# A Novel Brain-Like Navigation Based on Dynamic Attention with Modified Unet

Yu Zhang[1,2] and Xiyuan Chen[1,2(✉)]

[1] Key Laboratory of Micro-Inertial Instrument and Advanced Navigation Technology, Ministry of Education, Southeast University, Nanjing 210096, China
chxiyuan@seu.edu.cn

[2] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China

**Abstract.** In cognitive navigation system, animals show an inborn ability of spatial representations and correct self-positioning errors at every fired cell. Inspired by navigation mechanism of animals, we propose a novel strategy to improve the navigation accuracy of brain-like navigation based on UAV. Firstly, we employs encoder-decoder structure based on Unet to solve semantic segmentation tasks. Unet are able to encoder detailed information of images by constantly pooling and upsampling operations with less training parameters, while it often ignores high-level spatial information. Hence, we propose "dynamic attention with modified Unet" structure, which learns high-level information maintaining less training parameters. Specifically, multi-scale atrous convolutions are adopted in dynamic modules between encoder and decoder to extract features at different resolution. Secondly, the pixels with maximum probability segmentation are extracted, and they will be mapped to satellite map to obtain actual position coordinate of UAV. Finally, positioning errors are corrected at each place cells in the brain-like navigation of UAV. Our results show that proposed segmentation model improve performance by 9.64% compared with conventional Unet, and the positioning accuracy is improved by 90.52%.

**Keywords:** Semantic segmentation · Dynamic attention · Multi-scale atrous convolution

## 1 Introduction

With the development of artificial intelligence technology, next-generation navigation devices are endowed with self-correction abilities of positioning to fulfil the demands of applications. The unmanned aerial vehicle (UAV) is an emerging technology where positioning accuracy and robustness are critical for safe guidance and stable control. Conventional UAV navigation system is dominated by the loose-coupled with Inertial Navigation System (INS) and Global Navigation System (GPS). However, UAVs cannot maintain high accuracy because of GPS signal of low quality when flying through complex environments, such as urbans, canyons and electromagnetic interferences, and


W. Fu et al. (Eds.): ICMTEL 2021, LNICST 387, pp. 528–540, 2021.
https://doi.org/10.1007/978-3-030-82562-1_52

the results measured by INS will be divergent over time [1]. Thus, we need to investigate a steady strategy that obtain high position accuracy without relying on traditional measurement devices.

"Place cell" property of hippocampus is found as it have a firing rate when rodent is at a specific place in extensive environment which inspired the proposal of brain-like navigation strategy. Furthermore, there is a phenomenon that mammals such as rats or primates show an inborn ability of spatial representations, and they take advantage of it to implement space navigation [2]. More precisely, once the rodents reach at a specific dot, the cell node on hippocampus will be activated, and the actual position information will be recorded as reference. Also, the path integration of rats can be reset to acquire new information when they are placed in a familiar environment, so that the accuracy of bionic navigation can be remarkably improved by eliminating accumulated errors. Thus, the challenge of this navigation lies in whether it can find correct scenes in "Memory" to match current visual scenes [3].

For one thing, the typical application of scene recognition is on visual navigation. Images captured by camera are used to compare with referenced images to obtain actual navigation information. For another, convolutional neural networks is widely applied in a number of classification tasks, where every image output a predicted single class label. However, we need to focus on localization information in USV-assisted visual navigation. Hence, semantic segmentation with the aim to assign semantic labels to every pixel is proposed. As we all know that several network backbones with Fully Convolutional, U-net modules have shown striking improvements over strategies based on hand-crafted feature extraction [4]. Obviously, these modules have three drawbacks. First, consecutive pooling operations or convolution striding operations may impede dense prediction tasks, even though it allows deep learning model to extract feature representations. Second, we always tend to pay more attention to capture high-level semantic information without fusing low-level feature.

In this work, a novel semantic segmentation network combined with localization extraction is proposed to assist brain-like navigation based on UAV. Firstly, we design upon an elegant architecture, the so-called "U-net", which based on encoder-decoder structure. We modified and extend this architecture that can fuse high-level semantic information and low-level feature effectively with few training images and "dynamic attention" modules. We maintain a large number of feature channels in upsampling process, which allows model to propagate information to higher resolution layers. To improve the response capability of dense space, we add several "dynamic attention" modules consisting of atrous convolutions with different rates between downsampling part and upsampling part, which allows the seamless segmentation of images with different scales. In addition, this special attention module predict the pixels in the border region of the image precisely.

