IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-2010-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IGARSS52108.2023.10282198

VECTORIZING PLANAR ROOF STRUCTURE FROM VERY HIGH RESOLUTION REMOTE SENSING IMAGES USING TRANSFORMERS

Wufan Zhao, Claudio Persello, Xianwei Lv, Alfred Stein

Department of Earth observation science, Faculty of Geo-information Science and Earth Observation (ITC),
University of Twente,
Hengelosestraat 99, 7514 AE Enschede, the Netherland

ABSTRACT

Grasping the roof structure of a building is a key part of building reconstruction. Directly predicting the geometric structure of the roof from a raster image to a vectorized representation, however, remains challenging. This paper introduces an efficient and accurate parsing method based upon a vision Transformer we dubbed Roof-Former. Our method consists of three steps: 1) Image encoder and edge node initialization, 2) Image feature fusion with an enhanced segmentation refinement branch, and 3) Edge filtering and structural reasoning. The vertex and edge heat map F1-scores have increased by 2.0% and 1.9% on the VWB dataset when compared to HEAT. Additionally, qualitative evaluations suggest that our method is superior to the current state-of-the-art. It indicates effectiveness for extracting global image information and maintaining the consistency and topological validity of the roof structure.

Index Terms— Roof structure extraction, remote sensing image, Transformer, geometry reconstruction

1. INTRODUCTION

The creation of comprehensive 3D building models requires access to roof structure information. Such models are useful in applications such as building energy modeling, and urban planning [1]. Various remote sensing data, including monocular/stereo images, point clouds, and digital surface models, have been used to extract geometric building outlines and roof structures [2]. However, the procurement of 3D spatial data and the creation of accompanying 3D models are costly, especially over large areas. In contrast, roof structure extraction based on optical remote sensing images has the potential to offer low-cost and broad coverage advantages.

Roof structure is a topological collection of fine-grained geometric elements of building roofs that combine line and junction elements with their connections. Conventionally, geometric feature extraction from images is carried out using perceptual grouping of low-level cues. The advent of deep neural networks has introduced groundbreaking advances in spotting low-level primitives and recognizing high-level geometric structures. End-to-end trainable methods have achieved notable performance in detecting lines, points, wireframe, and floor plans [3, 1]. Despite this progress, research on automated extraction of the structured geometry (outline and roof structure) from optical remote sense images has been limited due to scene complexity and the large variety of roof top configurations [2, 4, 1]. Existing methods, however, reveal several false positive candidates in the extracted geometric primitives that are not positioned inside buildings. Additionally, they often drop adjacency relationships among the primitives.

Vision tasks have made significant use of the Transformers' sequence-to-sequence model [5]. The model denotes both input features and output targets as visual tokens, which engage in global interactions with one another via the attention mechanism of the Transformers. Based on DETR [5] developed for object detection, Holistic Edge Attention Transformer (HEAT) [6] was proposed to restructure a planar graph representing an underlying geometric structure. However, issues concerning the effective and efficient extraction of global image features persist due to insufficient single-scale feature maps and high computational costs [6].

We propose the Roof-Former, a Transformer network for efficient planar roof structure extraction from very high resolution remote sensing images. The Roof-Former ensures the consistency of spatial and topological relations of the extracted primitives within the roof structure by combining tokenized entity modeling, primitive detectors, and relationship inference. Our method is based specifically on HEAT [6], and we introduce an enhanced feature pyramid module to the Transformer, which enables the image encoder to learn multiscale features while reducing resource consumption during training. We also add a collaborative segmentation refinement branch to the existing framework, which ensures the spatial and topological relations of the extracted primitives within the roof structure by jointly learning the building masks. Different modality features are effectively fused based on an Attention Feature Fusion Module (AFFM). We evaluate our proposed method on the benchmark dataset and demonstrate its effectiveness in global structural reasoning compared to other

