Toward Integrating Semantic-aware Path Planning and Reliable Localization for UAV Operations

†Thanh Nguyen Canh[1,2], †Huy-Hoang Ngo[2], Xiem HoangVan[2] and *Nak Young Chong[1]

[1]School of Information Science, Japan Advanced Institute of Science and Technology
Ishikawa 923-1292, Japan ({thanhnc, nakyoung}@jaist.ac.jp) * Corresponding author
[2]University of Engineering and Technology, Vietnam National University
Hanoi 10000, Vietnam (ngoh52180@gmail.com, xiemhoang@vnu.edu.vn) † Equal contributions

Abstract: Localization is one of the most crucial tasks for Unmanned Aerial Vehicle systems (UAVs) directly impacting overall performance, which can be achieved with various sensors and applied to numerous tasks related to search and rescue operations, object tracking, construction, etc. However, due to the negative effects of challenging environments, UAVs may lose signals for localization. In this paper, we present an effective path-planning system leveraging semantic segmentation information to navigate around texture-less and problematic areas like lakes, oceans, and high-rise buildings using a monocular camera. We introduce a real-time semantic segmentation architecture and a novel keyframe decision pipeline to optimize image inputs based on pixel distribution, reducing processing time. A hierarchical planner based on the Dynamic Window Approach (DWA) algorithm, integrated with a cost map, is designed to facilitate efficient path planning. The system is implemented in a photo-realistic simulation environment using Unity, aligning with segmentation model parameters. Comprehensive qualitative and quantitative evaluations validate the effectiveness of our approach, showing significant improvements in the reliability and efficiency of UAV localization in challenging environments.

Keywords: Localization, Navigation, UAVs, Semantic Segmentation, Path Planning.

1. INTRODUCTION

In recent years, unmanned aerial vehicles (UAVs) have emerged as a significant research field and a top priority in the robotics industry, including tunnel navigation [1], surveillance, search operations [2], terrain mapping [3], disaster relief and accident response [4]. UAVs offer several advantages, such as simple structure and flexible flight capabilities. One of the key challenges in deploying UAVs effectively is ensuring accurate localization and navigation, particularly in complex and dynamic environments where traditional sensors like GPS and Inertial Measurement Units (IMUs) may be unreliable and unavailable. For example, IMUs often deliver suboptimal results for UAVs due to their difficulty adapting to environmental factors such as wind and air resistance. Similarly, GPS signals can be disrupted in areas with tall buildings or dense forests. The ability of UAVs to autonomously navigate and localize is therefore crucial to their operational success and safety.

As UAV applications become more widespread, ensuring the stability of UAV self-positioning has emerged as an important concern [5]. However, both real-time localization and mapping remain unresolved issues, because commonly used sensors can fail under adverse weather conditions [6], [7], or in the presence of mountains, tall buildings or water [8]. In these scenarios, UAVs require reliable methods to prevent localization system failures and determine optimal flight paths to fly toward their destination. Visual Simultaneous Localization and Mapping (V-SLAM) [9], [10] [11] and multi-sensor fusion [12] have emerged as potential solutions to address this challenge. However, these methods face significant performance issues, such as reduced robustness and accuracy, which are greatly affected by navigation environmental conditions. Typically, V-SLAM's accuracy degrades significantly in the presence of dynamic objects and areas lacking texture or with specular surfaces (*e.g.* oceans and lakes). Therefore, UAV localization and navigation systems often struggle in unstructured environments due to insufficient feedback information. Consequently, integrating spatial awareness capabilities could open up the potential to improve the accuracy of UAV localization and navigation systems.

On the other hand, semantic segmentation models [13] [14] [15] have shown promising performance using RGB images from monocular cameras, paving the way toward integrating semantic segmentation for UAVs. Current research focuses on segmenting environmental objects, particularly utilizing 3D reconstruction to recover shapes of occluded or partially visible dynamic objects. By leveraging semantic awareness, UAVs can detect and respond to environmental factors such as terrain, moving objects, and weather conditions. This enhances self-protection and collision avoidance capabilities, while improving navigation and localization accuracy and performance.

