IJE TRANSACTIONS B: Applications Vol. 37, No. 05, (May 2024) 1022-1035

International Journal of Engineering

Journal Homepage: www.ije.ir

Semantic Segmentation of Aerial Imagery: A Novel Approach Leveraging Hierarchical Multi-scale Features and Channel-based Attention for Drone Applications

E. Sahragard, H. Farsi*, S. Mohamadzadeh

Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran

P A P E R I N F O

Paper history:
Received 28 September 2023
Received in revised form 12 Novmber 2023
Accepted 19 November 2023

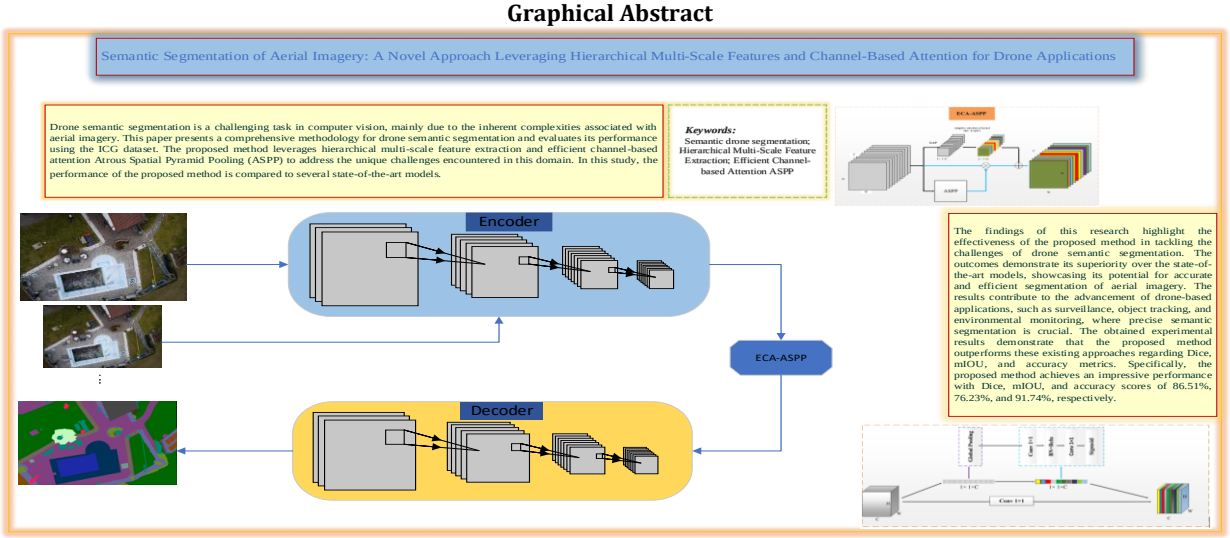
Keywords:
Semantic Drone Segmentation
Hierarchical Multi-scale Feature Extraction
Efficient Channel-based Attention
Atrous Spatial Pyramid Pooling

A B S T R A C T

Drone semantic segmentation is a challenging task in computer vision, mainly due to inherent complexities associated with aerial imagery. This paper presents a comprehensive methodology for drone semantic segmentation and evaluates its performance using the ICG dataset. The proposed method leverages hierarchical multi-scale feature extraction and efficient channel-based attention Atrous Spatial Pyramid Pooling (ASPP) to address the unique challenges encountered in this domain. In this study, the performance of the proposed method is compared to several state-of-the-art models. The findings of this research highlight the effectiveness of the proposed method in tackling the challenges of drone semantic segmentation. The outcomes demonstrate its superiority over the state-of-the-art models, showcasing its potential for accurate and efficient segmentation of aerial imagery. The results contribute to the advancement of drone-based applications, such as surveillance, object tracking, and environmental monitoring, where precise semantic segmentation is crucial. The obtained experimental results demonstrate that the proposed method outperforms these existing approaches regarding Dice, mIOU, and accuracy metrics. Specifically, the proposed method achieves an impressive performance with Dice, mIOU, and accuracy scores of 86.51%, 76.23%, and 91.74%, respectively.

***doi**: 10.5829/ije.2024.37.05b.18*

Graphical Abstract



*Corresponding Author Email: hfarsi@birjand.ac.ir (H. Farsi)

Please cite this article as: Sahragard E, Farsi H, Mohamadzadeh S. Semantic Segmentation of Aerial Imagery: A Novel Approach Leveraging Hierarchical Multi-scale Features and Channel-based Attention for Drone Applications. International Journal of Engineering, Transactions B: Applications. 2024;37(05):1022-35.

NOMENCLATURE

$x(i,j)$	Input feature map	C1	the weight of the first and second convolutions with kernel size 1
H	Height of the feature maps	TP	True positive
W	Width of the feature maps	FN	False-negative
IoU	Intersection over Union	FP	False positive
δ	ReLU activation function	σ	Sigmoid activation function

1. INTRODUCTION

Deep learning has emerged as a powerful technique in various domains, and it plays a crucial role in enabling drones to understand and accurately interpret their surroundings [1]. Semantic segmentation, a critical application, involves classifying objects or regions at the pixel level in drone images, enabling tasks such as aerial surveillance, infrastructure inspection, and precision agriculture(1). However, achieving accurate and reliable semantic segmentation for drone imagery poses significant challenges due to its unique characteristics (2),(3). The rapid advancement of drone technology has revolutionized industries by offering unprecedented capabilities for data acquisition and analysis (4). Drones, equipped with high-resolution cameras, provide a comprehensive aerial viewpoint, capturing valuable information that can be extracted through semantic segmentation. (5). This enables drones to make informed decisions based on their environmental understanding (6),(7). However, drone semantic segmentation encounters several challenges that musttion be addressed to achieve accurate results (8). The first challenge lies in the variability of scale and perspective inherent in aerial imagery. Objects of interest in drone images exhibit variations in distance, size, and orientation, making accurate segmentation challenging. Additionally, complex backgrounds often present in drone images, such as buildings, trees, and other objects, introduce occlusions and ambiguities in object boundaries, which further complicates the precise delineation from the surroundings. (9). The lack of readily available large-scale, diverse, and accurately annotated datasets poses a significant challenge when it comes to training robust semantic segmentation models that are specifically designed for drones. This limitation results in suboptimal performance and limited generalization capability. In this paper, we proposed a novel approach to address these challenges in drone semantic segmentation. Our method aims to improve the accuracy and robustness of semantic segmentation results by combining hierarchical multi-scale feature extraction with an efficient channel-based attention ASPP module. The proposed method contributes significantly to the field of drone semantic segmentation. Specifically, we introduced a hierarchical multi-scale feature extraction module that captures features at different scales and levels of granularity, enabling our model to handle scale and perspective variations prevalent in drone imagery. We also incorporated an efficient channel-based attention ASPP module that selectively focuses on informative features while suppressing irrelevant ones. This attention-based approach enhances the discriminative power of the model and improves segmentation accuracy. Furthermore, we proposed a feature fusion and integration step that combines the attention-guided features with the hierarchical multi-scale features, leveraging their complementary information to further enhance segmentation performance. Overall, our proposed method addresses the challenges of scale and perspective variability, complex backgrounds, and limited training data in drone semantic segmentation. Combining hierarchical multi-scale feature extraction and efficient channel-based attention ASPP provides a robust and accurate solution for interpreting drone-captured scenes.

In the following sections, we will describe the methodology in detail, present experimental results, and discuss the significance and implications of our findings. We believe these revisions provide a more targeted and logical overview of the existing work while highlighting the novelty and contributions of our approach in drone semantic segmentation.