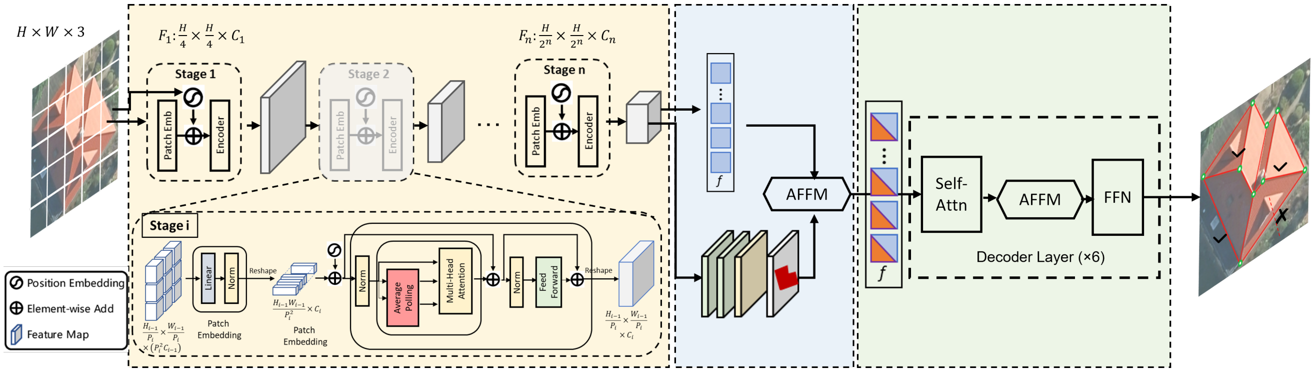

Fig. 1: The overall architecture of Roof-Former, which consists of three steps: 1) Image encoder and edge node initialization (yellow); 2) Image feature fusion with enhanced segmentation refinement branch (blue); and 3) Structural reasoning with Transformer decoders (green).

methods.

2. METHODOLOGY

The overall architecture of Roof-Former is designed on the basis of HEAT. It identifies vertices and categorizes edge candidates between vertices in an end-to-end manner (Figure 1). The model infers vectorized planar graphs (i.e., vertices and edges) representing a roof structure given a 2D raster image. The proposed Roof-Former comprises three modules: 1) Image encoder and edge node initialization, 2) Image feature fusion with enhanced segmentation refinement branch, and 3) Edge filtering and structural reasoning.

2.1. Image Encoder and Edge Node Initialization

Our Roof-Former network extracts the image feature map from a backbone with the reduced dimension using an input image of size $H \times W \times 3$. To create spatial relations, positional embeddings are concatenated with image features. Unlike HEAT, we introduce an enhanced pyramid structure into the Transformer framework, called the Feature Pyramid Transformer (FPT). The backbone consists of four stages that yield feature maps in varying scales, and each stage has a patch embedding layer and L_i Transformer encoding layers. The output resolution of the four stages gradually decreases from high with 4-stride to low with 32-stride, and P_i denotes the patch size of stage i. Each of the L_i encoder layers in the Transformer encoder's stage i is made up of an attention layer and a feed-forward layer. We utilize a linear spatial reduction attention layer to replace the encoder's multi-head attention layer, gaining a linear computational and memory cost.

In the vertex detection network, each 4×4 super-pixel is assigned as a node in our network, instead of a pixel in the 256×256 image space, to minimize memory costs. Each node's f_{coord} is built with an additional Multilayer perceptron (MLP) by summing the coordinate features of the 16 pixels that make up a super-pixel. A ConvNet decoder transforms the $64 \times 64 \times 256$ feature maps into the final 256×256 confidence map, consisting of convolution layers, upsampling layers, and a final linear layer for confidence map generation. To yield the final vertex detection results, we apply non-maximum suppression to the confidence map. Each pair of vertices functions as an edge candidate and becomes a Transformer node, with the feature f_{coord} initialized by the 256-dimensional trigonometric positional encoding. The vertex detection model undergoes combined training with edge classification. Overall, our Roof-Former network efficiently extracts roof structure from very high resolution remote sensing images by introducing an enhanced pyramid structure into the Transformer framework and using a vertex detection network for combined training with edge classification.

2.2. Image Feature Fusion with Enhanced Segmentation Refinement Branch

We adapt the deformable attention method used in HEAT to inject image features into each edge node. We generate sampling sites and attention-weights for image feature aggregation at each level of the feature pyramid in the image encoder using an 8-way multi-head attention strategy. To boost object accuracy for large-scale roof structure mapping, we add an additional semantic segmentation branch along the Transformer, producing the semantic binary label using the building outline in the building segmentation branch. After the backbone network, we convert the backbone feature into the embedding features of vertices and segmentation maps, predicting vertice heatmap and the segmentation mask for the building polygons. For the mask branch, we use two convolutional neural networks with a shared feature map from the backbone network, using a sigmoid activation function in the output layer for the aided segmentation map.

