applied sciences

MDPI

Article

Remote Sensing Image Segmentation for Aircraft Recognition Using U-Net as Deep Learning Architecture

Fadi Shaar [1], Arif Yılmaz [2,*], Ahmet Ercan Topcu [3,*] and Yehia Ibrahim Alzoubi [4]

[1] Computer Engineering Department, Ankara Medipol University, 06050 Ankara, Turkey; fadi.saar@ankaramedipol.edu.tr
[2] Department of Advanced Computing Sciences, Maastricht University, 6211 LK Maastricht, The Netherlands
[3] Department College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait
[4] College of Business Administration, American University of the Middle East, Egaila 54200, Kuwait; yehia.alzoubi@aum.edu.kw
* Correspondence: arif.yilmaz@maastrichtuniversity.nl (A.Y.); ahmet.topcu@aum.edu.kw (A.E.T.)

Abstract: Recognizing aircraft automatically by using satellite images has different applications in both the civil and military sectors. However, due to the complexity and variety of the foreground and background of the analyzed images, it remains challenging to obtain a suitable representation of aircraft for identification. Many studies and solutions have been presented in the literature, but only a few studies have suggested handling the issue using semantic image segmentation techniques due to the lack of publicly labeled datasets. With the advancement of CNNs, researchers have presented some CNN architectures, such as U-Net, which has the ability to obtain very good performance using a small training dataset. The U-Net architecture has received much attention for segmenting 2D and 3D biomedical images and has been recognized to be highly successful for pixel-wise satellite image classification. In this paper, we propose a binary image segmentation model to recognize aircraft by exploiting and adopting the U-Net architecture for remote sensing satellite images. The proposed model does not require a significant amount of labeled data and alleviates the need for manual aircraft feature extraction. The public dense labeling remote sensing dataset is used to perform the experiments and measure the robustness and performance of the proposed model. The mean IoU and pixel accuracy are adopted as metrics to assess the obtained results. The results in the testing dataset indicate that the proposed model can achieve a 95.08% mean IoU and a pixel accuracy of 98.24%.

Keywords: remote; image; segmentation; aircraft; deep learning; ensemble; U-Net

Citation: Shaar, F.; Yılmaz, A.; Topcu, A.E.; Alzoubi, Y.I. Remote Sensing Image Segmentation for Aircraft Recognition Using U-Net as Deep Learning Architecture. *Appl. Sci.* **2024**, *14*, 2639. https://doi.org/10.3390/app14062639

Academic Editors: Mirka Saarela, Lilia Georgieva and Vili Podgorelec

Received: 11 February 2024
Revised: 16 March 2024
Accepted: 19 March 2024
Published: 21 March 2024

1. Introduction

Images acquired by aerial platforms and satellites for remote sensing purposes are described as depictions of the Earth's surface from a vantage point in space. The semantic labeling of these images is widely recognized as a critical challenge [1]. This task involves assigning a class label to each pixel within an image. Over the past ten years, advancements in technology, such as compact imaging sensors and unmanned aerial vehicles, have yielded significant enhancements in image quality while reducing operational and equipment costs [2]. These cost-effective platforms provide flexible access to both multi-spectral and high-resolution images, along with an accelerated data acquisition rate [3]. Consequently, a significant challenge has emerged concerning the efficient management of extensive data collections and the swift retrieval of specific data of interest. Content-Based Image Retrieval (CBIR) is recognized as a valuable approach for rapidly accessing desired images within large-scale datasets [4].

In recent years, substantial effort has been dedicated to tailoring CBIR techniques for remote sensing images, leading to the emergence of a vibrant and complex field known as Remote Sensing Image Retrieval (RSIR) [5]. Within this field, researchers have placed

particular emphasis on the advancement of feature extraction methods. This focus is critical because the effectiveness of image retrieval largely hinges on the quality and efficiency of the features used [6]. As a result, numerous methodologies have been developed for the application of semantic segmentation in the context of aerial and satellite images. These techniques represent innovative approaches to enhance the retrieval and analysis of remote sensing data, addressing the unique challenges and complexities of this domain.

These techniques may be divided into three main categories. Methods using manual feature extraction methods, such as the Scale-Invariant Feature Transform (SIFT) method, fall under the first group [7], including Vector of Locally Aggregated Descriptors (VLAD) [8], improved Fisher kernel [9], bag of visual words [10], local binary patterns, and Gabor texture [11,12]. With these techniques, features are manually extracted from remote sensing photos. The second kind uses common classifiers that need previous knowledge of the feature extraction procedure to categorize individual pixels, such as Conditional Random Fields (CRFs) or Random Forest. Deep learning methods, which have become popular in the semantic segmentation of satellite pictures, are what distinguish the final category. These models, which frequently use neural networks, are capable of automatically learning contextual characteristics from the input data, such as higher-level specifics and shape aspects [13,14]. They have changed this area of research by enabling more automated and adaptable feature extraction, and they are known for their capacity to extract complex information [1]. In order to meet various criteria and circumstances for the semantic segmentation of satellite pictures, each of these categories provides unique benefits and trade-offs.

The automatic detection of aircraft in high-resolution satellite images represents a significant and intricate challenge in the realm of target identification. This challenge holds critical importance across both military and civilian domains, with dynamic airfield surveillance being one of its prominent applications [15]. High-resolution satellite images offer a wealth of spatial information, textures, and colors due to their superior quality. However, automating the recognition of aircraft in such images proves to be a highly demanding task, primarily because of the complex and diverse structures depicted in satellite imagery. Furthermore, aircraft exhibit variations in terms of shape, color, and size. Even for a single type of aircraft, the intensity and texture of its appearance can differ significantly across various scenarios [16]. This complexity underscores the substantial intricacies involved in this area of research and development.

When it comes to recognizing a single label for RSIR, several freely available benchmark datasets have been established [6]. However, it is worth noting that, within the domain of remote sensing, there is a scarcity of studies that have provided public datasets specifically tailored for image-segmentation-based aircraft detection. This scarcity of publicly available datasets has posed a limitation on the development of innovative solutions in this area. Consequently, if one intends to train a network using real data for aircraft detection, the available training dataset may be relatively small. In response to this challenge, and by leveraging the advancements in Convolutional Neural Networks (CNNs), researchers have introduced various CNN architectures, such as U-Net [17], which have demonstrated remarkable performance even with limited training data. In this study, we propose a pixel-wise image segmentation model that draws inspiration from the specialized architecture of CNNs. The primary framework for our model is based on U-Net, which was selected for its effectiveness in achieving the rapid and high-quality recognition of aircraft objects. To evaluate the performance of our aircraft recognition model, we employ the dense labeling remote sensing dataset (DLRSD) [4]. This dataset serves as a valuable resource for assessing the capabilities of the planned model in the context of aircraft recognition [18]. The contributions of this study can be summarized as follows:

1. The developed architecture presents an end-to-end semantic segmentation model with robust expandability. This model can be effectively employed for the automatic recognition of aircraft, including those that are not present in the initial training

dataset. Importantly, this recognition can be achieved without the need to retrain the model.
2. This study integrates several strategic approaches, including data preprocessing and data augmentation, with the model designed for aircraft semantic segmentation. These strategies collectively contribute to a substantial improvement in accuracy, enhancing the model's performance.
3. To the best of our knowledge, this study represents the first effort to explore the DLRSD for the development of a semantic segmentation model tailored for aircraft recognition by employing the U-Net architecture. This exploration expands the scope of available resources and insights in the field of aircraft recognition, making a notable contribution to the literature.