As for the UAV-assisted brain-like navigation system, the location information corresponding to the feature points on segmentation image is extracted and fed back to correct the positioning of INS to improve the navigation accuracy. Furthermore, we endow the positioning errors with 'self-correction' property termed as the navigation error correction (NEC) module so that navigation accuracy can be improved with increasing

familiarity of the flight trajectory. Subsequently, the proposed strategy rectifies positioning errors periodically with low computation cost as the parameters trained in network is very less, which is critical to meet the requirements of real time analysis.

In summary, our contributions are as follow:

- We propose a novel semantic segmentation network combined with localization extraction which employs modified Unet as a backbone.
- Multi-scale dynamic attention modules consisting of atrous convolutions are added between encoder-decoder structures, which propagate information to higher resolution layers instead of focusing on low-level details. Also, the running time of our proposed model will not increase too much.
- Cell models in biological brain-like system are mapped to actual trajectory to assist UAV navigation.
- The actual position coordinates corresponding to pixel in segmentation images are extracted to assist brain-like navigation.

## 2 Related Work

With the discovery of the role of grid-like cells, place cells and several brain cognitive navigation cells, brain navigation technology is investigated widely, which provides a theoretical basis to study brain-like navigation of UAVs in complex flight environments.

In practice level, a novel brain-like model is inherited from a normal brain-based device (BBD) which helps understand how rodent cognition and behaviour work [5], which means that we can build relationship between the spatial mapping property of entorhinal cortex (EC) and location nodes of the actual UAV trajectory directly. It is noted that 'place cells' property of hippocampus can be mapped into location nodes in actual trajectory measured by INS, which can be seen in Fig. 1.

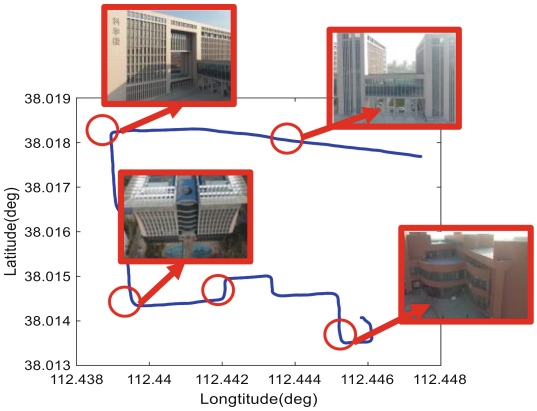


**Fig. 1.** The schematic diagram of place cells in actual trajectory.

Differ from traditional positioning strategy which applies integrated measurement, we rely on semantic segmentation network combined with localization extraction to finish navigation tasks. It has been proved that the global features or contextual interactions are vital in classifying pixels for semantic segmentation [7]. For semantic segmentation tasks, we consider three challenges. First, consecutive pooling operations or convolution striding operations may impede dense prediction tasks, even though it allows deep learning model to extract feature representations. Second, multi-scale images have to be unified into the same size, as the size of weight matrix in output layer is default. Hence, image pyramid, spatial pyramid pooling and atrous convolution methods is applied to multi-scale inputs to capture context at several ranges [8]. Third, too much parameters is calculated as several consecutive convolution operations, which give a pressure on GPU memory and cannot meet the requirements of real time analysis.

## 3 Methods

In this section, we discuss the proposed model structure for semantic segmentation. Then, we demonstrate that how the segmented image is used for brain-like navigation with UAVs.

### 3.1 Cell Model Based on Brain-Like System

We propose a novel strategy for UAV navigation under the scheme of brain-like model, which mainly contains three phases. In the first phase, an intelligent brain-like model is applied it to UAV navigation tasks inspired by rodent navigation mechanism. As we all know, 'place cell' and 'head direction cell' may have a high firing rate when rodents arrive at a particular location. Thus, several location cell nodes can be set in advance. Inputs to the brain-like model come from a camera, and it is used to record flight environment and collect scene images. Then the output from brain-like model goes to semantic segmentation with aim to extract the coordinates of the centroid location of image corresponding to the UAV. In the last phase, we establish a linear error equation based on the time series to model the accumulated errors between two cell nodes. Furthermore, the positioning errors accumulated at current flight can be compensated by the last error calculation, and the error model under the current trajectory is established at the same time to achieve the purpose of correcting the drift error in the next step. We endow the positioning errors with 'self-correction' property termed as the NEC module so that navigation accuracy can be improved with increasing familiarity with the flight trajectory.