2. LITERATURE REVIEW

During the past decade, deep learning has witnessed significant advancements in diverse domains, including machine vision (10), object tracking (11), and segmentation (12), (13). Deep learning methods have revolutionized this field by leveraging their ability to automatically learn and extract meaningful representations from large-scale datasets (14). In this section, we present a thorough review of research conducted on semantic segmentation of aerial imagery, particularly emphasizing its application in drone-related tasks. We examined the methodologies, techniques, and approaches utilized in previous studies, with a focus on how deep neural networks have been employed to achieve accurate semantic segmentation of aerial images. This literature review critically analyzes the accomplishments, limitations, and advancements in the field, laying the foundation for the proposed method. Additionally, it identifies the gaps that the present study aims to address. Moreover, this study highlights the significance of the proposed approach, which utilizes

hierarchical multi-scale features and channel-based attention. These advancements contribute to pushing the boundaries of aerial imagery analysis for drone applications. This research aims to advance semantic segmentation in aerial imagery by addressing identified gaps and introducing novel techniques. The intended outcome is an improved understanding and interpretation of aerial scenes. As a result, this advancement will enhance the capabilities and effectiveness of drones in various applications and domains such as surveillance, environmental monitoring, urban planning, and disaster response. We specifically examine the methodologies, techniques, and approaches utilized in previous studies, with a focus on the advancements made by prominent methods. These methods have made significant contributions to the field of semantic segmentation and have been widely adopted in various computer vision tasks(15).

FCN revolutionized semantic segmentation with end-to-end pixel-wise segmentation, serving as a foundational method (16). However, it has limitations in capturing fine details and object boundaries in aerial imagery (17). UNet introduced an encoder-decoder structure with skip connections and brought groundbreaking advancements in medical imaging (18). Nevertheless, it's important to note that the symmetric pathways in question may not fully cater to scale variations. This observation suggests that further considerations may be required to address the issue effectively. UNet++ was developed as an enhancement to UNet, incorporating nested skip and dense connections to improve segmentation accuracy. However, the increased complexity of UNet++ limit its feasibility in resource-constrained drone applications (19). DeepLab effectively addresses the challenge of capturing global and local contextual information using atrous/dilated convolutions and multi-scale contextual information (20). It dramatically improves segmentation accuracy, especially for objects of different scales and complex backgrounds. However, the reliance on dense dilated convolutions in DeepLab increases memory consumption and inference times, potentially introducing artefacts in segmentation masks. Lin et al. (21) utilize a feature pyramid network (FPN) for multi-scale object detection and segmentation, capturing local and global context. A limitation of this approach is its reliance on predefined anchor scales, which may encounter difficulties in handling diverse scale variations present in aerial imagery.

The PSPNet method, proposed by Zhao et al. (22) employs spatial pyramid pooling and dilated convolutions to capture contextual information at different scales in drone imagery. It utilizes a CNN backbone, like ResNet or VGG, to extract feature maps, followed by pyramid pooling modules. However, the pooling operations in PSPNet can cause information loss and reduced spatial resolution, leading to difficulties in accurately segmenting small objects and capturing fine details in aerial imagery.

By leveraging the insights and advancements from these methods, we propose a novel approach for semantic segmentation of aerial imagery in the context of drone applications. Our approach leverages hierarchical multi-scale features and channel-based attention mechanisms to enhance segmentation accuracy and improve the interpretability of aerial scenes. Through the integration of these innovative techniques, we aim to address the limitations and challenges faced in the field and contribute to the advancement of aerial imagery analysis for drone applications.

The following sections of this paper will provide a detailed description of our proposed method, including the architectural design, training strategies, and evaluation metrics. Additionally, we will present comprehensive experimental results to demonstrate the effectiveness and superiority of our approach compared to existing methods. Finally, we will discuss the implications of our findings and outline potential future research directions in the domain of semantic segmentation for drone applications.

3. PROPOSED METHOD

This section provides an overview of the methodology employed in this paper. We present a high-level description of the proposed method and explain its key components: hierarchical multi-scale feature extraction and efficient channel-based attention ASPP.

Furthermore, we discuss how these components effectively address the challenges encountered in drone semantic segmentation. The algorithm is introduced as follows:

Algorithm 1

1. **Input**: Drone image I, Ground truth segmentation map GT
2. **Preprocess** the input image I
3. **Define** the architecture of the proposed model, which includes the hierarchical multi-scale feature extraction and the Channel-based Attention ASPP module. The model consists of the following components:
   - Convolutional layers for feature extraction at different scales.
   - Efficient Channel-based Attention ASPP module for inter-channel dependencies.
   - Convolutional layers for refinement.
   - Up-sampling layers for restoring the original image size.
   - Softmax or sigmoid activation for obtaining pixel-wise predicted probabilities.
4. Initialize the model parameters.
5. Define the loss function, in this work as pixel-wise cross-entropy loss.
6. Set the number of training iterations and the learning rate for optimization.
7. Perform the training loop:
   a. For each iteration:
      - Perform forward propagation through the model:
        - Obtain multi-scale feature maps

- Apply the Efficient Channel-based Attention ASPP module to the multi-scale feature maps
- Fuse the multi-scale feature maps to generate a final feature representation.
- Apply convolutional layers and activation functions to refine the feature representation.
- Up-sample the refined feature representation to the original image size.
- Apply softmax activation to obtain pixel-wise predicted probabilities P for each class.

- Calculate the loss L between the predicted probabilities P and the ground truth segmentation map GT.
- Perform backpropagation to compute and update the gradients of the model's parameters.
- Update the model parameters using an optimizer with the defined learning rate.

b. Repeat the training loop for the specified number of iterations.

8. Evaluate the trained model on validation or test data:
   - Preprocess the validation/test images in the same way as the training images.
   - Perform forward propagation through the trained model to obtain predicted probabilities for the validation/test images.
   - Evaluate the segmentation performance using metrics such as intersection over union (IoU), Dice score and accuracy.
9. Output: Trained model for drone semantic segmentation.

The proposed method aims to improve the accuracy and robustness of drone semantic segmentation by combining hierarchical multi-scale feature extraction with efficient channel-based attention ASPP. The method influences the unique characteristics of aerial imagery and addresses the challenges posed by scale and perspective variability, complex backgrounds, and limited training data. The following explanation provides an overview of the components used in the proposed method.

3. 1. Hierarchical Multi-scale Feature Extraction

The hierarchical multi-scale feature extraction component captures features at multiple scales and levels of granularity (23), (24). It involves extracting features from different layers of the network architecture, allowing the model to incorporate information from various scales. By considering features at multiple scales, the model can handle the variations in object sizes, orientations, and perspectives often presented in drone imagery. This multi-scale feature extraction enables the model to capture both fine-grained details and global context, leading to improved segmentation accuracy. This component addresses the challenge of scale and perspective variability in drone imagery. The model can adapt to variations in object sizes, orientations, and perspectives by capturing features at different scales. This allows accurate segmentation of objects in aerial scenes, regardless of their scale or spatial arrangement. The use of multi-scale feature extraction in the model allows for the capture of both local details and global context. This, in turn, leads to improved segmentation accuracy, particularly when dealing with variations in scale.