The remainder of this article is divided into the following sections: To provide context and insights into the body of knowledge that is already available in the field of aircraft recognition utilizing U-Net and deep learning, Section 2 explores the background and relevant works in the literature. This study's materials and methods are described in depth in Section 3, including the research strategies, data pretreatment methods, and the use of U-Net and deep learning for aircraft recognition. In Section 4, the findings are discussed, the experimental data are presented and analyzed, and the performance and effectiveness of the suggested semantic segmentation model are evaluated. Section 5 discusses the results. Finally, Section 6 serves as the conclusion, summarizing the key findings and contributions and extending the discussion to potential future research directions within the scope of aircraft recognition using U-Net and deep learning.

2. Background and Related Literature

2.1. Classical Methods for Aircraft Recognition

In the realm of classical methods for aircraft recognition, many studies in the literature rely on the utilization of rotation-invariant features, often achieved by applying binarization techniques to images. Prominent examples of such features include Fourier descriptors and moment invariant features [19]. These methods typically leverage shape features extracted from either the object's contour or a binary representation of the object. They assume that accurate edge detection or region delineation can be readily obtained, which is a task that can be challenging in real-world scenarios. Additionally, these approaches often rely on threshold-based image segmentation applied to the entire shape or silhouette of the target [20]. Features related to rotation invariance, such as Zernike moments, wavelet moments, and Hu moments, are subsequently extracted for recognition purposes [19].

For example, in [20], Zernike invariant moments were effectively combined with an independent component algorithm to facilitate aircraft recognition. Meanwhile, in [21], the authors adopted histogram equalization to identify airports in images. Subsequently, they employed segmentation to isolate the areas containing planes, followed by the generation of binary images. A Principal Component Analysis (PCA) was applied to determine the main axes of each airplane, and aircraft recognition was accomplished through template matching [21]. However, these methods encounter certain limitations highlighted by [16]: (1) The alignment of segmentations or aircraft often necessitates direction estimation, which is constrained by the methods' capacity and may result in inaccuracies in direction prediction. (2) The coarse shape representations employed by these techniques tend to overlook crucial details that are essential for discriminating between different aircraft types.

2.2. Aircraft Recognition Using Image Features Directly

Several methods have been introduced for aircraft recognition, focusing on direct image feature utilization. In [22], aircraft recognition was accomplished through a combination of backpropagation neural networks and the moment invariant method. Similarly, [23] employed principal component features to train a neural network-based classifier. To perform classification, they adopted a directed acyclic graph support vector machine model. However, these methods exhibit certain limitations, as highlighted by [15]. They are notably

sensitive to the distribution of training data and necessitate a substantial amount of data for effective performance. Moreover, they often yield suboptimal recognition accuracy, particularly when dealing with imbalanced data distributions. In comparison to template-matching methods, these approaches possess two key drawbacks. First, they deviate from the task of detecting targets in natural images, as they require individuals experienced in interpretation techniques to label each aircraft promptly. Second, they demand a substantial volume of training examples and are sensitive to data distribution.

2.3. CNNs for Semantic Segmentation

Recent developments in the field of image analysis, driven by the utilization of clusters and Graphics Processing Units (GPUs) along with advancements in algorithms, have substantially reduced processing times. Consequently, deep learning methods have garnered significant attention, as noted by [24]. Specifically, the study of semantic segmentation using CNNs can be categorized into two primary sets of techniques: patch-based and pixel-based approaches [25]. In patch-based techniques, patches surrounding each pixel within the input image are extracted, and a single label is predicted for each patch using a CNN, effectively classifying the entire image [26]. While this category of techniques has contributed significantly to image segmentation, it is associated with certain drawbacks, including limited receptive fields, computational inefficiency, and substantial memory requirements.

In contrast, pixel-based techniques involve predicting a label for each individual pixel within the entire image. A notable advancement in this realm is the Fully Convolutional Network (FCN) [27]. An FCN represents a classical model that deploys convolutional layers instead of fully connected layers, making it adept at upscaling coarse segmentation outputs into more precise results. This is achieved through the utilization of transposed convolutional layers. Various CNN models have adopted the FCN architecture, characterized by an encoder–decoder structure, and have achieved commendable results. However, it is worth noting that one drawback associated with the FCN architecture is the potential for significant factor up-sampling, which can introduce classification ambiguities.

To address the challenge of classification ambiguities due to large factor up-sampling, various models have been proposed. One such model is the DeepLab-CRF, presented by [28], which incorporates a fully connected CRF. This model introduces "atrous" convolutions to mitigate the impact of eliminating pooling layers. To achieve smoother raw segmentation results, the fully connected CRF is integrated with the responses from the final CNN layer. Another approach is the DeconvNet [29], which employs a multi-layer deconvolution network to replace bilinear interpolation in the upscaling stage. This network incorporates un-pooling (the reverse of max-pooling) and deconvolution layers to enhance the upscaling process. Similarly, the SegNet [30], which is conceptually similar to DeconvNet but offers a simplified architecture, has the advantage of being trainable end-to-end and features a significantly smaller parameterization, making it a computationally efficient choice for image segmentation tasks.

2.4. CNNs for Semantic Segmentation of Remote Sensing Images

Deep neural networks, particularly CNNs, have gained considerable attention in the large-scale processing of remote sensing images in recent years. These networks have proven to be the top-performing tools for high-resolution semantic labeling and have been applied in various remote sensing tasks. For instance, in [31], experiments were conducted using high-resolution remote sensing imagery to analyze several dense semantic labeling CNNs by developing a multi-layer perceptron CNN model. In another work presented in [32], a network architecture with an hourglass shape was designed, complemented by a down-sampled network followed by an up-sampled network. It introduced an inception module, spanning from the encoder to the decoder, facilitating the direct flow of information across different network layers. This design choice aimed to enhance the network's ability to capture and utilize information effectively during the semantic labeling process.

In [16], deep CNNs were offered as a framework for classifying different types of aircraft. The method was a multi-step procedure. First, a specialized network was created to segment airplanes, yielding complex findings that captured the specifics needed to differentiate between various aircraft types. The segmentation outcomes were subsequently improved by adding a network for keypoint detection to recognize aircraft bounding boxes and orientations. The IoU measure was used to evaluate the similarity between segmentation outputs and specified templates, hence confirming the findings. Finally, a template matching approach was utilized for aircraft recognition. While such CNN models are typically resilient to label noise, it is vital to remember that they are in fact data hungry. When given a significant amount of training instances, frequently ranging from hundreds to millions or even billions, they tend to function at their best. This emphasizes the need to have a substantial and rich training dataset to enable the efficient training and generalization of CNN-based models for image segmentation tasks.

