Supplementary Document of RNGDet++

Zhenhua Xu, *Student Member, IEEE*, Yuxuan Liu, *Student Member, IEEE*,
Yuxiang Sun, *Member, IEEE*, Ming Liu, *Senior Member, IEEE*, and Lujia Wang, *Member, IEEE*

I. Instance Segmentation

The instance segmentation head predicts the heatmap of the ahead road of the current vertex v_t. The instance segmentation head can help the network to better capture the information of road networks, which facilitates the training process and improves the reasoning ability of the network. The GT (Ground Truth) instance segmentation map marks the road ahead, which is a segment of the GT road. Suppose the agent current position is v_t, and the GT vertex in the next step is $(v_{t+1}^i)^*$ ($(v_{t+1}^i)^*$ is already on the GT road). Note that the GT instance segmentation is not the line connecting v_t and $(v_{t+1}^i)^*$. We first project v_t onto the GT road as v_t^*, and then use the road segment ahead connecting v_t^* and $(v_{t+1}^i)^*$ as the GT instance segmentation mask. In this way, the instance segmentation head can train the network to capture the correct road information when it deviates from the right track, which improves the final performance. An example is visualized in Fig. 1 to show how we define instance segmentation labels.

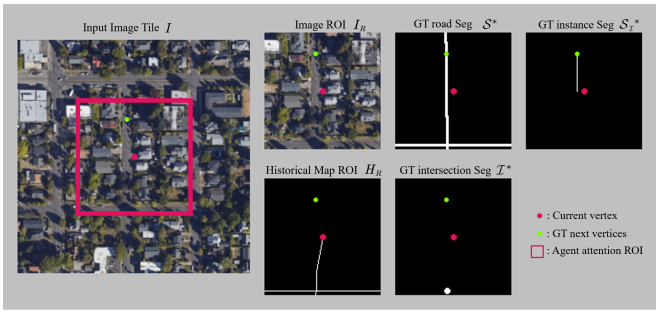


Fig. 1: Visualization of the GT instance segmentation mask (top right mask). The pink point is the current vertex v_t, and the green node is the ground-truth vertex in the next step $(v_{t+1}^i)^*$. We first project v_t onto the ground-truth road as v_t^*. Instead of connecting v_t and $(v_{t+1}^i)^*$, we use the segment connecting v_t^* and $(v_{t+1}^i)^*$ as the groud-truth instance segmentation of the road ahead of v_t. When the agent v_t (pink point) is away from the right track, the instance segmentation head still supervises the agent to capture the correct road information, which improves the final performance.

II. Training Data Sampling

We propose a BFS (Breath First Search) graph traversal algorithm as the imitation learning expert. The expert traversals the GT road network graph vertex by vertex. At each step, the expert generates a training sample $(I_R, H_R, \mathcal{S}^*, \mathcal{I}^*, \{[(\mathcal{S}_\mathcal{I})_{t+1}^i]^*\}_{i=1}^N, \{(v_{t+1}^i)^*\}_{i=1}^N)$, where I_R is cropped aerial image ROI, H_R is cropped historical image ROI, $\mathcal{S}^*$ is the GT road segment segmentation map, $\mathcal{I}^*$ is the GT road intersections segmentation map, $\{[(\mathcal{S}_\mathcal{I})_{t+1}^i]^*\}_{i=1}^N$ are the GT instance segmentation maps, and $\{(v_{t+1}^i)^*\}_{i=1}^N$ are the GT vertices in the next step. To enhance robustness, we add noise to the trajectory of the expert during the sampling phase. Some examples of sampled training data are visualized in Fig. 2.

III. Detailed Network Structure of RNGDet and RNGDet++

The idea of RNGDet++ that utilizes multi-scale features is quite similar to that of UNet [1] and FPN [2]. The backbone network predicts multiple feature tensors at different levels (e.g., in our work, four feature tensors f_1, f_2, f_3, f_4 at different levels are obtained by the ResNet backbone). RNGDet only uses the deepest layer (i.e., f_4) for vertex prediction, while tensors at other levels are ignored, which causes information loss. RNGDet++ can make the prediction based on the feature extracted at all four levels, which can better capture the deep feature of the input images to predict the vertex in the next step. Thanks to the multi-scale feature fusion, RNGDet++ can better process the input image and make more appropriate predictions than RNGDet. The network structures of different models are visualized in Fig. 3.