In the proposed method, we utilize the ResNet-50 backbone, which is a widely used convolutional neural network architecture known for its effectiveness in feature extraction. The hierarchical feature extraction process begins with the initial convolutional layer of the ResNet-50 backbone, which captures low-level features such as edges and textures. These features provide a basis for subsequent layers to extract more complex and informative features.The ResNet-50 backbone consists of several stages, each containing multiple residual blocks. The feature extraction layers at different scales are determined by the stages and blocks within the ResNet-50 architecture. In the ResNet-50 architecture, the first stage consists of a single convolutional layer that captures low-level features. The subsequent stages, each contains a varying number of residual blocks. These blocks consist of multiple convolutional layers, including bottleneck layers that reduce the spatial dimensions and increase the number of channels. The hierarchical multi-scale feature extraction process with the ResNet-50 backbone enables the capture of both local and global information in drone imagery. The earlier stages of the ResNet-50 capture local information and fine-grained details, which are crucial for segmenting small objects or objects with intricate textures. These features preserve object boundaries and capture local variations effectively. As the network progresses through the stages of the ResNet-50, the scale of the features increases, incorporating more global context. The later stages capture features at coarser scales, enabling the model to consider the relationships between objects and their surroundings. This global context is essential for accurately segmenting larger objects and handling complex scenes with multiple objects and backgrounds. The hierarchical multi-scale feature extraction process provides a comprehensive representation of the input scene by combining features from different stages of the ResNet-50 backbone. The model can leverage these multi-scale features to make informed decisions during the segmentation process, effectively addressing the challenges of scale and perspective variability in drone imagery.

3. 2. Efficient Channel-based Attention ASPP

The efficient channel-based attention Atrous Spatial Pyramid Pooling (ASPP) component incorporates an attention mechanism that selectively focuses on informative features while suppressing irrelevant ones. This attention-based approach enhances the discriminative power of the model by assigning attention weights to different channels of the feature maps. By emphasizing relevant features and de-emphasizing less informative ones, the model becomes more adept at discriminating objects from complex backgrounds and handling occlusions. The efficient channel-based attention ASPP enables the model to exploit local and global contextual information effectively, leading to enhanced segmentation accuracy.

A vital advantage of the ASPP module with efficient channel attention lies in its ability to model interdependencies among different channels of feature maps. The channels in feature maps represent various aspects or semantic features of objects in the image. However, not all channels contribute equally to the segmentation task. The incorporation of efficient channel attention resolves this issue by dynamically adjusting the importance of each channel through learned attention weights. By emphasizing informative channels and suppressing noise or low-value channels, the ASPP module ensures accurate and reliable segmentation results. The utilization of channel-level attention allows the model to leverage limited training data more efficiently, resulting in enhanced generalizability. This component tackles the challenges of complex backgrounds and limited training data in drone semantic segmentation. By applying an attention mechanism, the model can selectively focus on informative features while suppressing irrelevant ones. This attention-based approach aids in discriminating objects from complex backgrounds and handling occlusions, leading to improved segmentation accuracy. During the training process, the attention weights are learned through backpropagation, optimizing the model to attend to the most informative features for semantic segmentation. These weights are adjusted iteratively, allowing the module to adaptively focus on relevant channels depending on the specific characteristics of the input data. The dynamic adjustment of attention weights ensures that the module can adapt to different semantic segmentation tasks and handle variations in object appearance and context. It enables the model to effectively capture both local and global information while suppressing noise and irrelevant details. Different components of the module are explained in the following sub-sections.

3. 2. 1. ASPP Module In CNN architectures, the ASPP module plays a crucial role in capturing multi-scale contextual information w ithin images effectively (20), (25). Its fundamental purpose is to aggregate features from distinct receptive fields of the CNN's convolution kernel, enabling the extraction of comprehensive multi-scale information from an image. The ASPP module consists of parallel branches that apply atrous spatial convolutions to the input image. Each branch operates at a specific dilation rate. The dilation rate determines the spacing between kernel elements, resulting in an expanded receptive field that enhances the ability of the model to capture contextual details while keeping the computational cost low (26). Figure 1 provides a visual illustration of the ASPP module, showcasing its architecture and functionality.

3. 2. 2. Efficient Channel-based Attention The attention module can be implemented using techniques like squeeze-and-excitation blocks (27), (28). These techniques enable the model to learn and adaptively adjust the channel-wise attention weights based on the input data. Figure 2 illustrates the architecture of the Efficient Channel-based Attention module.

We can formulate the ECA module as follows: Let X represent the input feature maps with dimensions H×W×C, where H and W are the height and width of the feature maps, and C is the number of channels.

- Compression: Apply GAP to X to obtain a channel descriptor z of dimensions 1×1×C:

$$z(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x(i,j) \quad (1)$$

- Re-weighting: Apply a two-layer convolution with kernel size 1 to z to obtain an attention of weights s of dimensions 1×1×C:

$$s = F_{ex}(z, C1) = \delta(C1(\sigma C1(z))) \quad (2)$$

where δ is a ReLU activation function, and C1 is the weight of the first and second convolutions with kernel size 1, respectively. Besides, σ is a sigmoid activation function that scales the 1×1 conv output to the range [0,1], ensuring that the weights are positive and sum to 1.

- Scaling: Apply the weights s to the original feature maps X to obtain the scaled feature maps Y of dimensions H×W×C:

$$Y = s \otimes X \quad (3)$$

After performing the element-wise multiplication denoted by ⊗, the resulting scaled feature maps,

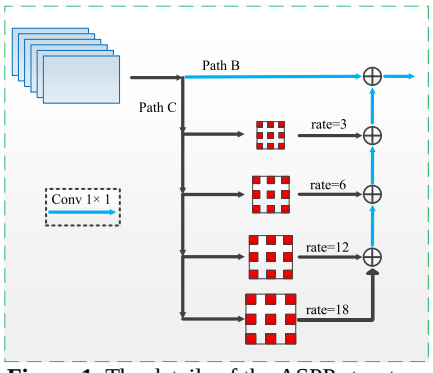


Figure 1. The details of the ASPP structure

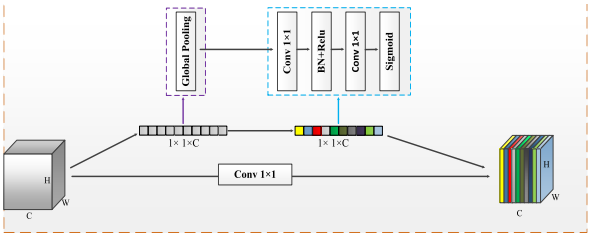


Figure 2. The details of the ECA structure

represented as Y, are subsequently passed to the next layer in the network. Using GAP to extract information from each channel of the input feature map, the ECA mechanism enabled the model to prioritize important features and enhance its overall performance. The resulting feature map utilized one-dimensional (1-D) convolutional cross-channel interaction instead of 1×1 convolutions to minimize the computational complexity of the model (29).

3. 2. 3. Efficient Channel-based Attention ASPP We introduce the Efficient Channel-based Attention ASPP (ECA-ASPP) module as an innovative component within our proposed method, offering an alternative to the concatenation operation utilized in the DeepLab architecture. Figure 3 provides a visual representation of the various components and operations involved within the ECA-ASPP structure. These details are crucial for understanding how the architecture functions and how it leverages its unique features to enhance image analysis.

Our module focuses on modeling interdependencies between channels presented in feature maps, dynamically adjusting the importance of each channel using attention weights. By employing this attention mechanism, we enhance feature representation, resulting in improved discriminative power and more accurate segmentation. The main advantage of our method is that it is able to selectively focus on useful information channels while reducing the importance of less relevant channels.This selective attention enables the network to efficiently utilize features, leading to enhanced segmentation performance and improved efficiency compared to the traditional concatenation operation.

By combining hierarchical multi-scale feature extraction with the ECA-ASPP, the proposed method leverages the complementary strengths of both approaches. The multi-scale feature extraction captures a wide range of spatial details, while the channel attention module enhances the discriminative power of the extracted features. This integration aids the model in effectively handling the challenges of drone semantic segmentation, such as variations in object scales, complex backgrounds, and occlusions.