The U-Net architecture [17] represents a specific type of FCN that has gained significant attention for its effectiveness in segmenting both 2D and 3D biomedical images [17,33]. Subsequently, it was recognized that this model also performs exceptionally well in the context of pixel-wise satellite image classification. Recently, U-Net achieved great performance not only for images related to the biomedical processing field but also for object segmentation from satellite images [2,34–37]. In the study conducted in [38], an approach was applied to address the challenge of semantic segmentation in satellite images. They devised an architecture akin to U-Net, leveraging ResNet-34 weights within the encoder component. This algorithm demonstrated impressive results in the detection of roads from satellite images sourced from the DeepGlobe database, achieving a notable public score of 0.64.

To facilitate the semantic segmentation of remote sensing images, the authors in [3] introduced a contextual U-Net architecture. Within this framework, they integrated three collaborative enhancements: a module designed to extract boundary features, enabling the fusion of both semantic and adaptive characteristics; an adaptive feature selection module to prioritize important semantic channels, especially for handling irregular objects; and, to effectively combine hierarchical features while utilizing dynamic inter-layer feature guidance, they incorporated a recursive feature fusion module. To detect edge maps in optical remote sensing images, the authors of [5] created a spatial channel attention U-Net model. This model effectively highlights aircraft within the images. The spatial channel attention U-Net provides significant edge cues, while the usage of encoders ensures a robust representation of salient object features. Subsequently, the decoders receive the output from the encoders. Within the decoders, a feature-merge module focuses on the positions of prominent objects. As a result, the final output includes the identified aircraft.

The authors of [39] developed a multi-scale residual U-Net with an attention (MSRAU-Net) scheme for multi-scale aircraft segmentation in remote sensing photos to solve this issue for U-Net if used for multi-scale segmentation. Two types of attention components, two adjusted Respaths, and a multi-scale convolution component were created and added to MSRAU-Net to gather the multi-scale features and improve the efficiency of the features' fusion. MSRAU-Net performs better than the other networks, especially when it comes to recognizing tiny aircraft, according to the trials conducted on the RSI dataset. MSRAU-Net obtained an F1 score of 93.10% and an accuracy of 93.12%. The authors claim that MSRAU-Net performed better than FCN, U-Net, AU-Net, and MultiResUNet [39]. Using the U-Net model, the study demonstrated a real-time method for object segmentation in aerial drone photos. The model's efficiency is increased by the use of data augmentation and transfer learning approaches. Several base architectures, such as VGG 16, ResNet-50, and MobileNet, were tested using the U-Net segmentation model. The model's performance is enhanced by data augmentation, as demonstrated by the experimental findings, which yield a segmentation accuracy of 92% for VGG-16, 93% for ResNet, and 95% for MobileNet [40].

As demonstrated in Table 1, our model surpassed all existing models in the domain of detecting and segmenting moving aircraft. This highlights the superior performance

of our proposed approach compared to previous methods. Our research stands out from previous work by employing the U-Net architecture for aircraft recognition through semantic segmentation. Additionally, our study may represent the first attempt to utilize the DLRSD specifically for the development of a semantic segmentation model tailored for aircraft recognition using the U-Net architecture. This innovative approach highlights the potential of U-Net in the context of aircraft image segmentation and opens new possibilities for enhancing the accuracy and efficiency of remote sensing applications.

Table 1. Selected literature on image detection.

Study	Domain	Innovation/Technique	Accuracy Achieved
[31]	Image semantic labeling	Multi-layer perceptron (MLP) was proposed, which outperformed other techniques like CNN	88.92%
[32]	Aerial image semantic segmentation	Hourglass-shaped network (HSN)-based semantic segmentation was proposed	89.42%
[16]	Aircraft classification	Keypoints' detection network	95.60%
[17]	Biomedical image segmentation	U-Net architecture was deployed	92.03%
[33]	Cardiac MR segmentation	U-Net architecture was deployed	___
[34]	Satellite image segmentation	U-Net architecture was deployed	87.61%
[2]	Satellite image segmentation	Improved U-Net architecture was deployed	97.56%
[35]	Forest change detection	U-Net architecture was deployed	99.00%
[36]	Brain tumor segmentation in magnetic resonance imaging (MRI)	U-Net architecture was deployed	99.00%
[37]	Glaucoma detection in retinal fundus images	U-Net and supervised ML algorithms	100%
[38]	Satellite image segmentation	ResNet-34 weights within the encoder component	64.00%
[3]	Remote image segmentation	Contextual U-Net architecture	___
[5]	Aircraft image segmentation	Spatial channel attention U-Net model	___
[39]	Multi-scale aircraft segmentation	MSRAU-Net	93.12%
[40]	Aerial drone object segmentation	U-Net segmentation model	95.00%
This study	Aircraft image segmentation	U-Net segmentation model	98.24%

3. Materials and Methods

The methodology used in this research involved utilizing the U-Net framework for the semantic segmentation of aircraft satellite images. To further improve feature localization, we integrated skip connections into the U-Net architecture, allowing for the combination of deep and high-resolution features. Additionally, NVIDIA CUDA technology was utilized for GPU acceleration to expedite the training process and optimize the model. The recognition of aircraft was accomplished through a three-stage workflow consisting of data preprocessing, training, and testing phases. We used the OpenCV library and an alpha compositing algorithm to transparently overlay the color mask over the original grayscale image.

3.1. Architecture of U-Net Used for Aircraft Recognition

As discussed in the previous section, U-Net is introduced as a modified version of the FCN, with two primary architectural distinctions. First, in contrast to FCN, which employs a 1×1 convolution layer at the last layer of the encoder for utilizing pertained models, U-Net omits this layer from the encoding stage. Second, to enhance the localization accuracy, U-Net leverages high-resolution features by effectively merging the encoding and decoding layers. Originally intended for biomedical segmentation tasks, U-Net has become renowned for its remarkable performance in a variety of image segmentation applications. One of the pivotal features contributing to the efficacy of the U-Net architecture is the incorporation of skip connections. These connections facilitate the fusion of both low-level and higher-level feature maps, a mechanism that plays a crucial role in achieving

precise object localization during the segmentation process. This ability to capture detailed information at multiple levels of abstraction is a key factor behind U-Net's success in producing accurate segmentation results.