IV. Additional Visualizations

Additional visualizations of the comparison experiments on the city-scale dataset is shown in Fig. 4.

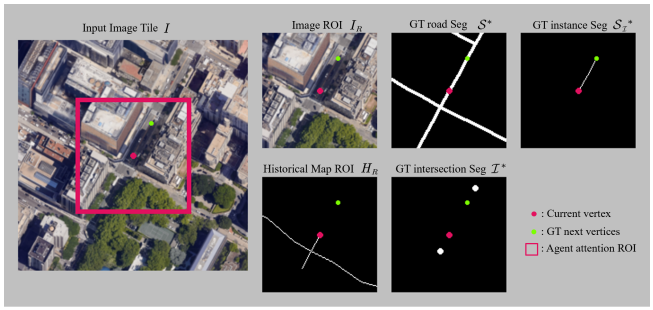
References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

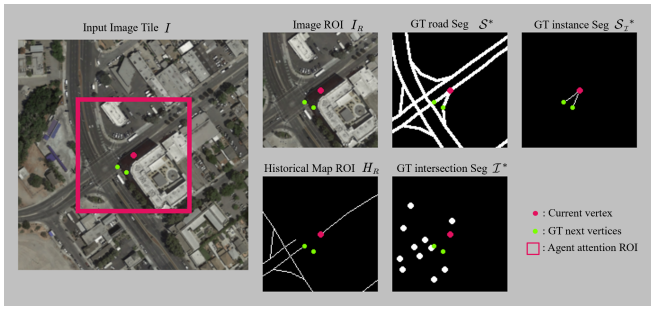
[2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[3] Z. Xu, Y. Liu, L. Gan, Y. Sun, X. Wu, M. Liu, and L. Wang, "Rngdet: Road network graph detection by transformer in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

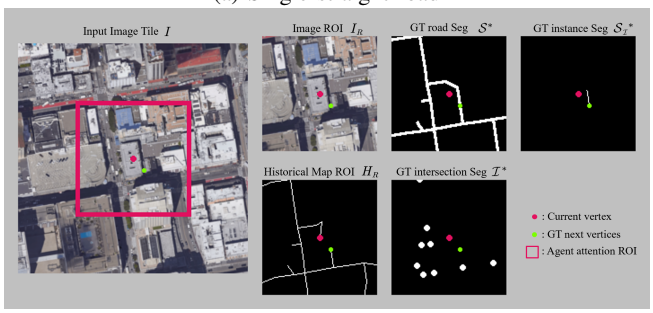
[4] S. He, F. Bastani, S. Jagwani, M. Alizadeh, H. Balakrishnan, S. Chawla, M. M. Elshrif, S. Madden, and M. A. Sadeghi, "Sat2graph: road graph extraction through graph-tensor encoding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 51–67.