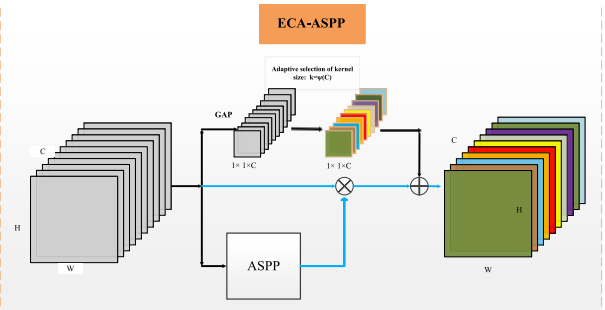


Figure 3. The details of ECA-ASPP structure

4. EXPRIMENTAL RESULT

4. 1. Datasets and Data Augmentation The designers of the Semantic Drone Dataset have specifically aimed to enhance the safety of autonomous drone flight and landing procedures by focusing on a semantic understanding of urban scenes. The dataset comprises imagery of over 20 houses captured from a bird's eye view at heights ranging from 5 to 30 meters above the ground.The images were acquired using a high-resolution camera with a size of 6000×4000 pixels (24 megapixels). The dataset includes pixel-accurate annotations for semantic segmentation. Detailed labels for 20 classes are assigned to the training and test sets. These classes include various elements such as trees, grass, other vegetation, dirt, gravel, rocks, water, paved areas, pools, persons, dogs, cars, bicycles, roofs, walls, fences, fence poles, windows, doors, and obstacles. The complexity of the dataset is constrained to these 20 classes, allowing researchers to focus on the specific semantic understanding of urban scenes. This carefully annotated dataset provides a valuable resource for developing and evaluating algorithms in semantic segmentation in the context of autonomous drone flights (30).

4. 1. 1. Data Augmentation In the proposed method, we utilize data augmentation techniques to improve the performance of our semantic segmentation model on the semantic drone dataset. Data augmentation is a widely used approach in computer vision tasks, including semantic segmentation, to address challenges such as limited labelled data and variations in environmental conditions. The data augmentation process involved applying a set of transformations to the original dataset, resulting in the creation of new and diverse training samples. These transformed samples facilitated an increase in quantity and variety of the training data, leading to improved model performance and generalization ability (31), (32). Specifically, we applied several standard data augmentation techniques to the semantic drone dataset:

- Random Cropping: This technique involves randomly selecting a portion of the image and using it as a new image. It aids in introducing variations in the position and composition of objects within the image. By cropping different parts of the image, the model can learn to recognize objects from various perspectives and locations.
- Horizontal Flipping: During horizontal flipping, the image undergoes a horizontal flip, creating a mirror image of the original. This technique is effective when the orientation of objects in the image does not affect their interpretation. It helps the model learn to recognize objects regardless of their left-right orientation.

- Vertical Flipping: Similar to horizontal flipping, vertical flipping involves flipping the image vertically, resulting in an upside-down version of the original image. It can be helpful in certain applications where the orientation of objects is not critical, such as text recognition or specific types of image classification.
- Rotation: This involves rotating the image to a certain degree. By applying the rotation, the model becomes more robust to changes in the orientation of objects in the image. It aids the model in learning to recognize objects from different angles and improves its ability to generalize to rotated images.
- Random Brightness and Contrast Adjustments: This technique involves randomly adjusting the brightness and contrast of the image. By modifying the brightness and contrast, the model can handle variations in lighting conditions. It helps the model become more resilient to changes in illumination and enhances its ability to generalize to images with different lighting levels.
- Contrast-Limited Adaptive Histogram Equalization (CLAHE): CLAHE is an image enhancement technique that improves the contrast of an image. It redistributes pixel intensities in a way that enhances details in both bright and dark regions of the image. The CLAHE aids the model in capturing fine-grained details and improves its performance in low-contrast images.
- Grid Distortion: Grid distortion applies a distortion effect to the image by manipulating a grid overlay. It introduces local deformations to the image, which can help the model learn to handle geometric transformations. The Grid distortion is particularly useful for tasks that require the model to be robust to deformations, such as object detection or image segmentation.
- Optical Distortion: Optical distortion simulates lens distortion effects in the image. It applies non-linear transformations to mimic the distortions introduced by different camera lenses. This technique is useful in scenarios where the images are captured by wide-angle lenses. By training the model with optically distorted images, it becomes more robust to lens distortions in real-world scenarios.

By applying these data augmentation techniques, we augmented the semantic drone dataset with transformed images, effectively expanding the size of the training dataset and introducing variations representative of real-world scenarios. This enabled our model to learn from a broader range of conditions and improved its ability to segment objects in unseen drone images accurately. Data augmentation played a crucial role in our semantic segmentation pipeline, enhancing the performance and generalization ability of our model on the semantic drone dataset.

Figure 4 shows the original image alongside four augmented versions generated from it. The original image serves as the base or reference image, while the four augmented images are created by applying augmentation techniques to the original image. Each augmented image has undergone a specific data augmentation technique. These techniques include random cropping, horizontal flipping, vertical flipping, rotation, random brightness, and contrast adjustments, contrast-limited adaptive histogram equalization, grid distortion, and optical distortion.

4. 2. Evaluation Metric

4. 2. 1. Intersection over Union Semantic segmentation tasks commonly employ the Intersection over Union (IoU) metric as the primary evaluation measure (33). This widely adopted metric quantifies the quality of a predicted segmentation mask by calculating the ratio between the intersection and the union of the predicted and ground truth masks. The IoU metric yields a value between 0 and 1, where a score of 1 denotes a flawless segmentation. The IoU is mathematically defined as follows:

$$IoU = \frac{TP}{TP+FP+FN} \quad (4)$$

where TP stands for true positive (the number of correctly classified pixels), FP stands for false positive (the number of incorrectly classified pixels), and FN stands for false negative (the number of pixels that should have been classified as belonging to the class but were not). Mean Intersection over Union (mIoU) is a commonly used evaluation metric for semantic segmentation tasks, defined as the average IoU score across all classes:

$$meanIoU = \frac{1}{C}\sum_c IoU_c \quad (5)$$

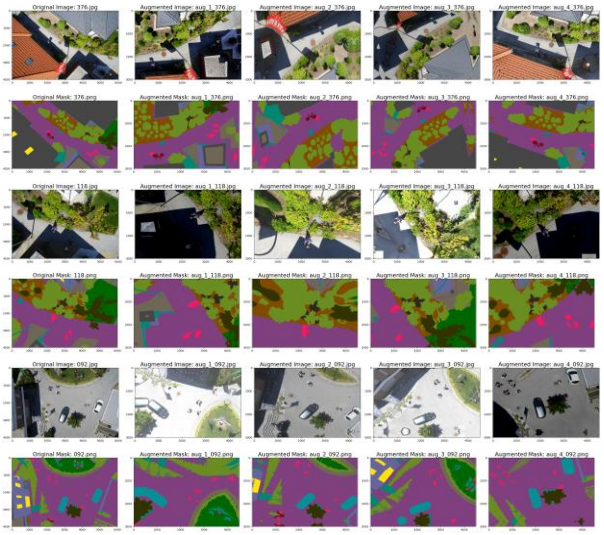


Figure 4. Examples of augmented images in the ICG dataset

The meanIoU (mIoU) score evaluates the overall accuracy of a segmentation model across all classes. A higher mIoU score signifies improved segmentation accuracy (34).