### 3.2 Dynamic Attention Mechanism

The network structure is illustrated in Fig. 2. The whole structure is inherited from classical U-net, which consists of the encoder-decoder benchmark. To be concrete, an encoder module gradually increases receptive field by reducing the size of feature maps with convolution operations, capturing higher semantic information. Meanwhile, a decoder module recovers the spatial information. However, traditional U-net structure focus too much on detailed information by constantly downsampling operations, so that it only captures

deeper feature instead of higher spatial information. Hence, we explore the multi-scale atrous convolutions between encoder and decoder during upsampling process, which is denoted as "dynamic attention" module. The encoder module encodes multi-scale spatial information with these dynamic atrous convolutions at multiple scales.

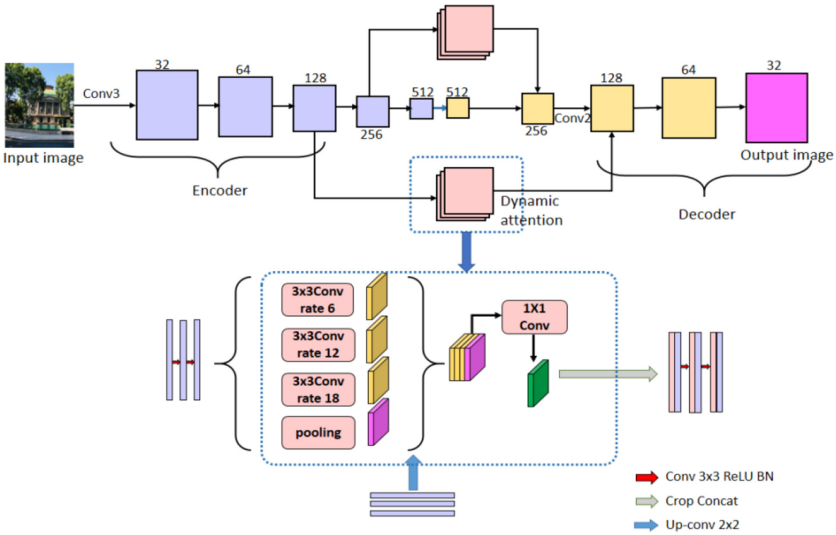


**Fig. 2.** The structure of proposed segmentation model.

We set the initial feature channels as 32. It comprises of two repeated 3 × 3 convolutions with no-padding, each followed by a rectified linear unit (ReLU) and batch normalization in the encoder process. In addition, 2 × 2 max pooling operations with stride 2 are chosen for downsampling. It is noted that we double the number of feature channels at each downsampling step. Secondly, 2 × 2 convolutions with stride 2 are conducted for upsampling, and the feature channels are halved at each step. Thirdly, features produced by upsampling is concatenated with the corresponding cropped low-level features. Also, the proposed dynamic attention modules are introduced into the copy of features produced by downsamping before cropping. Here, *output_stride* is denote as the ratio of image spatial resolution to output resolution, and we adopt rate = 6, 12, 18 to the copy of features by appling strous convolution correspondingly. In addition, 1 × 1 convolution is added to reduce the feature channels and parameters. By introducing this module, we can extract the features at different resolution. At the final layer, a 1 × 1convolution is used to map each upsampled features to the default number of classes.

### 3.3 Error Compensation Model

The core module for brain-like navigation with UAV is illustrated in Fig. 3. We construct the semantic segmentation methods under the frame of brain-like system to eliminate the accumulated errors between two place cell nodes. From the Fig. 3 we can see that the place cell is simulated just shown as A, B, C, and D.