The architecture employed in this study for performing the semantic segmentation of aircraft satellite images utilizes the U-Net framework, as illustrated in Figure 1. The U-Net architecture comprises two main sections: the left part, referred to as the encoder or contracting path, and the right part, known as the decoder or expansive path. The encoder section follows a conventional CNN structure and consists of four blocks of layers. Each block incorporates two convolutional layers with multiple 3×3 filters applied sequentially for feature extraction. After each convolution operation, a Rectified Linear Unit (ReLU) is employed as a nonlinear activation function for the feature maps. To achieve down-sampling, max-pooling with a filter size of 2×2 is applied, reducing the spatial dimensions by a factor of 2. At each step of dimension reduction, the number of channels is doubled, increasing the representational capacity of the network. Both the encoder and decoder sections consist of an equal number of blocks. However, to recover the size of the feature maps in the decoder section, each block includes an up-sampling operation, reducing the number of channels through a 2×2 filter, effectively implementing deconvolution. This structure enables the decoder to reconstruct detailed information from the reduced feature maps, facilitating accurate semantic segmentation of the aircraft satellite images.

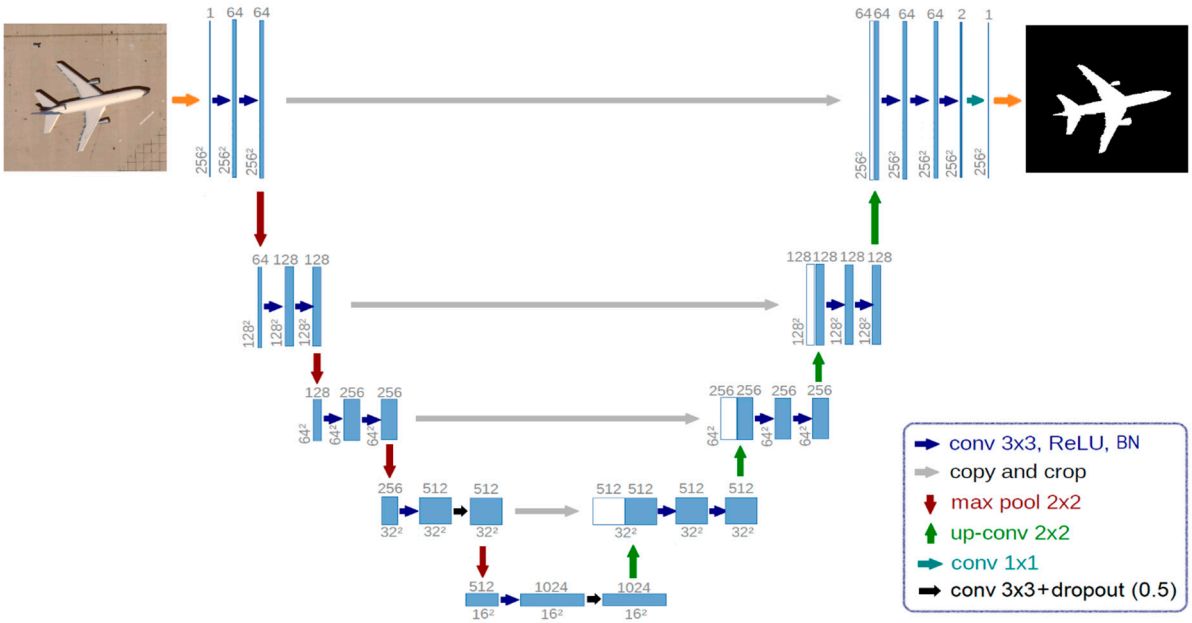


Figure 1. The U-Net-architecture for aircraft recognition.

Furthermore, to enhance the localization of up-sampled features, the skip-connections technique is employed. This involves combining the features from the deep expansive path with the high-resolution features from the shallow contracting path. As a result, the expansive branch effectively increases the resolution of the output. ReLU activation functions are applied after each convolution operation, with the exception of the last layer. The final layer of the network employs a sigmoid activation function to produce a pixel-wise probability map. To facilitate convergence during training, batch normalization is employed after each convolutional layer, except for the last layer. The last layer of the network is responsible for generating the model's output using a 1×1 convolution operation to map each pixel to a class, distinguishing between "Aircraft" and "Background". In summary, the architecture of the network encompasses a total of 23 ReLU activation functions, 24 convolutional layers, 19 batch normalization operations, 2 dropout operations, 4 max-pooling operations, 4 merging operations, and 4 up-sampling operations. To obtain the final predicted aircraft extraction result, the pixel-wise probability map needs to be

binarized using a specified threshold, typically set to 0.5, yielding the output of the semantic aircraft segmentation network.

3.2. Semantic Workflow of Proposed Model

The proposed workflow for aircraft recognition is divided into three primary stages, as illustrated in Figure 2: the data preprocessing phase, the training phase, and the testing phase. In the data preprocessing phase, various techniques are applied to prepare the input dataset for training and testing. This contains procedures like grayscale picture conversion and average division standardization. Additionally, techniques for data augmentation are used to diversify the training dataset and reduce the danger of model overfitting. These augmentation techniques include horizontal flipping, rotation, zooming, shearing, width and height shifting, and picture rotation. The U-Net model is used as the fundamental architecture in the training phase, and this model is trained using the enhanced training dataset created in the preceding step. In order to evaluate the trained model's prediction accuracy and confirm the performance of the suggested model, test pictures are fed through the trained model in the testing phase. This stage acts as the last assessment of the model's capability to spot airplanes in satellite photos.