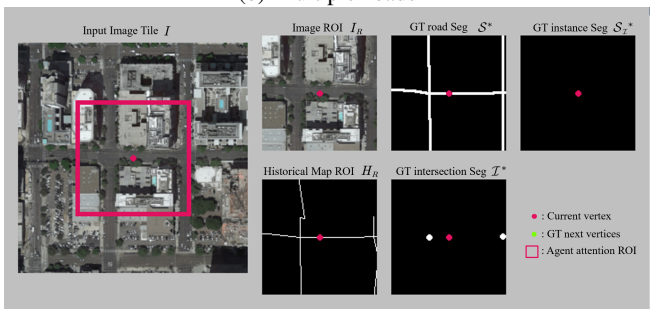
(a) Single straight road



(b) Multiple roads

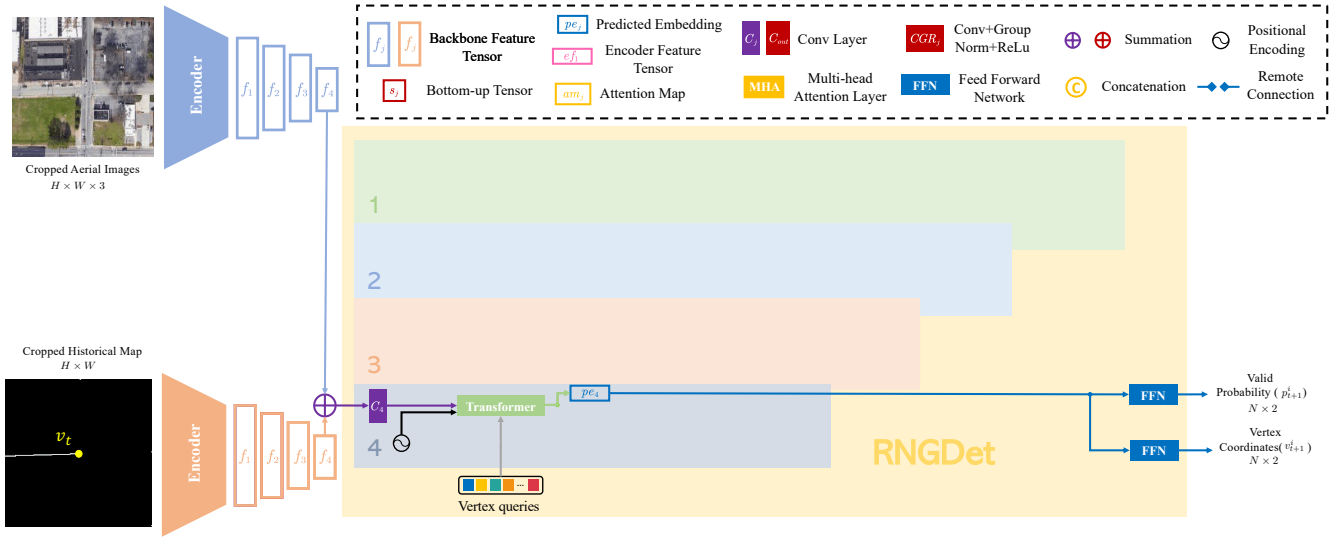


(c) Road loop

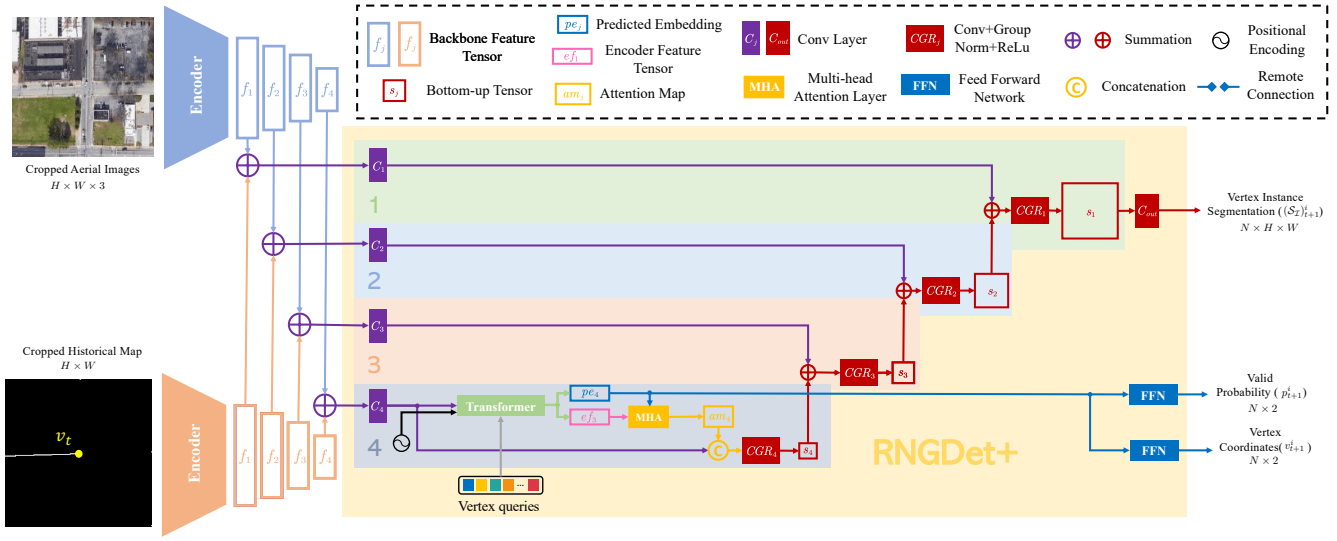


(d) Stop

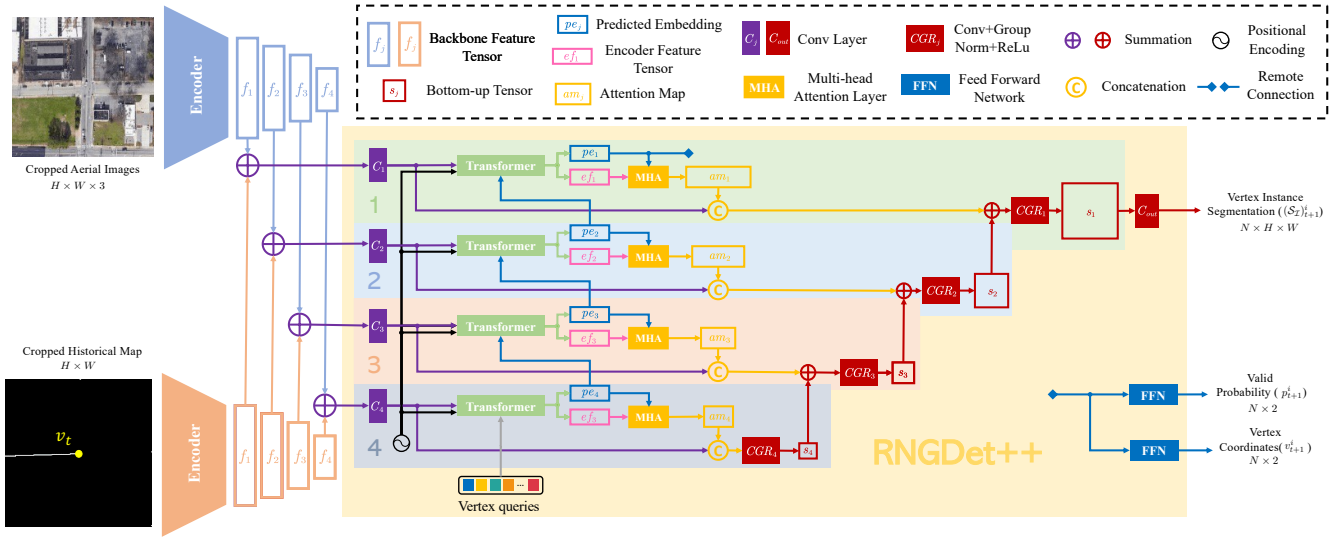
Fig. 2: Visualizations of sampled training data. For better visualization, we add the instance segmentation labels of all queries into a single mask (top right). Note that this is the visualization of the expert sampling process, not the inference results. (a) A sample on the straight road. (b) A sample on intersections. There are two vertices in the next step. (c) A sample on the loop road. The agent should predict and align with previously predicted vertices to close the road loop. (d) A sample on stop. There is no unexplored road ahead, thus the agent should stop. The proposed expert sampling algorithm can generate correct expert demonstrations/trajectories for imitation learning with high efficiency.