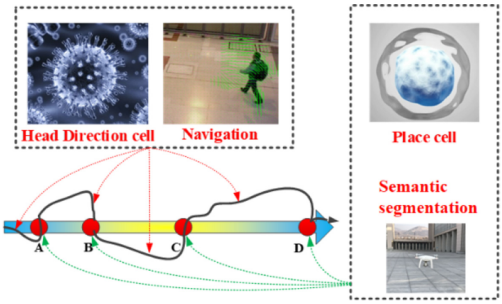


**Fig. 3.** The structure of the NEC module.

In the actual UAV flight process, due to the drift error of the INS, the error will linearly diverge with time in a short time, so it is necessary to model and analyze the error divergence between two cell nodes. Therefore, the flight trajectory between the two location cell nodes can be approximated as a linear motion for simplify. Then, a linear error equation with time series is established under the scheme of the navigation. The positioning errors at current flight can be compensated by the last flight, and the error model under the current trajectory is established at the same time to achieve the purpose of correcting the drift error in the next step, which can be seen in (1) and (2). In each process of compensating the current trajectory error, it is necessary to adjust the coefficient value of the linear function to achieve self-correction compensation for the error.

$$error(i) = k * (trj_{\mathrm{Ref}}(i) - trj_{\mathrm{INS}}(i)) + b \tag{1}$$

$$trj_{\mathrm{INS}}(i+1) = error(i) + trj_{\mathrm{INS}}(i) \tag{2}$$

Where $error(i)$ represents flight error function, which can be simplify as linear motion; $k$, $b$, represent coefficients of linear function; $trj_{\mathrm{Ref}}(i)$, $trj_{\mathrm{INS}}(i)$ represent the reference and INS trajectory of the i-th flight, respectively; $trj_{\mathrm{INS}}(i+1)$ represents the INS trajectory of the i + 1-th flight.

Gradually, we endow the positioning errors with 'self-correction' property and navigation accuracy can be improved with increasing familiarity with the flight trajectory, just like the work mechanism of path integrator of rats in brain-like system.

## 4 Training

In this section, we discuss the training details of our proposed segmentation model. Our implementation is built on Pytorch.

The proposed segmentation model is evaluated on PASCAL VOC2012 semantic segmentation benchmark, which contains 21 classes. It is noted that the origin dataset covers 1464 samples for training, 1449 for validation and 1456 pictures for test, respectively. One background class and 20 object classes is calculated.

**Initialization.** The features encoded are downsampled by a factor of 2 and then concatenated with the corresponding decoded features. We set initial feature channels as 32 with two repeated $3 \times 3$ convolutions (no-padding) and a rectified linear unit (ReLU). In addition, $2 \times 2$ max pooling operation with stride 2 is chosen for downsampling. Also, each modified module on top of basic structure of Unet all includes batch normalization to reconstruct parameters. Basically, we set batch size $= 8$, and the batch normalization parameters are trained with decay $= 0.9997$. Accordingly, we set learning rate as simple 0.001 and use a momentum (0.9), which can combine more trained samples to update in the current optimization step.

**Dynamic Attention.** The core of dynamic attention modules is multi-scale atrous convolutions, which are added between downsamling and upsampling to capture higher spatial information. The characteristic of atrous convolution is that it can modify the filter's field-of-view and control the dense of feature response adaptively. Based on some evaluation discussed in [8], the *output_stride* $= 16$ sacrifices some accuracy to obtain faster calculation speed compared with *output_stride* $= 8$ since the inter-mediate feature maps are spatially four times smaller. However, it has been verified that 16 *output_stride* strikes the best trade-off between speed and precision. In addition, the performance when setting *output_stride* $= 1$ is equal to that without any operations for features. Thus, we set the *output_stride* $= 6$, 12, 18, and all output of atrous convolution operations are concatenated. Meanwhile, we apply $1 \times 1$ convolution after concatenation to reduce the number of feature channels (e.g. 512 or 1024), which can make the training easier. We apply $1 \times 1$ convolution before the finally output to fuse all feature and match the target number of classes. The reason why we call this multi-scale module as "dynamic attention" is that it can be added between encoder and decoder to increase the sensitivity to high-level spatial information and overcome the defect of being only sensitive to detailed features with Unet. Also, the whole training parameters will not increase even adding this module.

## 5 Experiment Evaluation

To evaluate the proposed semantic segmentation-assisted brain-like navigation, two specific experiment is proceeded for validation.

### 5.1 Experiment 1

In experiment 1, the trajectory of UAV is shown in Fig. 4 and the start coordinate is 32.0302, 118.8792(deg). The length of the trajectory is approximately 10 km. From the Fig. 4 we can see that four location cell nodes corresponding to specific scenes are marked, which can be used as ground truth. The performance of segmentation can be verified at every location cell node. In order to improve the segmentation accuracy, we modify the training set by adding some specific scene images captured by camera equipped with UAV, and the specific images at each location node are used for validation. In order to verify the effectiveness of proposed model, different flight motions including long straights, turning and sudden accelerations are all performed.

