e.*, the Faster-RCNN provided by ConvMPN [6]) in our ablation studies on the outdoor reconstruction benchmark, the number of training epochs is reduced from 800 to 500 as we only need to train the edge classification part alone, and the corner BCE is discarded. All other training settings are exactly the same as the full HEAT. For all experiments, we simply take the checkpoint from the last training epoch for evaluation.

**Training data preparation.** We conduct random flipping and random rotation for data augmentation when training the models on the outdoor reconstruction task. However, only random flipping is applied for indoor reconstruction since random rotation always makes the planar graph surpass the image boundary.

For producing corner labels, we first produce a label map with the same resolution as the input image, and then apply a Gaussian blur (with sigma=2) to the label map to alleviate the class imbalance.

For producing edge labels, we follow three steps: 1) Match ground-truth corners with detected corners. A detected corner and a ground-truth corner are matched if their distance is smaller than 5 pixels and both of them are not matched with other corners; 2) Generate training-time corner candidates. We produce the set of corner candidates by merging all the detected corners and ground-truth corners, but removing the ground-truth corners that are matched by detected corners; 3) Enumerate corner pairs and assign edge labels. We enumerate all pairs of corner candidates to generate the edge candidates. The label of an edge candidate is true if and only if each of its endpoints is either a ground-truth corner or a matched detected corner, otherwise the label is false.

**Running competing approaches.** We explain how we run the competing approaches to evaluate them on the two structured reconstruction benchmarks:

- HAWP [5]: We adapt the official HAWP implementation[1] for the two structured reconstruction tasks with two modifications: 1) we change the image resolution according to the experimental setups and 2) we increase the number of training epochs to make it the same as HEAT and LETR. We found that increasing the number of training epochs does not improve HAWP on its original wireframe parsing task, but can clearly improve its performance on the structured reconstruction tasks.
- LETR [4]: We adapt the official implementation of LETR[2] for the two structured reconstruction tasks with three modifications: 1) we change the image resolution based on the experimental setups; 2) we search for the best hyper-parameter

[1] https://github.com/cherubicXN/hawp

[2] https://github.com/mlpc-ucsd/LETR

for the number of "dummy query nodes" and change it from 1000 to 100, which is also consistent with the dataset stats in Sect.3 of the main paper (the model with default settings cannot converge on our two benchmarks); and 3) we run a simple post-processing to merge neighbouring corners within 10 pixels to get a cleaner planar graph. Note that we strictly follow LETR's three-stage training pipeline.

• ConvMPN [6] and Exp-cls [7]: We run the released official checkpoints[3] to get the quantitative evaluation results and corresponding qualitative visualizations.

• Others: For other domain-specific approaches (*i.e*., IP [2], MonteFloor [3], Floor-SP [1]), we directly borrow their evaluation results from previous papers.

## References

[1] Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floor-sp: Inverse cad for floorplans by sequential room-wise shortest path. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2661–2670, 2019. 3

[2] Nelson Nauata and Yasutaka Furukawa. Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference. In *ECCV*, 2020. 3

[3] Sinisa Stekovic, Mahdi Rad, Friedrich Fraundorfer, and Vincent Lepetit. Montefloor: Extending mcts for reconstructing accurate large-scale floor plans. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[4] Yifan Xu, Weijian Xu, David Cheung, and Zhuowen Tu. Line segment detection using transformers without edges. In *CVPR*, 2021. 2

[5] Nan Xue, Tianfu Wu, Song Bai, Fudong Wang, Guisong Xia, Liangpei Zhang, and Philip H. S. Torr. Holistically-attracted wireframe parsing. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2785–2794, 2020. 2

[6] Fuyang Zhang, Nelson Nauata, and Yasutaka Furukawa. Conv-mpn: Convolutional message passing neural network for structured outdoor architecture reconstruction. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2795–2804, 2020. 1, 2, 3

[7] Fuyang Zhang, Xiang Xu, Nelson Nauata, and Yasutaka Furukawa. Structured outdoor architecture reconstruction by exploration and classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[3] https://github.com/zhangfuyang/Conv-MPN and https://github.com/zhangfuyang/search_evaluate