Additionally, linking perception with SLAM [16] [17] to integrate semantic information, create a semantic map, and enhance localization performance has shown the potential to address the problem of active perception. However, these methods are hindered by the high computational complexity, therefore the implementation in outdoor environments is challenging. The most similar work to ours was committed by Bartolomei *et al.* [18], propos-

This work was supported by the Asian Office of Aerospace Research and Development under Grant/Cooperative Agreement Award No. FA2386-22-1-4042.

ing a method that applies semantic segmentation to enhance the localization quality of UAVs by designing a perception-aware navigation system based on the VTNet model, along with the A* Kinodynamic and B spline. In this paper, we proposed a reliable localization system for UAVs based on semantic segmentation, which integrates semantic segmentation information into a UAV path-planning framework to evaluate the quality of candidate areas for localization systems. The main contributions of this work are summarized as follows:

- A semantic-aware localization system for UAVs in challenging environments.
- A hierarchical planner that integrates the Dynamic Window Approach (DWA) algorithm with a cost map.
- Demonstration of the performances of our proposed system in active perception through photo-realistic simulations.

The remainder of this paper is organized as follows: Section 2 presents our proposed system based on the semantic segmentation and path planning algorithm. The experiments conducted and the analysis of the results are detailed in Section 3. Finally, Section 4 concludes the paper with discussion of future work.

2. METHODOLOGY

Our proposed pipeline is illustrated in Fig. 1, which takes the RGB image as input and progressively navigates to the goal based on semantic information to avoid unreliable localization areas. To achieve this, the RGB images undergo initial processing via key frame decision to determine whether semantic segmentation in this frame is necessary or not (Section 2.1). Subsequently, we introduced an efficient semantic segmentation to extract semantic masks from individual frames and calculate the cost map (Section 2.2). The localization module determines the UAV's pose, which can be estimated based on visual odometry, GPS, IMUs, or integration. Finally, a hierarchical planner is performed that integrates the Dynamic Window Approach (DWA) algorithm with a cost map to provide a potential trajectory to achieve the goal and avoid unreliable localization areas (Section 2.3).

2.1 Key Frame Decision

To ensure the ability of the real-time performance of our proposed system, we first introduce a keyframe decision architecture as shown in Fig. 1 to evaluate the usefulness of RGB images. Additionally, semantic segmentation is a challenging task that may require considerable time to execute and infer results, leading to non-continuous UAV flight while the segmentation of all frames is unnecessary. Therefore, our key frame decision module determines whether the image contains sufficient or rich semantic information to be fed into the semantic segmentation module, which requires much less execution time than the semantic segmentation module.

To achieve this goal, we segment the image into distinct regions to capture comprehensive information from various areas. This segmentation is akin to the patching process utilized in contemporary vision transformer models, where the input image $\mathbf{I}$ of size $H \times W$ is divided into non-overlapping smaller patches of size $P \times P$. Mathematically, this segmentation can be expressed as

$$\mathbf{I} = \left\{ \mathbf{I}_{i,j} | i \in [1, \frac{H}{P}], j \in [1, \frac{W}{P}] \right\} \tag{1}$$

where $\mathbf{I}_{i,j}$ denotes the patch located at the $i-th$ row and $j-th$ column of the patch grid.

Subsequently, data from each patch is globally average pooled to condense the information, with the global average pooling operation for a given patch $\mathbf{I}_{i,j}$ computed as:

$$GAP(\mathbf{I}_{i,j}) = \frac{1}{P^2} \sum_{u=1}^{P} \sum_{v=1}^{P} \mathbf{I}_{i,j}(u,v) \tag{2}$$

The pooled image is then flattened into a one-dimensional vector for further processing, described as $F(I) = concat\big(GAP(\mathbf{I}_{1,1}, GAP(\mathbf{I}_{1,2}, \cdots, GAP(\mathbf{I}_{\frac{H}{P}, \frac{W}{P}})\big)$. This vector is passed through a fully connected layer for evaluation, modeled by $\mathbf{y} = \sigma(\mathbf{W}\mathbf{x} + b)$, where $\mathbf{W}$ is the weight matrix, b is the bias vector, and σ denotes the activation function. The architecture ensures meticulous assessment of pixel value distribution across image regions. By employing only a patching layer and a fully connected network, the system achieves minimal execution time, enabling real-time operation. This module is crucial for meeting the system's requirements, ensuring efficient and accurate UAV localization by leveraging the robustness of semantic segmentation and enhancing computational efficiency, making it suitable for deployment in dynamic and complex environments.