(a) RNGDet

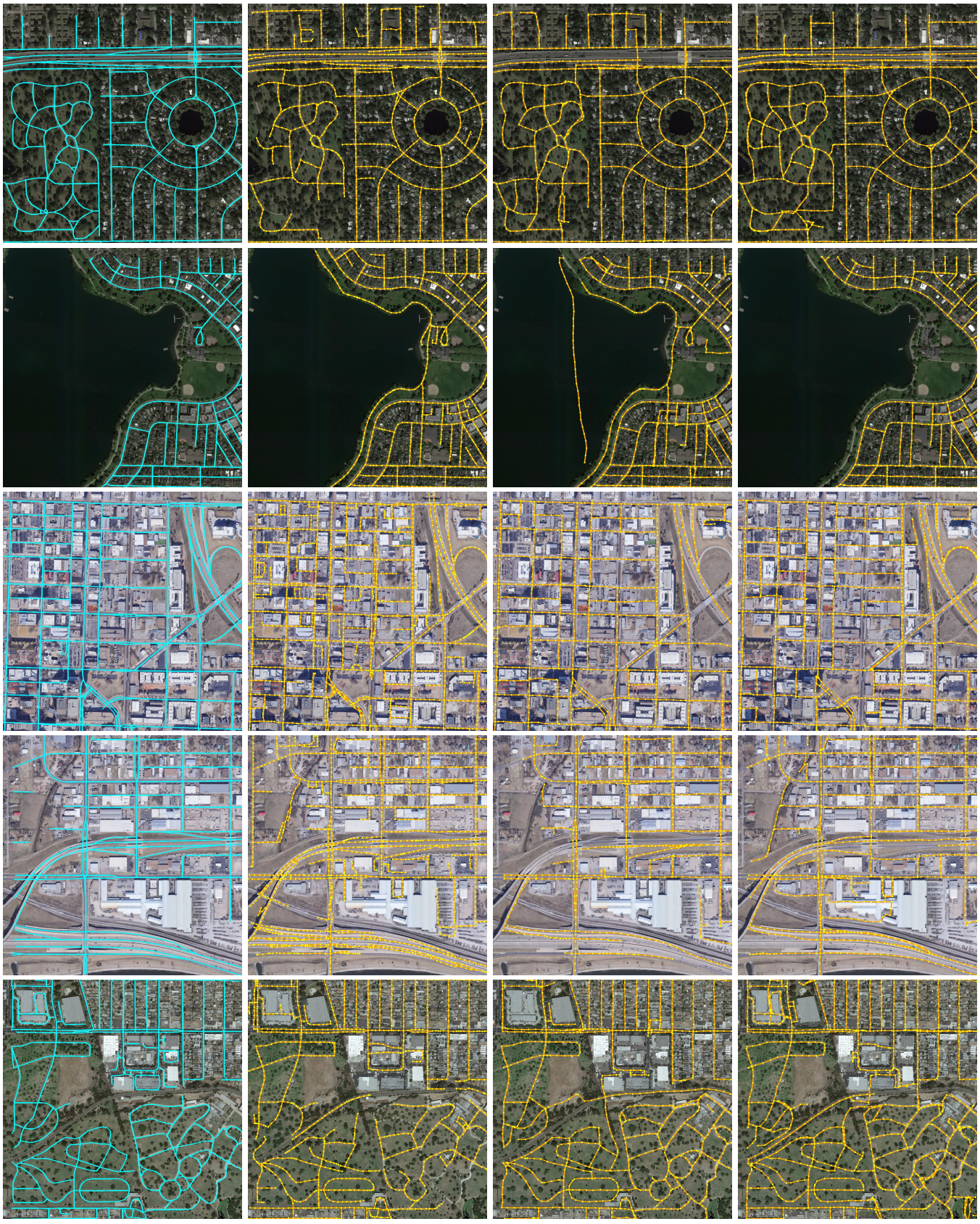


(b) RNGDet++ with a single feature layer



(c) RNGDet++

Fig. 3: Network structures of different approaches. (a) RNGDet [3]. RNGDet only leverages the deepest feature layer and does not have the instance segmentation head. (b) RNGDet++ with a single feature layer (the deepest layer). Compared with RNGDet, this model adds the instance segmentation module. (c) RNGDet++. It can utilize multi-scale features for road network detection, and can facilitate the training process by the instance segmentation head.



(a) Ground truth (b) Sat2Graph [4] (c) RNGDet [3] (d) RNGDet++

Fig. 4: Qualitative visualization. (a) Ground truth road networks (Cyan lines). (b) Road network graph detected by Sat2Graph. (c) Road network graph detected by RNGDet. (d) Road network graph detected by RNGDet++. For (b)-(d), yellow points represent graph vertices and orange lines represent graph edges. For the visualization, it is found that RNGDet++ can output road network graphs with more accurate structure and correctness compared with previous works. This figure is best viewed in color. Please zoom in for details.