4. 2. 2. Dice Metric The Dice coefficient, also known as the Sørensen–Dice coefficient and F1 score, is a widely used metric for evaluating the performance of binary image segmentation models on a given dataset. It effectively captures the balance between false positives (FP) and false negatives (FN) (35). The Dice coefficient measures the degree of overlap between the predicted and ground truth segmentations. It ranges from 0 to 1, where 1 indicates a perfect overlap, and 0 specifies no overlap. The computation of the Dice coefficient relies on the counts of true positives (TP), false negatives (FN), and false positives (FP), which can be derived from the confusion matrix of the model's predicted outcomes.The TP count represents the number of correctly identified positive pixels, while the FN count reflects the number of incorrectly identified negative pixels. Conversely, the FP count denotes the number of pixels erroneously classified as positive. The Dice coefficient is mathematically defined as follows:

$$Dice = \frac{2|X \cap Y|}{|X|+|Y|} = \frac{2TP}{2TP+FP+FN} \quad (6)$$

The Mean Dice (mDice) coefficient represents the average Dice coefficient score across all classes within a dataset. We can calculate the Dice coefficient as follows, which provides a measure of overall segmentation performance:

$$mean\ Dice = \sum_c Dice_c \quad (7)$$

4. 2. 3. Pixel-wise Accuracy The evaluation of performance and accuracy in semantic segmentation models often involves utilizing the pixel-wise accuracy metric for pixel-level predictions. Pixel-wise accuracy calculates the ratio of correctly classified pixels to the total number of pixels in the image. To calculate this metric, we compare the model's estimated outcomes with the ground truth labels on a pixel-by-pixel basis. Each pixel in the predicted segmentation is compared to its corresponding pixel in the ground truth segmentation. If the predicted label matches the ground truth label, the pixel is classified correctly. Utilizing pixel-wise accuracy provides valuable insights into the overall performance of semantic segmentation models.

4. 3. Exprimental Result Table 1 presents the experimental results for evaluating a semantic drone segmentation model. The metrics used for evaluation are IoU, Dice coefficient, and Accuracy. Each row is related to a specific class. Table 1 demonstrates the performance of the proposed method across different classes. The model achieves high IoU, Dice, and Accuracy scores for several classes such as paved-area, grass, pool, and car.

TABLE 1. Performance Evaluation of Proposed Method Across Multiple Classes

Class	IoU	Dice	Acc
Unlabeled	50.14	66.7910	88.02
Paved-area	95	97.4359	97.42
Dirt	62.71	77.0819	79.26
Grass	93.04	96.3945	97.45
Gravel	79.98	88.8765	96.17
Water	92.35	96.0229	98.52
Rocks	85.89	92.4095	95.67
Pool	96.62	98.2809	98.54
Vegetation	74.4	85.3211	85.37
Roof	94.9	97.3833	97.66
Wall	84.85	91.8042	82.89
Window	69.28	81.8526	87.9
Door	47.03	63.9733	63.51
Fence	59.65	74.7260	73.93
Fence-pole	42.4	59.5506	60.86
Person	79.16	88.3679	89.72
Dog	68.05	80.9878	75.01
Car	94.25	97.0399	98.58
Bicycle	67.58	80.6540	75.83
Tree	78.17	87.7477	84.03
Bald-tree	79.99	88.8827	89.5
Ar-marker	88.17	93.7131	95.28
Obstacle	78.23	87.7854	93.46

These high scores indicate that the proposed model successfully captures the boundaries and details of these classes, resulting in accurate segmentation. However, it is worth noting that certain classes such as unlabeled, dirt, fence-pole, and door exhibit lower scores, indicating areas where the model faces challenges in achieving accurate segmentation. These classes may pose challenges due to their complex or ambiguous visual characteristics, resulting in lower performance than other classes. Figure 5 demonstrates the performance of the proposed method for drone semantic segmentation using augmented images from the ICG dataset. The figure comprises three columns, each providing crucial information about the segmentation process. The first column showcases the original images from the augmented ICG dataset. The drone captures these images, and we apply augmentation techniques such as rotation, scaling, flipping, or adding noise to enhance them.

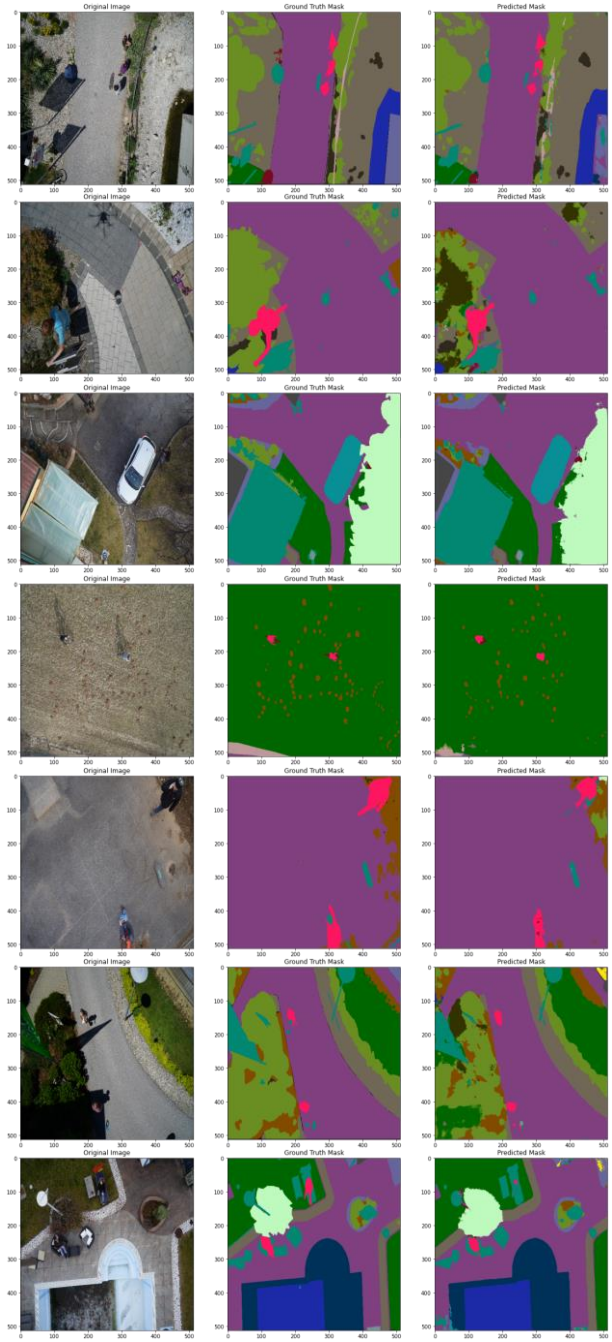


Figure 5. Qualitative results. From left to right: input, ground truth, our method on ICG

Augmenting the dataset enhances its diversity and enables the model to handle a broader range of real-world scenarios. The second column presents the ground truth annotations, which serve as the reference labels for each image pixel. These annotations are meticulously handcrafted masks, accurately outlining the boundaries and regions of interest in the augmented images. They represent the accurate segmentation information and provide a benchmark against which the performance of the proposed method can be evaluated. The third column displays the predicted masks generated by the proposed method. These masks result from applying our trained model to the augmented images from the dataset. The experimental results indicate that the proposed method performs well in handling the augmented ICG dataset for drone semantic segmentation. The predicted masks exhibit a high degree of agreement with the ground truth annotations, suggesting that the proposed model successfully captures and classifies the objects in the augmented drone imagery. These results highlight the robustness and generalization capability of the proposed method, showcasing its potential for real-world applications. Figure 6 illustrates the accuracy of metrics' measurement results for all the compared methods obtained during the validation phases.

Figures 7 and 8 illustrate the measurement results of the Dice coefficient and mIOU metrics for all the compared methods during the training and validation phases. We obtained these results by training for 20 epochs. By analyzing the results in Figures 7 and 8, it is evident that the proposed method outperforms the other compared techniques in terms of both the Dice coefficient and mIOU metrics. The proposed method achieves significantly higher scores, indicating its superiority in accurately segmenting objects in the given dataset.