2.2 Semantic Segmentation

After selecting the keyframe, semantic segmentation plays a vital role in extracting meaningful areas of information from the surrounding environment. Prior research, including PSPNet [19] and DeepLab [20] has demonstrated effective semantic segmentation using pretrained backbone networks such as VGG16 [21], and ResNet [22] variants (*e.g.* ResNet50, ResNet100), and MobileNet [23]. These architectures enhance model accuracy by expanding the receptive field through techniques like the Pyramid Pooling Module and Atrous Convolutions, which are employed during the encoding phase to capture broad semantic context. In the decoding phase, conditional random fields are used to smooth the output, with the aim of improving the accuracy and execution time. However, the execution time of both the DeepLab and PSPNet models remains relatively high due to their sequential nature, traversing each layer sequentially in the encoding and decoding phases. On the other hand, edge information is crucial for semantic segmentation, as it delineates boundaries between semantic regions within an image. Accurate preservation of edge information, typically found in the initial layers of backbone networks such as ResNet, InceptionNet, and VGG16, is essential

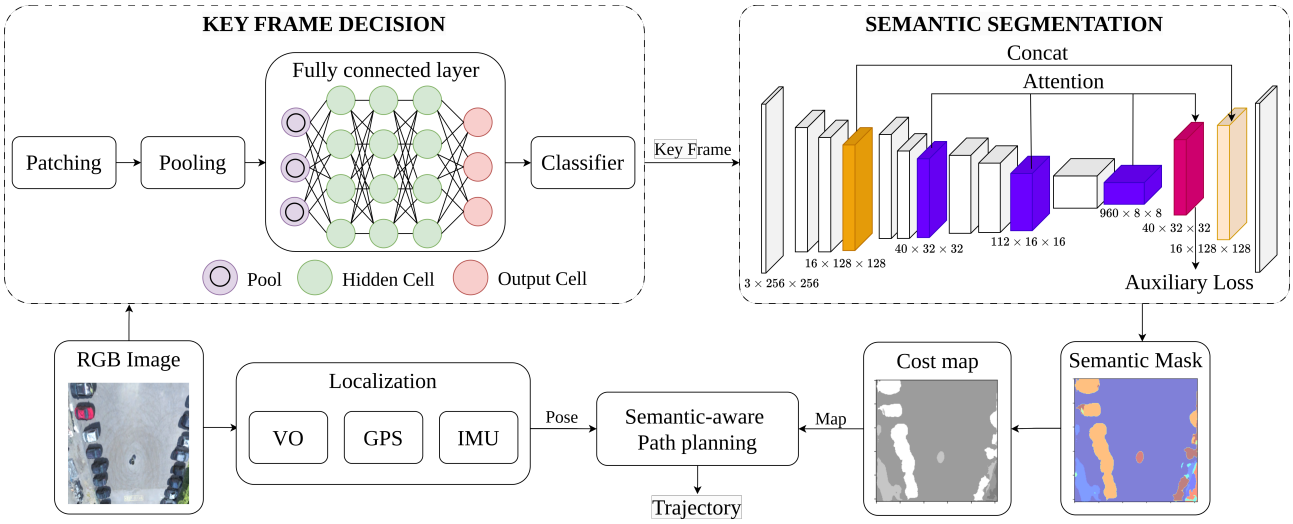


Fig. 1.: **Overall architecture of our proposed system:** The system is composed of three main units: Key Frame Decision Module, Semantic Segmentation Module, and Integration for Semantic Information and Path Planning.

for optimal model performance. Models like Unet and PsPNet prioritize edge information by using skip connections to transfer features from the encoding to the decoding module, enhancing the final segmentation accuracy.