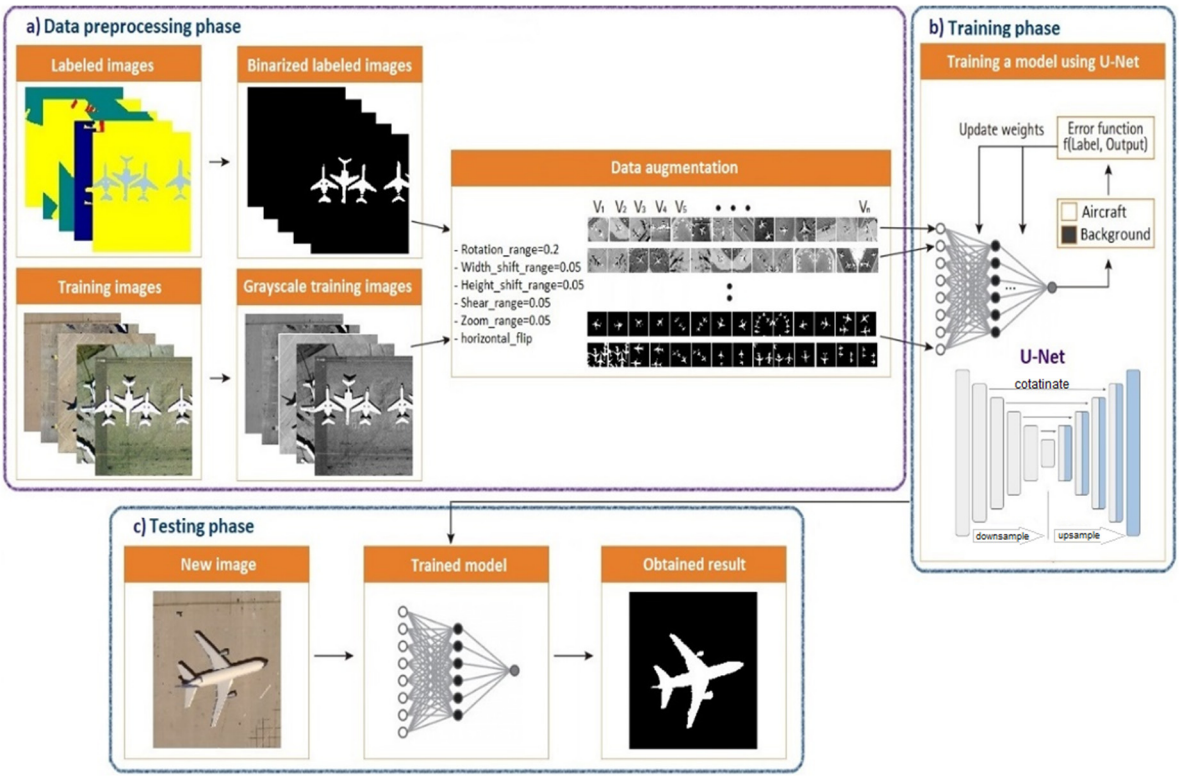


Figure 2. Semantic workflow of proposed aircraft recognition model.

In this study, the dataset is partitioned into three segments for training, validation, and testing. The training and validation portions of the dataset are utilized for training and validating the proposed model [41]. Subsequently, the trained model is employed to evaluate its performance using the testing dataset. It is worth noting that CNN-based approaches often require substantial computational resources. To expedite the computational processes within the neural network, the training and validation operations were conducted on a GPU utilizing NVIDIA CUDA technology. CUDA is a parallel computing technology that leverages numerous independent streams to accelerate operations. It is widely supported by modern NVIDIA graphics cards and is compatible across various platforms. Additional details about the dataset and the proposed model are provided in the subsequent sections.

3.3. Overview of Dataset Used

The DLRSD was introduced as a resource for research related to solving semantic segmentation challenges, which involve pixel-level labeling in RSIR. This dataset serves as an extension of the multi-label UC Merced dataset wherein multiple labels are assigned to each image. The DLRSD provides imagery with a one-foot pixel resolution and labels each pixel in the images with one of seventeen distinct class labels. These class labels encompass a range of objects and land cover types, including fields, tanks, chaparral, grass, airplanes, ships, pavement, water, cars, bare soil, buildings, docks, seas, mobile homes, sand, courts, and trees. The structure of the UC Merced dataset is mirrored by the 100-image limit for each class. Each image in the UC Merced dataset was semantically segmented using the eCognition 9.0 program to provide these labels. Each image was divided into areas, and depending on its features and content, each region was given one of the seventeen pre-defined class names. The densely labeled dataset known as DLRSD was produced as a result of this labeling process, and it is an important tool for research on semantic segmentation in the field of remote sensing.

Labeling masks and related pictures from the DLRSD are shown in Figure 3. The ground truth photographs were tagged into two main categories in order to use this dataset for our research: the image backdrop, which is represented by the color black, and the aircraft, which is represented by the color white. The dataset consists of a total of 100 images, each with a pixel size of 256 × 256. These images were partitioned into different sets for training, validation, and testing purposes. Specifically, 80 images were designated for the training set, 16 for the validation set, and 4 for the testing set. During the training stage, several data augmentation techniques were applied to enhance the diversity of the training dataset and prevent model overfitting. These augmentation methods are outlined in Table 2 within this study.

Figure 3. Example images and corresponding labeling masks from DLRSD [4].

Table 2. Data augmentation method applied during training phase.

Augmentation Methods	Value
Horizontal flip	True
Height shift	0.05
Width shift	0.05
Image zoom	0.05
Image shear	0.05
Image rotation	0.2

3.4. Implementation Tools

The experiments were implemented on an Ubuntu 18.04 platform, utilizing Python 3 as the programming language. For deep learning functionalities, the experiments employed Keras [42] along with TensorFlow [43] as the backend framework, as described in the study. The hardware configuration used to conduct the experiments involved an Intel CPU with 8 cores, specifically an i7 950 running at 3.07 GHz. Additionally, the experiments made use of a Nvidia GeForce GTX 960 GPU with 4 GB of memory. This hardware setup was utilized to facilitate the training, validation, and testing phases of the research.

4. Results

This study involves comprehensive experiments designed to assess the performance of the proposed model. These experiments were conducted using the DLRSD to test the segmentation model. This section discusses the training settings applied in the experiments, the evaluation metrics used, and the performance analysis.

4.1. Training and Evaluation of U-Net Model for Aircraft Semantic Segmentation

To train the proposed model, the Adam optimization algorithm was selected as a method for efficient gradient-based stochastic optimization. The aim of the training process is to reduce the loss, which involves learning all parameters associated with the model. This optimization process is applied over the entire 256 $\times$ 256 pixel input image, representing the input to the U-Net architecture. The Dice coefficient serves as the loss function in this context, helping to quantify the error between the predicted results and the provided ground truth mask. The weights for the model are initialized randomly, and the ReLU activation function is applied to the hidden layers' neurons. During the training phase of this study, considering the limited memory capacity of the GPU, a batch size of eight is employed. A set of image patches is fed into the network during each iteration to facilitate the backpropagation process. Additionally, the learning rate is set to 0.00001 to regulate the update step size during optimization.

The training of the model is conducted over 20 epochs, and during this training process, a dropout rate of 0.5 is applied. After the completion of these 20 epochs, the mean IoU was calculated, yielding a value of 95.08%. The output of the neural network is a 256 $\times$ 256 mask, representing the semantic segmentation of the input image. To ensure that the pixel values in this mask fall within the range of [0, 1], the sigmoid activation function is utilized. The pixel accuracy and mean IoU scores are employed as performance metrics to assess the model's effectiveness. Training is stopped when the performance score for the validation dataset no longer exhibits improvements, indicating that the model has reached a stable and optimal state.

4.2. Evaluation Metrics

The evaluation of our aircraft recognition results relies on commonly used pixel-based metrics, similar to those employed in studies involving building recognition. Specifically, we utilize the pixel accuracy and mean IoU scores to assess the model's performance, as shown in Formulas (1) and (2). The terms True Negative (TN), False Positive (FP), True Positive (TP), and False Negative (FN) are used to assess the performance of our model. TP represents the count of pixels that have been correctly predicted as aircraft when the

actual labeled pixels are also aircraft. FP represents the count of pixels that have been incorrectly predicted as aircraft when the actual labeled pixels represent the background. The term "TN" refers to pixels that have been appropriately identified as not falling inside the relevant class, for example, the backdrop or non-aircraft pixels. Last but not least, FN stands for the number of pixels that have been wrongly classified as background when they really represent an airplane. These words are crucial for determining if the model correctly identifies airplanes in the semantic segmentation outputs. The mean IoU quantifies the intersection over union for each class and computes their mean, while the pixel accuracy assesses the proportion of properly predicted pixels to all pixels. These metrics aid in assessing the precision and potency of our model's ability to identify airplanes from the semantic segmentation outcomes.