Firstly, we verify the performance of semantic segmentation. The dataset for validation is VOC2012 which has been mentioned before. Figure 5 represents the compared segmentation results on validation set between the proposed multi-scale dynamic attention model with modified Unet network backbone and the conventional Unet. Figure 6 represents visualization results on *val* set with proposed segmentation model.

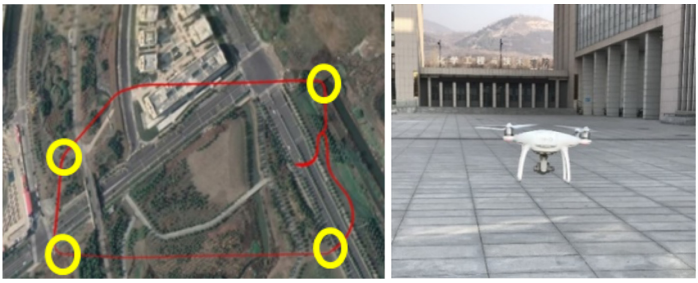

**Fig. 4.** (1) The trajectory of UAV; (2) The material object of UAV.

The results demonstrate that the detail of images is segmented more comprehensive by applying multi-scale atrous convolution operations. The performance of different models dealing with segemention is shown in Table 1. Unet with 32 employing the proposed multi-scale dynamic attention (multi-scale atrous convolutions), attains the performance of 85.91% on the validation set. We notice that decreasing the initial feature channels of Unet and adding multi-scale atrous convolutions between encoder and decoder inevitably improve the performance by 9.64%. Furthermore, this proposed model performs better than RetNet-101 banckmark with less training parameters. As we all know that conventional Unet is effective to deal with medical images, thanks to atrous convolution, this proposed model also obtains better segmentation results.

**Table 1.** The results of different segmentation methods. Unet with 32 refers to the initial feature channels of Unet model is 32; Unet with 64 refers to the initial feature channels of Unet mdoel is 64.

| Backbone | Dynamic module (multi-scale atrous convolution) | Decoder | mIOU |
|---|---|---|---|
| Unet with 32 | ✓ | ✓ | 85.91% |
| Unet with 64 | | ✓ | 77.62% |
| ResNet-101 | | ✓ | 84.7% |

In order to verify the navigation accuracy with semantic segmentation, we conduct an actual flight experiment. All images captured at each location node during flight are recorded for validation. The experiment system is conducted on a hybrid MEMS-INS/GPS platform, where an Ublox NEO-M8T GPS receiver with a STIM202 and 1521L integrated IMU are connected to a data acquisition module running Windows

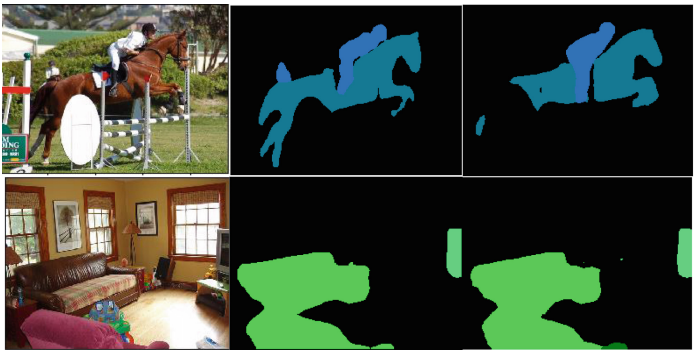


**Fig. 5.** (a) Segmentation performance based on the proposed multi-scale dynamic attention model with modified Unet; (b) Performance based on conventional Unet.

**Fig. 6.** Visualization results on *val* set.

10 operation system. A NovAtel ProPak6 receiver is used for reference offline. The sampling frequency of INS is set as 200 Hz. The experimental setup is shown in Fig. 7, and specific parameters of sensors are given in Table 2.

As the flying height of the UAV is very high, the proportion of target features in the image will be reduced, which will cause some difficulties in segmentation. Thus, we

scale the size of input image proportionally at the expense of resolution. The visualization results at location node are shown in Fig. 8. As we all know that each pixel is categorized as one class with the highest probability output, and each pixel will be masked in the meantime, which makes up the outline of the final target object. Hence, the pixels with maximum probability segmentation are extracted, and the actual location on the map is calculated as the current position coordinate of UAV. Precisely, we select the pixel block corresponding to the maximum probability output as the referenced location, and this location will be mapped to satellite map to obtain actual position coordinate, which also shown in Fig. 8.