To address these challenges, we propose a model (Semantic Segmentation module in Fig. 1) that supports parallel computation while leveraging edge information and maintaining rapid processing capabilities suitable for real-time systems. Our approach integrates the strengths of DeepLabv3, particularly its use of atrous convolutions, which accelerate processing and minimize the loss of critical features during pooling. Our model extracts edge information and incorporates it into the feature map at the final working layer. Additionally, features from other intermediate layers are combined with the feature map, thus requiring only two decoding layers compared to the more extensive structures in SegNet and DeepLabv3. This approach significantly improves execution time while maintaining high accuracy.

Moreover, the attention mechanism shown in Fig. 2 enhances the segmentation performance by focusing on the relevant parts of the images, capturing fine details and context. The attention architecture involves several steps. First, global pooling aggregates contextual information across the entire feature map:

$$\mathbf{g} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{f}_{ij} \tag{3}$$

where $\mathbf{g}$ is the global context vector, H and W are the height and width of the feature map, and $\mathbf{f}_{ij}$ represents the feature vector at position (i, j).

Next, the context vector $\mathbf{g}$ is passed through 1×1 convolution $\mathbf{g'} = \mathbf{W}_1\mathbf{g} + b_1$ with $\mathbf{W}_1$ and b_1 is the weight matrix and bias of the 1×1 convolution. The original feature map $\mathbf{f}$ is processed through a 3×3 convolution to emphasize important features: $\mathbf{f'} = \mathbf{W}_2 * \mathbf{f} + b_2$ with $\mathbf{W}_2$ and b_2 the weight matrix and bias of the 3×3 convolution, respectively. The attention weights $\mathbf{w}$ are computed as $\mathbf{w} = \mathbf{f'} \odot \mathbf{g'}$ with $\odot$ denoting element-wise multiplication. Finally, the refined feature map $\mathbf{f}_{out}$ is obtained by adding these attention weights back to the original feature map: $\mathbf{f}_{out} = \mathbf{f}_{i,j} + \mathbf{w}_{i,j}$.

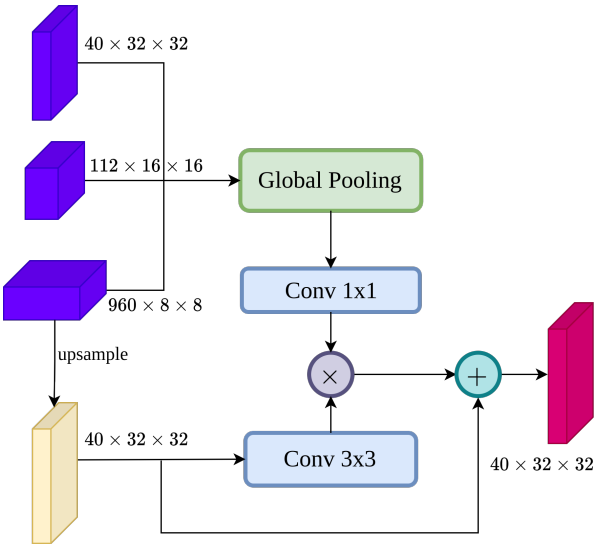


Fig. 2.: Attention mechanism.

To further boost performance, we employ an auxiliary loss technique, where the combined loss function is:

$$L = \lambda\mathcal{L} + (1 - \lambda)\mathcal{L}_{aux} \tag{4}$$

where λ balances the primary loss, $\mathcal{L}$ represents the discrepancy between the predicted output and the reference, and $\mathcal{L}_{\text{aux}}$ is the auxiliary loss function computed from the top layer of the decoder model.

2.3 Semantic-aware Path Planning

Upon receiving the semantic map, we generate a cost map based on the normalized semantic information.

Since we have 23 labels, the cost map value ranges from 0 to 1 corresponding to the semantic mask value from 0 to 22. We then present a hierarchical planner that incorporates the Dynamic Window Approach (DWA) [24] algorithm with this cost map. This algorithm aims to determine the optimal pair of linear and angular velocity values that describe the best achievable trajectory for the robot within the next time step. This is achieved through two main steps:
1. Determining feasible velocities based on constraints related to acceleration, deceleration, and obstacle avoidance.
2. Selecting the velocities that optimize an objective function.