In our study on semantic segmentation, we employed the following hyperparameters to train and evaluate the models. These choices were made based on prior research in the field and empirical observations. We set the batch size to 8, determining the number of samples propagated through the network in each training iteration. This value strikes a balance between memory consumption and convergence speed. To complete one epoch, we used 200 steps per epoch. This value ensures that the model is exposed to a diverse

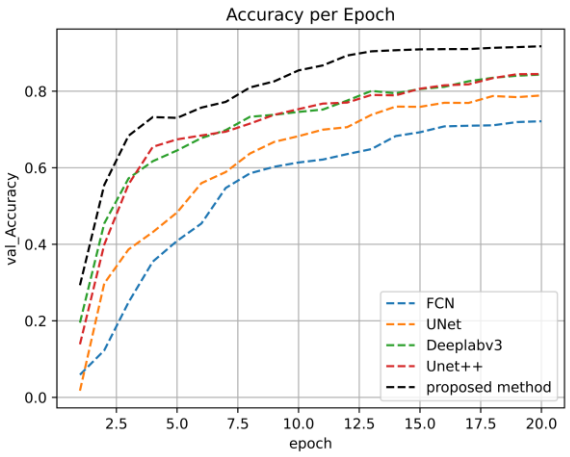


Figure 6. Comparison of accuracy for different methods validation phases

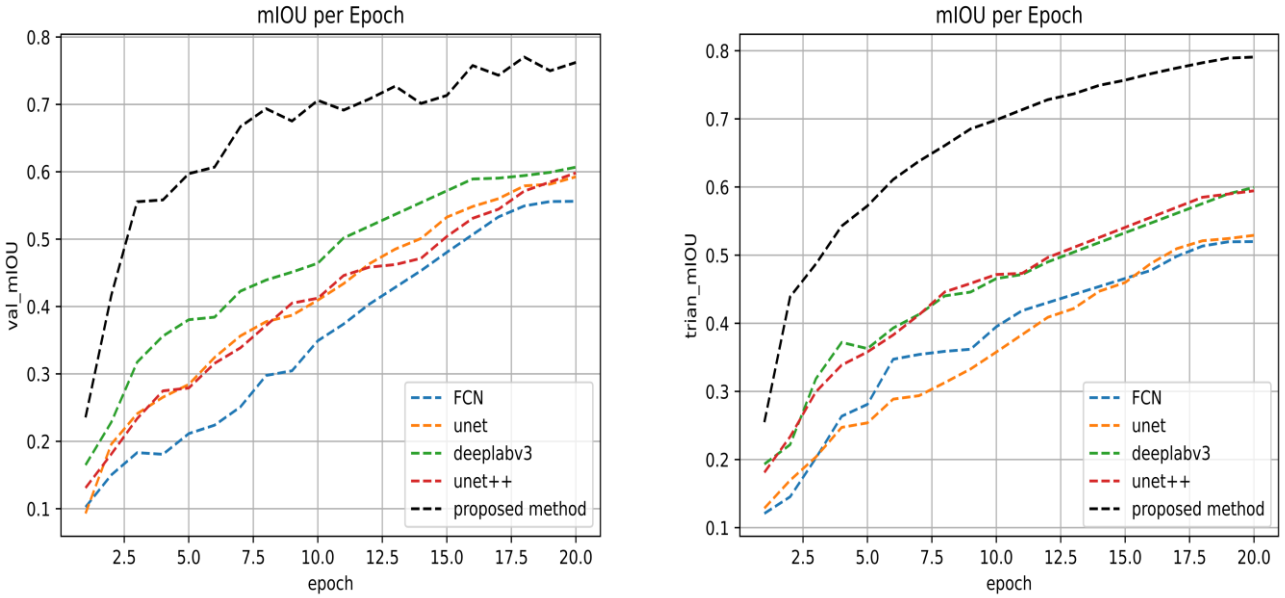


Figure 7. Comparison of mIOU metrics for different methods in training and validation phases

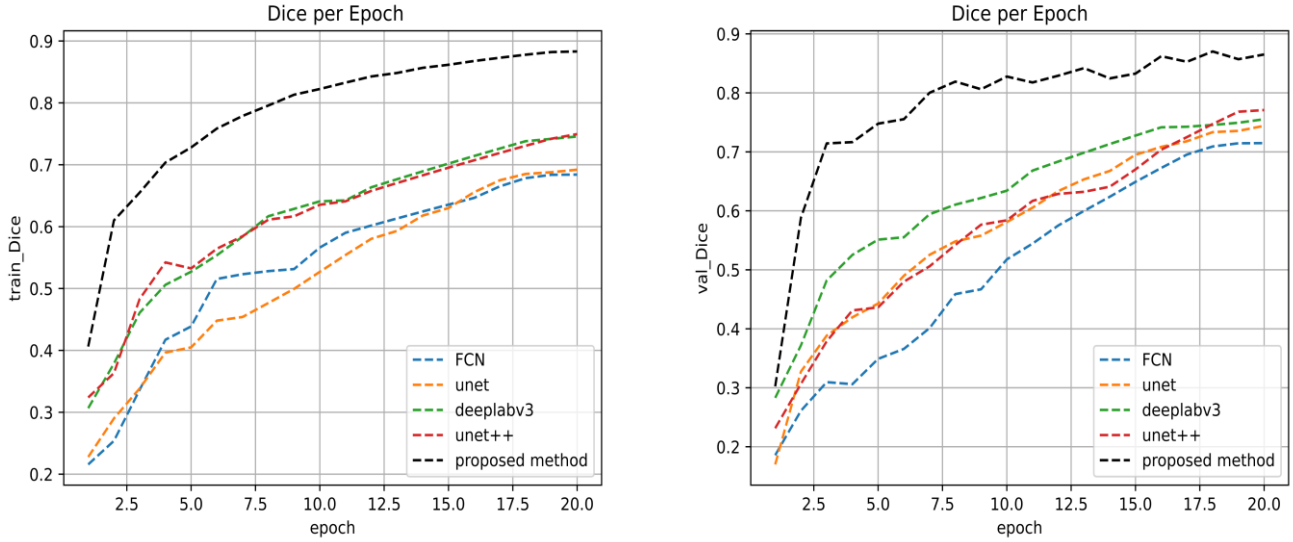


Figure 8. Comparison of dice metrics for different methods in training and validation phases

range of samples during training, facilitating better generalization. For validation, we utilized ten validation steps per epoch, allowing us to evaluate the performance of the model on a separate set of samples.

The input shape of images was (512, 512, 3). We initialized the learning rate to 0.0001, determining the step size during gradient descent optimization. We chose this initial learning rate to strike a balance between convergence speed and accuracy.To further refine the learning process, we employed an Exponential Decay Learning-Rate-Scheduler callback. This scheduler gradually reduced the learning rate over time, aiding the model in refining its parameters effectively. The proposed model was trained for 20 epochs, allowing the model to learn from the dataset multiple times. The number of epochs affects both training time and the capacity of the model to generalize, which refers to its ability to accurately classify or predict unseen or new data that was not a part of the training set. A model with good generalization can effectively extrapolate patterns and provide an accurate prediction on unfamiliar data, indicating its robustness and ability to handle real-world scenarios.

Additionally, we utilized two callbacks during training: Model-Checkpoint and Early-Stopping. The Model-Checkpoint callback saved the best model weights based on a specified metric, enabling the retrieval of the best-performing model. The Early-Stopping callback monitored a specified metric and stopped training early if the metric did not improve for a certain number of epochs, preventing overfitting.