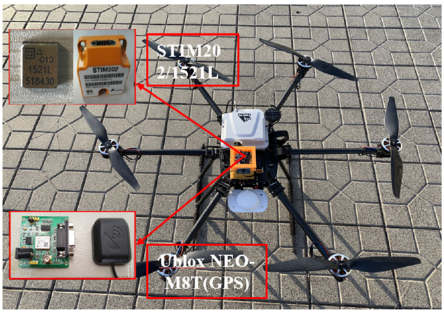


**Fig. 7.** The experimental setup based on experiment 2.

**Table 2.** The specific parameters of sensors.

| | | |
|---|---|---|
| Gyroscope (STIM202) | Bias | 0.5°/hr |
| | Scale factor | 200 ppm |
| | Random walk | $0.2°/\sqrt{hr}$ |
| Accelerometer (1521L) | Calibration error | 0.5–1% |
| GPS (Ublox NEO-M8T) | Position accuracy | 2.5 m |
| | Velocity accuracy | 0.05 m/s |
| | Time accuracy | 60 ns |
| GPS (NovAtel ProPak6) | Position accuracy | 1cm + 1 ppm |
| | Velocity accuracy | 0.03 m/s |
| | Time accuracy | 20 ns |

Subsequently, in order to improve positioning accuracy and evaluate the fault-tolerant of UAVs, we verify the performance of proposed NEC module with designed segmentation model. As we discussed before, NEC module corrects accumulated errors by iterative compensation at each flight process. Hence, we conduct three flight experiment, the flight trajectory is shown in Fig. 9(a), flight altitude is 38.4 m, and the flight trajectory of UAV is measured by a NovAtel ProPak6 receiver, which is used for absolute reference.