We introduce an attention feature fusion module (AFFM) that uses segmentation branches to enhance feature fusion across tasks. The efficient attention mechanism enriches the feature fusion of mask and edges at both global and local scales by establishing opposing spatial pooling sizes and selectively executing channel attention at distinct scales. To keep the weight light, the local context is simply tacked onto the global context inside the attention module. A local channel context aggregator, named point-wise convolution (PWConv), relies only on point-wise channel interactions at each spatial position, conserving parameters using the bottleneck structure.

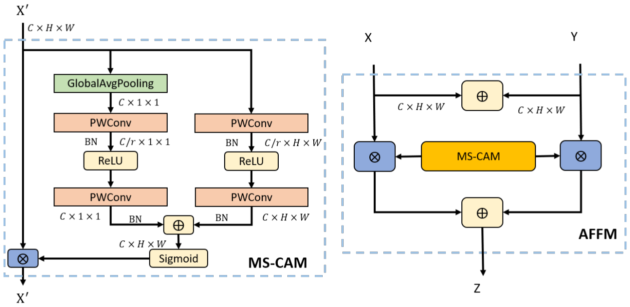

Fig. 2: Illustration of the proposed AFFM.

The feature maps X and Y are then taken into consideration. After fusing these features, the network can apply mask-level guidance to limit the scope of candidate primitives. Additionally, the AFFM is used to combine the edge candidates with the aided segmentation map in order to generate line proposals. When deciding whether or not to keep a primitive, the aided segmentation map is fused with the candidate primitives from the primitive detection branch. This is performed by taking the primitive's location and state relative to other features in the image. Once the retrieved candidate primitives satisfy the segmentation layer's range, they are activated. They will be otherwise suppressed.

2.3. Edge Filtering and Structural Reasoning

After integrating image and mask feature with edge node, we generate a fused feature by integrating a conventional addnorm layer and a feed-forward network (FFN), as in the original Transformer. Following that, we eliminate unsuited candidates by putting f through a 2-layer MLP followed by a sigmoid function and generating a confidence score. Top-K candidates are kept, where K is three times the number of vertex candidates.

Two weight-sharing Transformer decoders are used to categorize every edge candidate as either correct or not. Each edge candidate is modeled as a node and assigned the fused feature f by the image feature decoder. An 8-way multi-head attention mechanism is incorporated into the network's six layers of self-attention, edge image feature fusion module, and feed forward network. The geometric decoder has the same architecture and shares the weights without using image information. It is used to enhance the global geometric reasoning and performance of the image-aware decoder. We use the same masked training and iterative inference as in [6].

3. EXPERIMENT AND RESULTS

3.1. Dataset and Evaluation Metrics

We performed experiments on the Vectorizing world building dataset (VWB) [7] to verify the performance and robustness of our method. The VWB dataset is a part of the SpaceNet challenge 2 with a spatial resolution of $\sim$ 30 cm. The entire dataset contains 2001 patches in total. We separated them into 1601 and 400 for training and testing, respectively. The image patches are with the size of 256×256 pixels.

We applied two evaluation schemes for evaluation, including pixel-wise and vector-wise metrics. Specifically, We compute a heat map based average precision (AP^H) and an F1-score (F^H) for each of the vertex and edge primitives. We also apply the mean structural Average Precision ($msAP$), metrics defined on vectorized wireframes for both vertex ($msAP^V$) and edges ($msAP^E$) [1].

3.2. Experimental Setup

The PyTorch environment was used for all experiments. All training and testing were carried out on a single GTX 2080Ti GPU with 12 GB of memory. The loss balancing weights for the three edge Binary cross-entropy (BCE) losses all equal 1.0, while the vertex prediction BCE weight equals 0.05. The Adam optimizer is used to train our model. We set an initial learning rate as $2e^{-4}$, and a weight decay factor of $1e^{-5}$. For the last 25 epochs, the learning rate decays by a factor of 10. Our network is trained for 400 epochs. Roof-Former can be trained from end to end without the requirement for a separate preparatory extraction phase.

3.3. Results and Evaluations

We compared our method against four competing methods: ConvMPN [4], HAWP [3], RSGNN [1], and HEAT [6]. Each model was trained and evaluated using the same split.