A key criterion for assessing the precision of semantic aircraft picture segmentation is the mean IoU score. It calculates the accuracy by dividing the union area, which includes both the ground truth and the detected aircraft, by the intersection area between the ground truth and the detected aircraft masks. The IoU's "mean" refers to the average score determined over all classes, which are normally "aircraft" and "background". This measure reveals how successfully the model recognizes and segments aircraft in the photos.

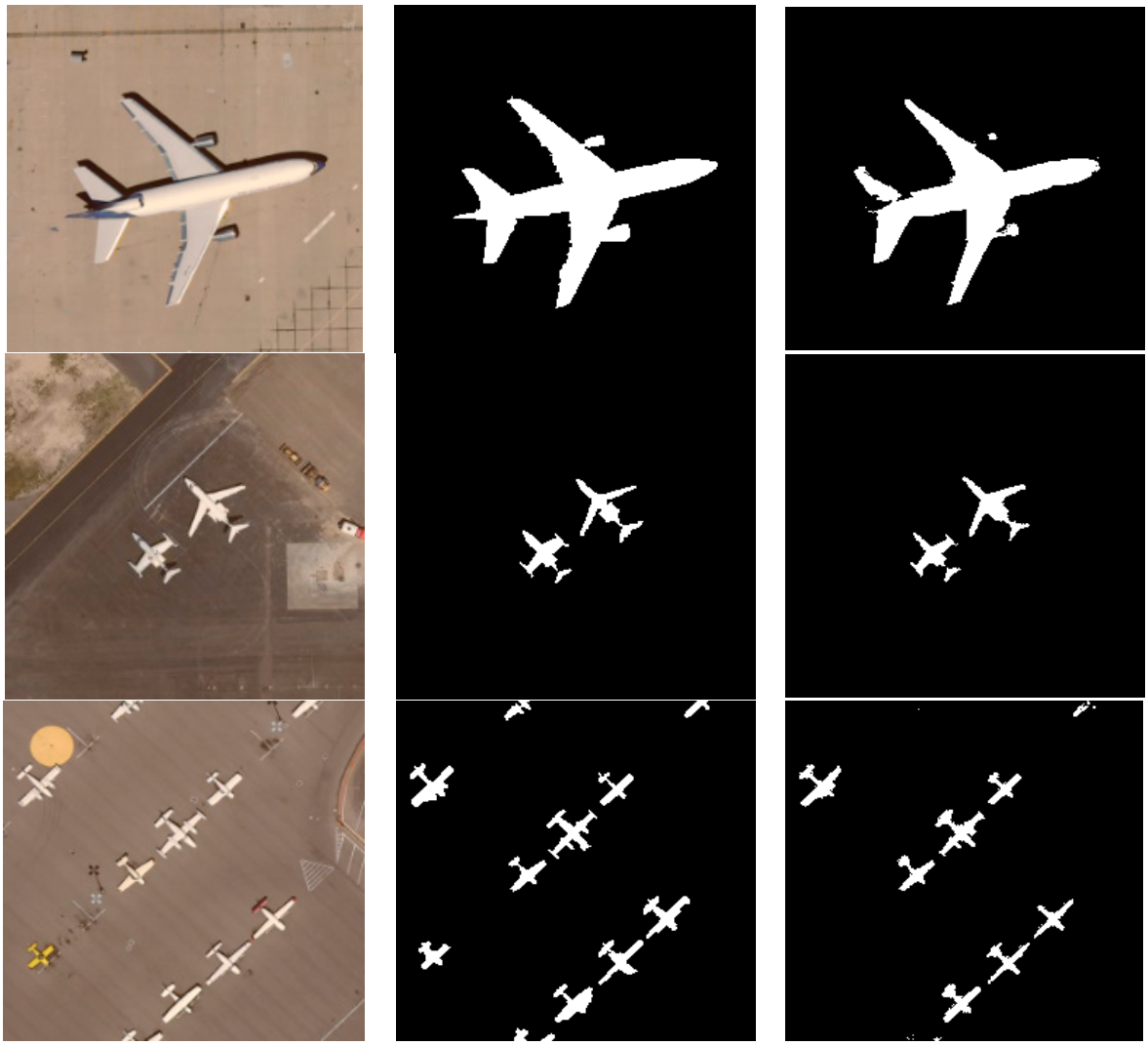
$$\text{Pixel} - \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\begin{aligned} \text{Mean} - \text{IoU} &= \sum_{c \in (\text{Aircraft,Background})} \frac{\text{target(c)} \cap \text{prediction(c)}}{\text{target(c)} \cup \text{prediction(c)}} \\ &= \sum_{c \in (\text{Aircrft,Background})} \frac{\text{TP(c)}}{\text{TP(c)} + \text{FP(c)} + \text{FN(c)}} \end{aligned} \tag{2}$$

4.3. Performance Analysis of Semantic Segmentation

Comparing the anticipated masks to the given ground truth masks is crucial to determine how accurate picture segmentation is. The validation dataset is essential for evaluating and validating the model's performance during the training process. We use the validation dataset to capture important measures, such as pixel-based accuracy and the mean IoU score, which are often employed in semantic segmentation tasks. Additionally, we carry out a rigorous assessment by contrasting the given ground truth mask with the anticipated binarized aircraft mask produced by the segmentation model. The model's accuracy and segmentation effectiveness are continually checked and improved throughout training thanks to this thorough validation approach.

Figure 4 illustrates the results of testing the proposed method, presenting the best segmentation results achieved by the model using three test images. In this representation, the first column displays three original images, the second column shows the corresponding ground truth images, and the third column reveals the outcomes of the proposed aircraft semantic segmentation method. As observed in Figure 4, the results of semantic aircraft segmentation closely resemble the provided ground truth images, with only minor variations in some small regions. The overall pixel-wise accuracy, representing the percentage of correctly predicted pixels, impressively reaches 98.24%. Additionally, the mean IoU score achieves a high value of 95.08%. These metrics emphasize the precision and effectiveness of the proposed segmentation model in accurately identifying aircraft in the images.