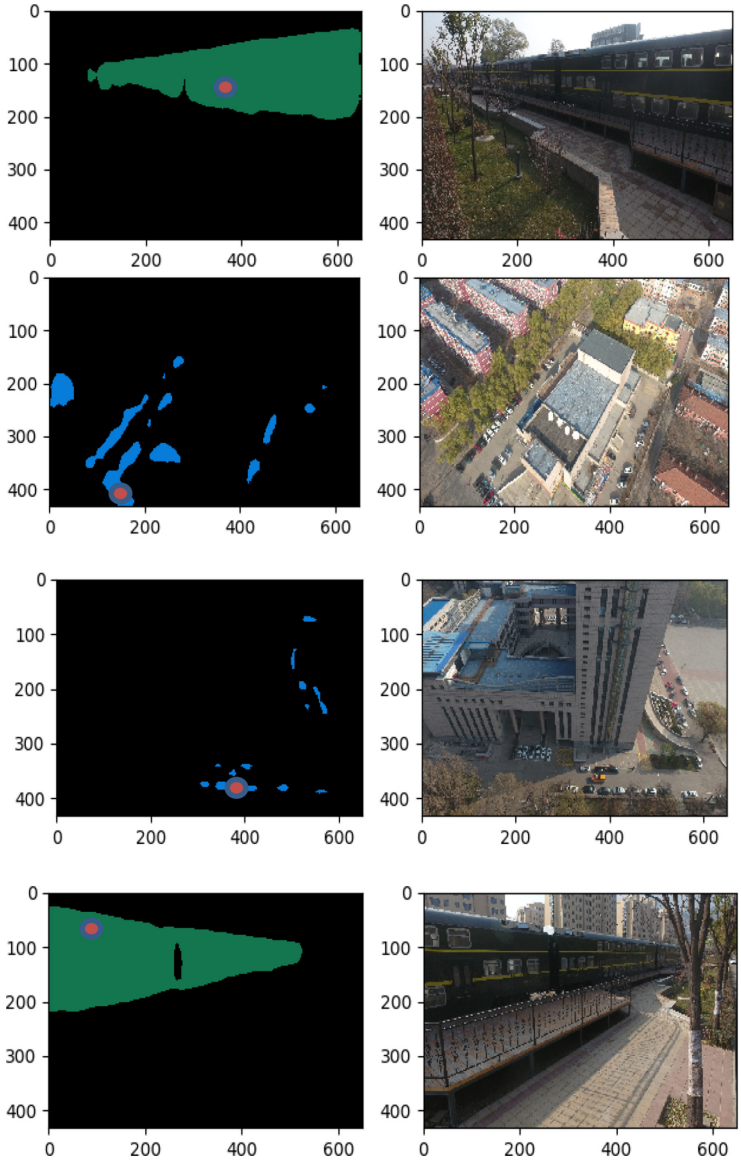


**Fig. 8.** The visualization results at location node.

Two-dimensional trajectory corresponding to Fig. 8 is shown in Fig. 9(b), and three place cells are marked in results. The Root Mean Square (RMS) results of position errors are shown in Table 3. From the results shown in Fig. 9(b) we can see that the positioning error can be compensated at current flight after the navigation error was compensated last flight, and the whole trajectory is closer to reference trajectory. From the Table 3 we can see that the RMS errors decrease from 206.44 m to 1.93 m at the third flight, which

are improved by 90.52%. Eventually, INS cumulative drift error re-accumulates from 0 after correction at each node by segmentation, so the INS can maintain high accuracy for a certain period of time after correction calculation. Eventually,.

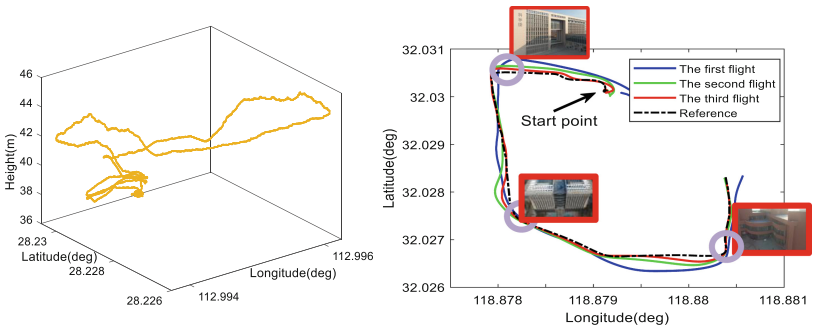


**Fig. 9.** (a) Flight trajectory; (b) The compensated results at each node with NEC module.

**Table 3.** The RMS of position errors

| NEC module | The RMS of position errors (m) |
|---|---|
| The first flight | 206.44 |
| The second flight | 20.36 |
| The third flight | 1.93 |

## 6 Conclusions

In this work, with the aim to improve brain-like navigation accuracy of UAV, our proposed model "dynamic attention with modified Unet" employs the encoder-decoder where dynamic attention is used to capture rich spatial information without increasing extra much training parameters. Precisely, multi-scale atrous convolutions, as the core of dynamic modules are adopted between encoder and decoder to extract features at different resolution. The pixels with maximum probability segmentation are extracted, and the actual location on the map is calculated as current position coordinate of UAV. Finally, our results show that proposed segmentation model sets a start-of-art performance on VOC2012 datasets compared with conventional Unet, and the final navigation accuracy is improved by NEC module combined with maximum probability of pixel extraction.

## References

1. Atia, M.M., Waslander, S.L.: Map-aided adaptive GNSS/IMU sensor fusion scheme for robust urban navigation. Measurement **131**, 615–627 (2019)
2. Huajin, T., Weiwei, H., Aditya, N., Rui, Y.: Cognitive memory and mapping in a brain-like system for robotic navigation. Neural Networks **87**, 27–37 (2017)
3. Shen, C., Liu, X.C., Cao, H.L., Zhou, Y.C.: Brain-Like navigation scheme based on MEMS-INS and place recognition. Appl. Sci. **9**, 1708 (2019)
4. Jiawang, B., Wenyan, L., Yasuyuki, M.: GMS: grid-based motion statistics for fast, ultra-robust feature correspondence. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2528–2837 (2017)
5. Krichmar, J.L., Aitz, N.D., Gally, A.J., Edelman, G.M.: Characterizing functional hippocampal pathways in a brain-based device as it solves a spatial memory task. Proc. Natl. Acad. Sci. **102**, 2111–2116 (2004)
6. Weijer, V.D., Gevers, J.T., Gijsenij, A.: Edge-based color constancy. IEEE Trans. Image Process. **16**, 2207–2214 (2007)
7. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
8. Chen, L.C., George, P., Florian, S., Hartwig, A.: Rethinking atrous convolution for semantic image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)