Table 2 provides information on the training schedule and time for a semantic segmentation model. The model was trained for 20 epochs, with a maximum (initial) learning rate of 10e4 and a minimum learning rate of 1.12× 10e5. The total training time for the model was 22981 seconds. Additionally, it is worth noting that the best weights were obtained at the 18th epoch, indicating the optimal point of model performance during training.

Table 3 presents the evaluation results of various semantic segmentation methods for drone images, including the proposed method. We assessed the performance using three metrics: Dice coefficient, mean Intersection over Union (mIOU), and accuracy. The proposed method achieved an impressive Dice coefficient of 86.51%, indicating a strong agreement between the predicted and ground truth segmentations. This demonstrates the accuracy of the proposed method in capturing the shapes and boundaries of objects in the drone images. Additionally, our method achieved an mIOU of 76.23%, showcasing its ability to represent the spatial extent of the objects accurately.The high mIOU suggests that the method effectively captures the overall quality of the segmentation output. Moreover, our method achieved an accuracy of 91.74%, indicating its effectiveness in correctly labelling pixels within the drone images. This showcases the reliability of the proposed method in accurately classifying the pixels. Comparing the proposed method to the other methods in the table, we outperformed them regarding Dice coefficient, mIOU, and accuracy.

The FCN method achieved a Dice coefficient of 71.44, an mIOU of 55.58, and an accuracy of 72.14. These results indicate that FCN performs reasonably well but has limitations in accurately capturing fine details and object boundaries in aerial imagery. The UNet-VGG16 method showed improvement with a Dice coefficient of 74.80, an mIOU of 59.70, and an accuracy of 78.44. This suggests that incorporating the VGG16 architecture in the UNet framework enhances the segmentation results. Further enhancing the UNet architecture with the ResNet50 backbone, the UNet-ResNet50 method achieved a Dice coefficient of 75.01, an mIOU of 60.01, and an accuracy of 78.91. This suggests that integrating a more advanced backbone network results in enhanced segmentation performance. The FPN method demonstrated even better performance with a Dice coefficient of 75.37, an mIOU of 60.48, and an accuracy of 80.04. This suggests that utilizing a feature pyramid network effectively captures multi-scale information and enhances segmentation accuracy. The PspNet method continued the trend of improvement, achieving a Dice coefficient of 76.52, an mIOU of 61.98, and an accuracy of 83.93. This indicates that the Pyramid Scene Parsing Network (PspNet) is effectively captures both local and global contextual information, leading to improved segmentation results. DeeplabV3-ResNet50 obtained a Dice coefficient of 75.94, an mIOU of 61.21, and an accuracy of 84.35. This suggests that the DeeplabV3 architecture, combined with the ResNet50 backbone, improves segmentation accuracy, particularly for objects of different scales and complex backgrounds. The Unet++ method improved the results with a Dice coefficient of 77.06, an mIOU of 62.68, and an accuracy of 84.48. combining incorporating nested skip and dense connections in the UNet architecture enhances segmentation accuracy. DeeplabV3+-VGG16 achieved a Dice coefficient of 76.87, an mIOU of 64.01, and an accuracy of 85.79. This suggests that utilizing the VGG16 architecture in the DeeplabV3+ framework improves segmentation accuracy, capturing finer details and object boundaries.DeeplabV3+-ResNet50 showed even better performance, with a Dice coefficient of 79.18, an mIOU of 65.27, and an accuracy of 86.11. This indicates that combining the DeeplabV3+ architecture with the ResNet50 backbone further enhances segmentation accuracy.

Finally, the proposed method outperformed all other methods, achieving a Dice coefficient of 86.51, an

TABLE 2. Training schedule and time with learning rate (best weights at 18th epoch)

Epochs	Max. (Initial) LR	Min. LR	Total Training Time
20	10e4	1.12× 10e5	22981 s

TABLE 3. Comparative evaluation of semantic drone segmentation methods

	Dice	mIOU	Accuracy
FCN (36)	71.44	55.58	72.14
UNet-VGG16 (37)	74.80	59.70	78.44
UNet-ResNet50	75.01	60.01	78.91
FPN (21)	75.37	60.48	80.04
PspNet (22)	76.52	61.98	83.93
DeeplabV3-ResNet50 (20)	75.94	61.21	84.35
Unet++ (19)	77.06	62.68	84.48
DeeplabV3+VGG16 (38)	76.87	64.01	85.79
DeeplabV3+ResNet50 (38)	79.18	65.27	86.11
Proposed	**86.51**	**76.23**	**91.74**

mIOU of 76.23, and an accuracy of 91.74. These results demonstrate that the novel method proposed in this paper achieves the highest segmentation accuracy, showcasing its effectiveness and superiority over existing methods in semantic drone segmentation.

Our method outperforms other evaluated methods in accurately segmenting objects in drone images. It combines hierarchical multi-scale feature extraction with an Efficient Channel-based Attention ASPP module, capturing local and global information. The proposed method achieves significantly higher segmentation accuracy by focusing on relevant features. These findings contribute to advancing drone semantic segmentation techniques and offer insights for future research. Our method shows superior performance, promising improved accuracy and reliability in drone image segmentation.

5. CONCLUSION

In this study, we proposed a novel method for drone semantic segmentation that combines hierarchical multi-scale feature extraction and an Efficient Channel-based Attention ASPP module. The superior performance of our proposed method can be attributed to its ability to capture both local and global information while efficiently focusing on relevant features, resulting in accurate object segmentation in drone images. The evaluation results demonstrate that the proposed method outperforms other existing methods regarding segmentation accuracy. These findings validate the effectiveness of the hybrid approach and its potential to advance the field of drone semantic segmentation. Furthermore, our proposed method offers significant advancements in drone imagery applications. Improving the accuracy and reliability of segmentation algorithms provides valuable insights for various tasks such as object detection, tracking, and scene understanding in drone-based systems. Looking ahead, there are promising prospects for the suggested hybrid approach. One potential direction for future research is to explore the scalability and efficiency of the method for real-time or near-real-time applications. This could involve optimizing the computational efficiency of the approach to enable its deployment on resource-constrained platforms. Moreover, further investigations can be conducted to evaluate the proposed method on large-scale, diverse, and challenging datasets specific to drone imagery. This would provide a deeper understanding of its performance and generalization capabilities across different environmental conditions and object classes. The findings contribute to the existing body of knowledge and provide a foundation for future research and development in drone-based computer vision systems.