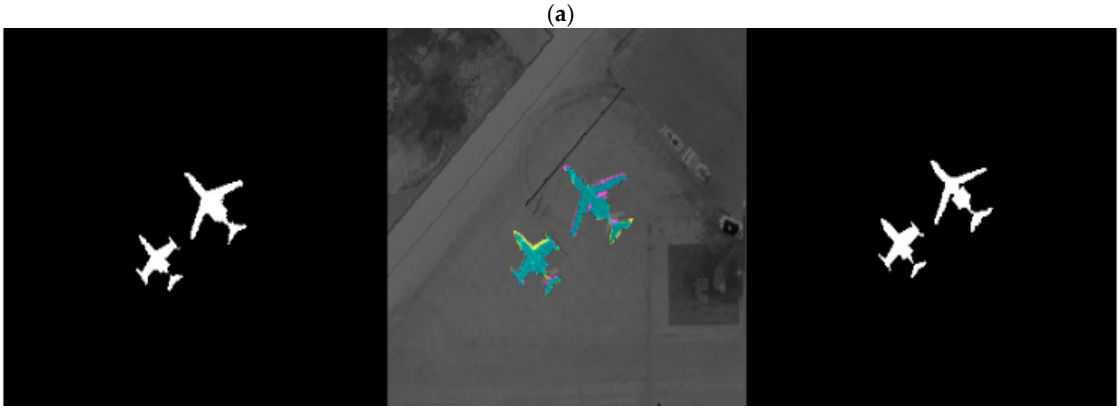
(**a**) (**b**) (**c**)

Figure 4. Results of experiments. (**a**) Original Images; (**b**) Ground Truth Images; (**c**) Predicted Images.

By conducting a thorough evaluation of the proposed method, we calculated the confusion matrix specifically for the two classes considered: aircraft and image background. Figure 5 displays the resulting confusion matrix for a single test image. We used the Scikit-learn package to obtain the TP, FP, FN, and TN components of the confusion matrix [44]. These confusion matrix components were assigned various colors for visualization purposes, allowing us to see how they are distributed throughout the picture. Since this makes it possible for us to tell the difference between TP and TN, TP, FP, FN, and TN are all mapped to the cyan, magenta, yellow, and black color spaces in Figure 5. We used the OpenCV library and an algorithm known as alpha compositing to transparently overlay the color mask onto the original grayscale image. Figure 5 illustrates the confusion matrix and its overlay on one of the test photos, providing details on how well the model performs in various areas of the image.

		Groundtruth: Foreground (255, aircraft)	Groundtruth: Background (0, background)
Predicted	Foreground (255, aircraft)	True Positive (TP) 1014	FP (False Positive) 124
Predicted	Background (0, background)	False Negative (FN) 124	TN (True Negative) 64274

(**a**)



(**b**) (**c**) (**d**)

Figure 5. Confusion matrix overlay mask (TP—cyan = 1014; TN—black = 64,274; FP—magenta = 124; and FN—yellow = 124). (**a**) Confusion matrix elements; (**b**) Predicted Image; (**c**) Overlaid Image; (**d**) Ground Truth.

The mean IoU curves for the U-Net model during training and testing across several epochs are shown in Figure 6. This visual depiction provides information about the model's functionality and development during the training and testing phases. Additionally, it is clear from the charts that the mean IoU first grows significantly as the number of epochs rises. The rate of improvement, however, declines as the training goes on, and the number of epochs exceeds 10, suggesting that the model's learning begins to plateau. Additionally, the graph shows that at epoch 20, which is the final epoch considered, the model achieved its highest mean IoU value. This information highlights the training process's effectiveness and indicates that, beyond a certain point, further training may have diminishing returns in terms of the mean IoU improvement. It provides valuable insights into the optimal training duration and when to conclude the training process for this specific model.

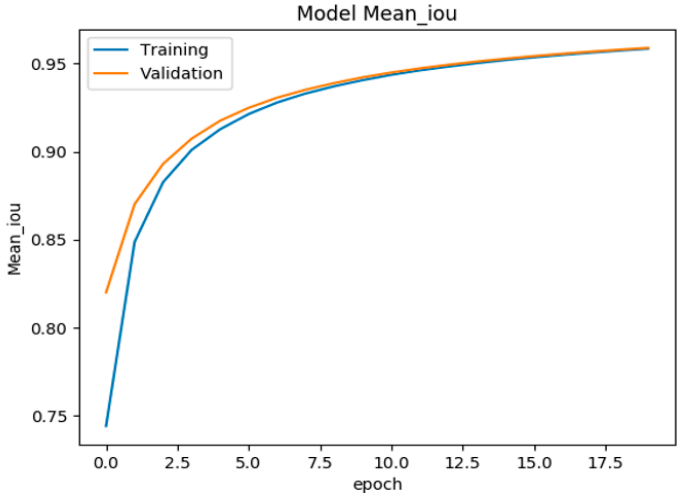


Figure 6. Graph of mean IoU versus epoch during validation and training stages.

5. Discussion and Future Works

The architecture we developed offers an end-to-end framework for aircraft semantic segmentation, characterized by strong scalability. Notably, it makes it possible for aircraft that were not in the initial training dataset to be automatically recognized, removing the requirement for the model to be retrained. For researchers struggling to obtain a sizable amount of labeled actual data for aircraft picture segmentation, our model offers a workable option. The findings imply that a fine-tuning strategy for initializing the encoder's weights in the network can improve the performance of the proposed U-Net model.

One limitation of our research pertains to the number of images utilized, which is determined by the DLRSD. While the number may appear modest compared to those employed in other studies, our proposed model mitigates the necessity for a large volume of labeled data and obviates the manual extraction of aircraft features. For instance, in the study presented in [40], the researchers assembled a dataset comprising 600 images, achieving the highest accuracy for U-Net at 95%, which is inferior to the accuracy attained in our investigation (i.e., 98.24%).

Similarly, the authors in [39] utilized the NWPUVHR-10 dataset, containing 650 target images and 150 background images, totaling 800 images across 10 target types, with 80 images featuring aircraft. Despite this larger dataset, their accuracy only reached 93.10%. Likewise, in the study conducted in [4], the DLRSD encompassed 21 broad categories, each with 100 images, resulting in a total of 2100 images. Despite this extensive dataset, their accuracy stood at 81.77%. Thus, while our dataset may be smaller in comparison, our study achieved superior accuracy levels.

One additional limitation of our study is the exclusive evaluation of the U-Net architecture without considering other architectures, such as VGG 16 or ResNet-50. While this choice may impact the generalizability of our findings, the results attained, when juxtaposed with those of other studies, can serve as a foundational framework for subsequent research endeavors.

By solely assessing the U-Net architecture, our study may have overlooked potential insights that could have been gleaned from exploring alternative architectures like VGG 16 or ResNet-50. However, the improvement to our work may require using transfer learning methods for more advanced pre-trained encoders like ResNet34 or VGG16. The use of ensemble learning methods is a further possible enhancement route. To construct an aggregate prediction, ensemble learning combines predictions from many pixel-wise classification networks. This strategy can lessen the bias brought on by certain models, which enhances the network's overall performance.