The objective function of the Semantic-aware DWA controller, considering the cost map, is defined as follows:

$$G(v,\omega) = \alpha \cdot H(v,\omega) + \beta \cdot D(v,\omega) + \gamma \cdot Vel(v,\omega) + \epsilon \cdot C(v,\omega) \tag{5}$$

where $\alpha, \beta, \gamma, \epsilon$ are positive weight coefficients, and $H(v,\omega), D(v,\omega), Vel(v,\omega)$ represent heading function, distance function and velocity function, respectively. The cost map function $C(v,\omega)$ is defined as:

$$C(v,\omega) = \sum_{i=0}^{k} e^{-0.2t} f_k(x,y) \tag{6}$$

where t is the flight time and $f_k(x,y)$ is the value at the $k-th$ discrete position (x,y) of the robot on the trajectory mapped onto the cost map.

Thus, the Semantic-aware DWA path planning system accounts for dynamic constraints, motion capabilities, and environmental semantic factors. The parameter $e^{-0.2t}$ indicates that initially, the UAV can choose a longer flight path to reach a better semantic area. However, as it approaches the destination, it needs to fly directly to the destination as quickly as possible, as positioning errors become less critical due to their cumulative nature.

3. EXPERIMENTS

The experimental evaluation of our proposed system was conducted in two different environments using photo-realistic Unity simulation: Baxall Village and Singapore, as depicted in Fig. 3. Creating realistic simulation environments that closely mimic real-world conditions is a crucial step in the development and testing process of our system. The Singapore Bay environment consists of four distinct areas: a grass area with low light reflection, which is a favorable area for UAV operation, and two texture-less structured areas: asphalt roads and water surfaces with strong reflections, which negatively impact UAV localization, will be considered unfavorable areas. Tall buildings in Singapore Bay disrupt the UAV's GPS signal, necessitating avoidance maneuvers. Similarly, Baxall Village features a grassy surface suitable for UAV operation, while asphalt roads, parked cars, and areas with strong reflections pose challenges for UAV flight. Finally, we utilized sockets to enable seamless communication between the Unity simulation and the semantic segmentation module.

(a) Singapore environment (b) Baxall environment

Fig. 3.: Unity environment simulation.

3.1 Semantic Segmentation Results

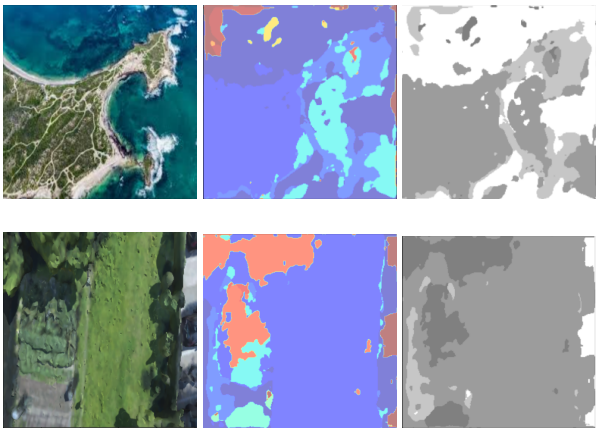
Our proposed semantic segmentation model is trained on the UAV Image Dataset, which comprises 3,600 images captured from UAVs at altitudes ranging from low to medium (20-30m). This dataset includes 23 semantic labels that encompass crucial labels utilized in this study, such as water bodies, grasslands, cars, and high-rise buildings.

(a) Label "Yes". (b) Label "No".

Fig. 4.: Example of the visual image rendered by Unity engine at training process.

We evaluated the efficacy of the keyframe decision module using accuracy metrics, considering the binary classification nature of the problem. The training dataset consisted of 200 images, with 137 images labeled as "Yes" and 63 images labeled as "No", as shown in Fig. 4. The key frame decision module achieved an accuracy of 72% and a true positive for the "Yes" label reached 0.83, demonstrating its capability to retain important cases that necessitate passing through the semantic segmentation module.