6. REFERENCES

1. Bhatnagar S, Gill L, Ghosh B. Drone image segmentation using machine and deep learning for mapping raised bog vegetation communities. Remote Sensing. 2020;12(16):2602. 10.3390/rs12162602
2. Lyu Y, Vosselman G, Xia G-S, Yilmaz A, Yang MY. UAVid: A semantic segmentation dataset for UAV imagery. ISPRS journal of photogrammetry and remote sensing. 2020;165:108-19.
3. Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review. 2021;54:137-78. 10.48550/arXiv.1910.07655
4. Benjdira B, Bazi Y, Koubaa A, Ouni K. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. Remote Sensing. 2019;11(11):1369.
5. Chakravarthy AS, Sinha S, Narang P, Mandal M, Chamola V, Yu FR. DroneSegNet: Robust Aerial Semantic Segmentation for UAV-Based IoT Applications. IEEE Transactions on Vehicular Technology. 2022;71(4):4277-86. 10.1109/TVT.2022.3144358
6. Zhang L, Wang M, Ding Y, Wan T, Qi B, Pang Y. FBC-ANet: A Semantic Segmentation Model for UAV Forest Fire Images Combining Boundary Enhancement and Context Awareness. Drones. 2023;7(7):456. 10.3390/drones7070456
7. Mahmudnia D, Arashpour M, Bai Y, Feng H. Drones and blockchain integration to manage forest fires in remote regions. Drones. 2022;6(11):331. 10.3390/drones6110331
8. Kumar S, Kumar A, Lee D-G. Semantic Segmentation of UAV Images Based on Transformer Framework with Context Information. Mathematics. 2022;10(24):4735. 10.3390/math10244735
9. Lobo Torres D, Queiroz Feitosa R, Nigri Happ P, Elena Cué La Rosa L, Marcato Junior J, Martins J, et al. Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution UAV optical imagery. Sensors. 2020;20(2):563. 10.3390/s20020563
10. Sezavar A, Farsi H, Mohamadzadeh S. A Modified Grasshopper Optimization Algorithm Combined with CNN for Content Based Image Retrieval. International Journal of Engineering. 2019;32(7):924-30. 10.5829/ije.2019.32.07a.04 https://www.ije.ir/article_87122_c98ece12fa7377fe34ea6c8a5f519352.pdf
11. Feizi A. Convolutional Gating Network for Object Tracking. International Journal of Engineering. 2019;32(7):931-9. 10.5829/ije.2019.32.07a.05 https://www.ije.ir/article_87123_97e73a696cb75dc070dd2dced8875f7d.pdf
12. Hassanpour H, Mortezaie Z, Beghdadi A. Sensing Image Regions for Enhancing Accuracy in People Re-identification. Iranian (Iranica) Journal of Energy & Environment. 2022;13(3):295-304.
13. Habibi M, Hassanpour H. Splicing Image Forgery Detection and Localization Based on Color Edge Inconsistency using Statistical Dispersion Measures. International Journal of Engineering. 2021;34(2):443-51. 10.5829/ije.2021.34.02b.16 https://www.ije.ir/article_124931_7626c7eef9e9dc3ae8dfa8e6c6adf7ac.pdf
14. Sakimalla G, Chilukuri P, Jamuna A. Picture Segmentation using changing Artifacts Identification and Bias Modification. 2023. 10.36227/techrxiv.22362469.v1
15. Niu R, Sun X, Tian Y, Diao W, Feng Y, Fu K. Improving semantic segmentation in aerial imagery via graph reasoning and

disentangled learning. IEEE Transactions on Geoscience and Remote Sensing. 2021;60:1-18.

16. Zagoruyko S, Komodakis N. Wide residual networks. arXiv preprint arXiv:160507146. 2016.
17. Kestur R, Farooq S, Abdal R, Mehraj E, Narasipura O, Mudigere M. UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle. Journal of Applied Remote Sensing. 2018;12(1):016020-.
18. Giang TL, Dang KB, Le QT, Nguyen VG, Tong SS, Pham V-M. U-Net convolutional networks for mining land cover classification based on high-resolution UAV imagery. Ieee Access. 2020;8:186257-73. 10.1109/ACCESS.2020.3030112
19. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging. 2019;39(6):1856-67. 10.48550/arXiv.1912.05074
20. Chiu W-T, Lin C-H, Jhu C-L, Lin C, Chen Y-C, Huang M-J, et al., editors. Semantic segmentation of lotus leaves in UAV aerial images via U-Net and deepLab-based networks. 2020 International Computer Symposium (ICS); 2020: IEEE. 10.1109/ICS51289.2020.00110
21. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S, editors. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
22. Zhao H, Shi J, Qi X, Wang X, Jia J, editors. Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. 10.48550/arXiv.1612.01105
23. Li S, Zhu X, Bao J. Hierarchical multi-scale convolutional neural networks for hyperspectral image classification. Sensors. 2019;19(7):1714. 10.3390/s19071714
24. Huo X, Sun G, Tian S, Wang Y, Yu L, Long J, et al. HiFuse: Hierarchical Multi-Scale Feature Fusion Network for Medical Image Classification. arXiv preprint arXiv:220910218. 2022. 10.1016/j.bspc.2023.105534
25. Megir V, Sfikas G, Mekras A, Nikou C, Ioannidis D, Tzovaras D, editors. Salient object detection with pretrained deeplab and K-means: Application to UAV-captured building imagery. Pattern Recognition ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII; 2021: Springer. 10.1007/978-3-030-68787-8_35
26. Liu Y, Wang L, Zhao L, Yu Z. Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Volume 1: Springer Nature; 2019.
27. Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations. arXiv preprint arXiv:180302155. 2018. 10.48550/arXiv.1803.02155
28. Hu J, Shen L, Sun G, editors. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.
29. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q, editors. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. 10.48550/arXiv.1910.03151
30. Semantic Drone Dataset. In: Technology GUo, editor. Institute of Computer Graphics and Vision, Graz University of Technology.
31. Prakash S, Shah P, Agrawal A. Exploiting CNNs for Semantic Segmentation with Pascal VOC. arXiv preprint arXiv:230413216. 2023. 10.48550/arXiv.2304.13216
32. Rebuffi S-A, Gowal S, Calian DA, Stimberg F, Wiles O, Mann TA. Data augmentation can improve robustness. Advances in Neural Information Processing Systems. 2021;34:29935-48. 10.48550/arXiv.2111.05328
33. Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, et al. Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging. 2019;38(10):2281-92. 10.48550/arXiv.1903.02740
34. Union GIO, editor A metric and a loss for bounding box regression. Rezatofighi, N Tsoi, J Gwak, A Sadeghian, I Reid, S Savarese//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA; 2019. 10.48550/arXiv.1902.09630
35. Wu K, Du B, Luo M, Wen H, Shen Y, Feng J, editors. Weakly supervised brain lesion segmentation via attentional representation learning. Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22; 2019: Springer. 10.1007/978-3-030-32248-9_24
36. Long J, Shelhamer E, Darrell T, editors. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015.
37. Ronneberger O, Fischer P, Brox T, editors. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18; 2015: Springer.
38. Guo Z, Xu J, Liu A, editors. Remote sensing image semantic segmentation method based on improved Deeplabv3+. International Conference on Image Processing and Intelligent Control (IPIC 2021); 2021: SPIE. 10.1117/12.2611930

Persian Abstract

چکیده

تقسیم‌بندی معنایی پهپاد یک وظیفه چالش‌برانگیز در حوزه بینایی کامپیوتر است. این وظیفه به دلیل پیچیدگی‌های مرتبط با تصاویر هوایی، به ویژه دشوار است. در این مقاله، یک روش جامع برای تقسیم‌بندی معنایی پهپاد ارائه شده است و عملکرد آن با استفاده از مجموعه داده ICG مورد ارزیابی قرار می‌گیرد. روش پیشنهادی از استخراج ویژگی‌های چندمقیاس سلسله‌مراتبی و استفاده بهینه از توجه مبتنی بر کانال استفاده می‌کند. این روش به چالش‌های منحصربه‌فرد موجود در این حوزه پاسخ می‌دهد. در این مطالعه، عملکرد روش پیشنهادی با چندین مدل مدرن مقایسه شده است. نتایج آزمایشی نشان می‌دهند که روش پیشنهادی نسبت به این رویکردهای موجود با معیارهای Dice، mIOU و دقت عملکرد بهبود قابل توجهی دارد. به طور خاص، روش پیشنهادی با معیارهای Dice، mIOU و دقت به ترتیب ۸۶.۵۱٪، ۷۶.۲۳٪ و ۹۱.۷٤٪ عملکرد چشمگیری را به دست می‌آورد. یافته‌های این پژوهش نشان می‌دهند که روش پیشنهادی در مقابله با چالش‌های تقسیم‌بندی معنایی پهپاد مؤثر است. این نتایج به پیشرفت برنامه‌های مبتنی بر پهپاد، مانند نظارت، پیگیری اشیا و نظارت محیطی که تقسیم‌بندی معنایی دقیق آنها ضروری است، کمک می‌کند.