3.3.1. Quantitative analysis

Table 1 shows the experimental results. Roof-Former surpasses all the competing methods on all the precision and F1-scores. Specifically, compared to HEAT, our method has greatly enhanced the vertex and line segment outcomes, and the vertex and edge heat map F1-scores have risen by 2.0 points on the VWB dataset. The $msAP$ for vertices and edges is 43.1, which is higher than the HEAT and other methods. The results indicate that our method also increases geometric accuracy.

Table 1: Quantitative results on the VWB dataset.

Datasets	Methods	$msAP^V$	$msAP^E$	AP_V^H	F_V^H	AP_E^H	F_E^H
VWB	ConvMPN	35.7	34.2	78.0	78.8	57.0	58.1
	HAWP	31.1	31.0	90.9	85.2	76.6	72.1
	RSGNN	34.8	34.6	89.6	85.7	76.4	75.7
	HEAT	41.6	40.3	91.7	87.1	80.6	76.2
	Ours	43.1	42.4	92.3	89.1	82.3	78.1

Non-Transformer methods rely predominately on image features and do not acquire global geometric reasoning across query nodes, resulting in a large number of false edges and building reconstructions that do not resemble buildings. The performance gap is especially noticeable for edges, which involve high-level geometry reasoning.

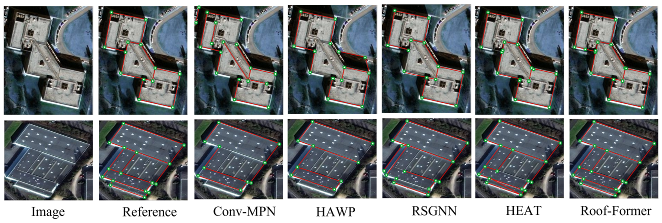

Fig. 3: Sample results on the VWB datasets.

3.3.2. Qualitative analysis

Figure 3 provides the qualitative comparisons. The reconstruction quality of Roof-Former is easily noticed to be superior to competing methods and closer to reality, which is true even when massive and complex buildings are considered. Carefully scrutinizing the structures, Roof-Former is particularly effective at determining global information and maintains the overall prediction consistency and geometric validity (e.g., less hanging edges, not distracted by background buildings). It can be easily observed that the addition of the building segmentation branch can significantly increase the accuracy of roof structure extraction.

3.4. Discussion

There are still some major failures of our methods. First, our method fails to recover from vertices overlooked by the vertex detector. Absent vertices result in absent incident graph structure or degraded geometry. Second, our method adopts a piece-wise linear structure and cannot deal with curved buildings. Third, our method may become less effective when transferring our method to oblique, relatively low-resolution satellite images. In addition, the segmentation outcomes are strongly reliant on the quality of the reference data. Future research will explore Generative adversarial networks (GANs) to improve instance data augmentation and more effective training processes to improve the proposed method.

4. CONCLUSION

This paper introduces an improved planar reconstruction method for vectoring 2D roof structures directly from a single image. Our method is built upon HEAT, and is enhanced by applying a feature pyramid Transformer and introducing a collaborative branch of semantic segmentation into primitives extraction. Compared with HEAT, the vertex and edge heat map F1-scores have risen by 2.0% on the VWB dataset. Qualitative evaluations also demonstrate that our method makes improvements over the existing state-of-the-art. Future research will continue to explore more efficient and effective training methods, such as the introduction of self-supervised learning and knowledge distillation.

5. REFERENCES

[1] W. Zhao, C. Persello, and A. Stein, "Extracting planar roof structures from very high resolution images using graph neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 187, pp. 34–45, 2022.

[2] B. Xiong, S. O. Elberink, and G. Vosselman, "A graph edit dictionary for correcting errors in roof topology graphs reconstructed from point clouds," *ISPRS Journal of photogrammetry and remote sensing*, vol. 93, pp. 227–242, 2014.

[3] N. Xue, T. Wu, S. Bai, F. Wang, G.-S. Xia, L. Zhang, and P. H. Torr, "Holistically-attracted wireframe parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2788–2797.

[4] F. Zhang, N. Nauata, and Y. Furukawa, "Conv-mpn: Convolutional message passing neural network for structured outdoor architecture reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2798–2807.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[6] J. Chen, Y. Qian, and Y. Furukawa, "Heat: Holistic edge attention transformer for structured reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3866–3875.

[7] N. Nauata and Y. Furukawa, "Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference," in *European Conference on Computer Vision*. Springer, 2020, pp. 711–726.