Table 1 compares our semantic segmentation model with previous models, using metrics such as mIoU (mean Intersection over Union) and average execution time (AET), averaged over 20 measurements. Our proposed model outperforms standard models like PSPNet [19], FPN [14], and FAN [13]. Specifically, it shows a 22.6% improvement in mIoU compared to PSPNet, 21.3% compared to FPN, and 15.4% compared to FAN. Additionally, it enhances execution time compared to these models. While maintaining similar execution times to DeepLabv3 [20], our model demonstrates an 8% accu-



(a) Input image (b) Semantic mask (c) Cost map

Fig. 5.: Results of semantic segmentation.

racy improvement. Compared to the Transformer model Segformer [15], our model ensures a satisfactory execution time with only a slight decrease in accuracy.

Table 1.: Results of segmentation models.

Model	mIoU	AET (ms)
PSPNet [19]	0.53	189
FPN [14]	0.46	221
FAN [13]	0.50	101
Segformer [15]	**0.71**	126
DeepLabv3 [20]	0.59	**91**
Our method	0.65	**91**

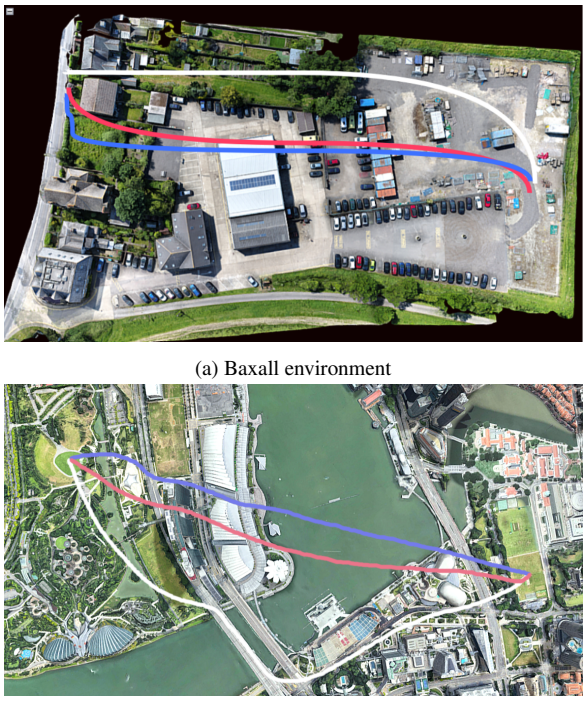
3.2 UAV Navigation Results

To demonstrate the ability of navigation, we calculated the UAV's flight distance and the distance flown into areas with unreliable localization. Specifically, unreliable localized areas in the Singapore environment include roads, water bodies, and the vicinity of high-rise buildings, while in the Baxall environment, poorly localized areas are defined as roads and cars. The results are summarized in Table 2:

Table 2.: Quantitative comparison for flight performance (Flight Distance (F-D), Unreliable Distance (U-D) meter).

Method	Baxall		Singapore	
	F-D	U-D	F-D	U-D
DWA [24]	17.3	5.8	50.6	20.3
Our method	19.3	3.8	72.1	10.2

The flight distance of the Semantic-aware DWA system slightly climbed over the standard DWA in both test environments, increasing by 11.6% in the Baxall environment and 42.5% in the Singapore environment. Furthermore, the unreliable distance of the UAV's flight path shows substantial improvement, with a 52.6% and 99.0% increase in the Baxall and Singapore environments, respectively. This highlights the system's precision and reliability. The integration of DWA with semantic segmentation also reduces positioning errors during the initial periods, proving beneficial for UAV systems by mitigating the accumulation of errors along the flight path.



(a) Baxall environment

(b) Singapore environment

Fig. 6.: Flight trajectories recorded: the white trajectory represents Semantic-aware DWA, the green and red trajectories represent DWA.

4. CONCLUSION

In this paper, we introduced a reliable localization system that integrates semantic segmentation with the Dynamic Window Approach (DWA) to enhance UAV navigation in texture-less and problematic areas using a monocular camera. Our proposed framework effectively addressed the challenge of navigating through areas with harder localization conditions, enabling enhanced perception of the environment for the path-planing algorithm. We demonstrated that our system can reach the assigned goal with high performance in terms of accuracy, execution time, flight distance, and unreliable distance when compared with other methodologies. Future directions could explore incorporating the uncertainty of labels in semantic segmentation and feature classifiers for UAV deployment, as well as integrating advanced machine learning techniques to further improve the UAV's perception capabilities.

REFERENCES

[1] S. S. Mansouri, P. Karvelis, C. Kanellakis, D. Kominiak, and G. Nikolakopoulos, "Vision-based mav navigation in underground mine using convolu-

tional neural network," in *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1, 2019, pp. 750–755.

[2] T. Tomic, K. Schmid, P. Lutz, A. Domel, M. Kassecker, E. Mair, I. L. Grixa, F. Ruess, M. Suppa, and D. Burschka, "Toward a fully autonomous uav: Research platform for indoor and outdoor urban search and rescue," *IEEE Robotics & Automation Magazine*, vol. 19, no. 3, pp. 46–56, 2012.

[3] S. S. Mansouri, C. Kanellakis, E. Fresk, D. Kominiak, and G. Nikolakopoulos, "Cooperative coverage path planning for visual inspection," *Control Engineering Practice*, vol. 74, pp. 118–131, 2018.

[4] J. Zègre-Hemsey, B. Bogle, C. Cunningham, K. Snyder, and W. Rosamond, "Delivery of automated external defibrillators (aed) by drones: Implications for emergency cardiac care," *Current Cardiovascular Risk Reports*, vol. 12, 09 2018.

[5] M. Dai, E. Zheng, Z. Feng, L. Qi, J. Zhuang, and W. Yang, "Vision-based uav self-positioning in low-altitude urban environments," *IEEE Transactions on Image Processing*, 2023.

[6] A. Bachrach, S. Prentice, R. He, P. Henry, A. S. Huang, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Estimation, planning, and mapping for autonomous flight using an rgb-d camera in gps-denied environments," *The International Journal of Robotics Research*, vol. 31, no. 11, pp. 1320–1343, 2012.

[7] G. Costante, J. Delmerico, M. Werlberger, P. Valigi, and D. Scaramuzza, "Exploiting photometric information for planning under uncertainty," *Robotics Research: Volume 1*, pp. 107–124, 2018.

[8] L. Bartolomei, L. Teixeira, and M. Chli, "Semantic-aware active perception for uavs using deep reinforcement learning," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3101–3108.

[9] J. Qian, K. Chen, Q. Chen, Y. Yang, J. Zhang, and S. Chen, "Robust visual-lidar simultaneous localization and mapping system for uav," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[10] S. Chen, W. Zhou, A. Yang, H. Chen, B. Li, C. Wen *et al.*, "An end-to-end uav simulation platform for visual slam and navigation aerospace," *Aerospace*, vol. 9, no. 2, 2022.

[11] M. Rizk, A. Mroue, M. Farran, and J. Charara, "Real-time slam based on image stitching for autonomous navigation of uavs in gnss-denied regions," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2020, pp. 301–304.

[12] T. N. Canh, T. S. Nguyen, C. H. Quach, X. HoangVan, and M. D. Phung, "Multisensor data fusion for reliable obstacle avoidance," in *2022 11th International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2022, pp. 385–390.

[13] D. Zhou, Z. Yu, E. Xie, C. Xiao, A. Anandkumar, J. Feng, and J. M. Alvarez, "Understanding the robustness in vision transformers," in *International Conference on Machine Learning*. PMLR, 2022, pp. 27 378–27 394.

[14] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6399–6408.

[15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[16] T. N. Canh, V.-T. Nguyen, X. HoangVan, A. Elibol, and N. Y. Chong, "S3m: Semantic segmentation sparse mapping for uavs with rgb-d camera," in *2024 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2024, pp. 899–905.

[17] T. N. Canh, A. Elibol, N. Y. Chong, and X. HoangVan, "Object-oriented semantic mapping for reliable uavs navigation," in *2023 12th International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2023, pp. 139–144.

[18] L. Bartolomei, L. Teixeira, and M. Chli, "Perception-aware path planning for uavs using semantic segmentation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5808–5815.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[24] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 4, no. 1, pp. 23–33, 1997.