6. Conclusions

The process of training models for the semantic segmentation of satellite images, especially for tasks like aircraft recognition, heavily relies on manually labeled datasets. Creating such datasets is both time-consuming and expensive, posing a significant challenge to automating the analysis of such images. Additionally, due to the diverse and intricate nature of aircraft images, obtaining an adequate representation for recognition through image segmentation techniques remains a complex endeavor. With the advent of CNNs, new architectures like U-Net have emerged, which are capable of achieving excellent performance even with small training datasets. This research introduced a method for aircraft identification based on the U-Net model, and its efficacy was demonstrated with a modest training dataset (e.g., 100 images used in this study). We adapted the U-Net architecture for aircraft extraction, showcasing its ability to generate high-quality aircraft masks. Our tests demonstrated the efficacy of our method, which yielded a mean IoU score of 95.08% for the overlapped region and a pixel accuracy of 98.24%. Our work shows a promising approach for data-efficient and highly successful aircraft detection in satellite photos, highlighting the promise of CNN-based architectures like U-Net for completing challenging image segmentation tasks with little training data. Future research

may emphasize refining and optimizing our model, paving the way for even more accurate and robust aircraft image segmentation in remote sensing applications.

Author Contributions: Conceptualization, F.S. and A.Y.; methodology, A.Y. and Y.I.A.; software, F.S. and A.Y.; validation, A.Y. and A.E.T.; data curation, A.Y.; writing—original draft preparation, F.S. and A.E.T.; writing—review and editing, A.Y. and Y.I.A.; supervision, A.Y. and A.E.T. All authors contributed to the study conception and design. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, B.; Gao, J.; Chen, S.; Lim, S.; Jiang, H. POI detection of high-rise buildings using remote sensing images: A semantic segmentation method based on multitask attention Res-U-Net. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
2. Hao, X.; Yin, L.; Li, X.; Zhang, L.; Yang, R. A multi-objective semantic segmentation algorithm based on improved U-Net networks. *Remote Sens.* **2023**, *15*, 1838. [CrossRef]
3. Shao, X.; Qiang, Y.; Li, J.; Li, L.; Zhao, X.; Wang, Q. Semantic segmentation of remote sensing image based on Contextual U-Net. In Proceedings of the 2nd International Conference on Applied Statistics, Computational Mathematics, and Software Engineering (ASCMSE 2023), SPIE, Kaifeng, China, 26–28 May 2023; pp. 370–380.
4. Shao, Z.; Yang, K.; Zhou, W. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sens.* **2018**, *10*, 964. [CrossRef]
5. Tummidi, J.R.D.; Kamble, R.S.; Bakliwal, S.; Desai, A.; Lad, B.V.; Keskar, A.G. Salient object detection based aircraft detection for optical remote sensing images. In Proceedings of the 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS), IEEE, Nagpur, India, 5–6 April 2023; pp. 1–6.
6. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
7. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
8. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
9. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In *Computer Vision—ECCV 2010. ECCV 2010. Lecture Notes in Computer Science*; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Crete, Greece; Berlin/Heidelberg, Germany, 2010; pp. 143–156.
10. Liu, C.; Wechsler, H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Process.* **2002**, *11*, 467–476.
11. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
12. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the 9th IEEE International Conference on Computer Vision, IEEE, Nice, France, 13–16 October 2003; Volume 1472, pp. 1470–1477.
13. Topcu, A.E.; Alzoubi, Y.I.; Elbasi, E.; Camalan, E. Social media zero-day attack detection using TensorFlow. *Electronics* **2023**, *12*, 3554. [CrossRef]
14. Alzoubi, Y.I.; Topcu, A.E.; Erkaya, A.E. Machine learning-based text classification comparison: Turkish language context. *Appl. Sci.* **2023**, *13*, 9428. [CrossRef]
15. Zhao, A.; Fu, K.; Wang, S.; Zuo, J.; Zhang, Y.; Hu, Y.; Wang, H. Aircraft recognition based on landmark detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1413–1417. [CrossRef]
16. Zuo, J.; Xu, G.; Fu, K.; Sun, X.; Sun, H. Aircraft type recognition based on segmentation with deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 282–286. [CrossRef]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*; Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds.; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
18. Topcu, A.E.; Alzoubi, Y.I.; Karacabey, H.A. Text analysis of smart cities: A big data-based model. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 724–733.

19. Zhang, F.; Liu, S.-q.; Wang, D.-b.; Guan, W. Aircraft recognition in infrared image using wavelet moment invariants. *Image Vis. Comput.* **2009**, *27*, 313–318. [CrossRef]
20. Liu, F.; Yu, P.; Liu, K. Research concerning aircraft recognition of remote sensing images based on ICA Zernike invariant moments. *CAAI Trans. Intell. Technol.* **2011**, *6*, 51–56.
21. Shao, D.; Zhang, Y.; Wei, W. An aircraft recognition method based on principal component analysis and image model matching. *Chin. J. Stereol. Image Anal.* **2009**, *3*, 7.
22. Fang, Z.; Yao, G.; Zhang, Y. Target recognition of aircraft based on moment invariants and BP neural network. In Proceedings of the World Automation Congress 2012, IEEE, Puerto Vallarta, Mexico, 4–28 June 2012; pp. 1–5.
23. Wang, D.; He, X.; Zhonghui, W.; Yu, H. A method of aircraft image target recognition based on modified PCA features and SVM. In Proceedings of the 9th International Conference on Electronic Measurement and Instruments, IEEE, Beijing, China, 16–19 August 2009; p. 4-177-174-181.
24. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
25. Pan, X.; Gao, L.; Marinoni, A.; Zhang, B.; Yang, F.; Gamba, P. Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network. *Remote Sens.* **2018**, *10*, 743. [CrossRef]
26. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In *Computer Vision—ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; Volume 8695, pp. 345–360.
27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
28. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
29. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the International Conference on Computer Vision, IEEE, Santiago, Chile, 11–18 December 2015; pp. 1520–1528.
30. Badrinarayanan, V.; Kendall, A.; SegNet, R.C. A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 2481–2495. [CrossRef]
31. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. High-resolution aerial image labeling with convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7092–7103. [CrossRef]
32. Liu, Y.; Minh Nguyen, D.; Deligiannis, N.; Ding, W.; Munteanu, A. Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. *Remote Sens.* **2017**, *9*, 522. [CrossRef]
33. Patravali, J.; Jain, S.; Chilamkurthy, S. 2D-3D fully convolutional neural networks for cardiac MR segmentation. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges, Proceedings of the 8th International Workshop, STACOM 2017, Quebec City, QC, Canada, 10–14 September 2017*; Lecture Notes in Computer Science; Pop, M., Sermesant, M., Jodoin, P.-M., Lalande, A., Zhuang, X., Yang, G., Young, A., Bernard, O., Eds.; Springer: Cham, Switzerland, 2018; Volume 10663, pp. 130–139.
34. Kim, J.H.; Lee, H.; Hong, S.J.; Kim, S.; Park, J.; Hwang, J.Y.; Choi, J.P. Objects segmentation from high-resolution aerial images using U-Net with pyramid pooling layers. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 115–119. [CrossRef]
35. Pyo, J.; Han, K.-j.; Cho, Y.; Kim, D.; Jin, D. Generalization of U-Net semantic segmentation for forest change detection in South Korea using airborne imagery. *Forests* **2022**, *13*, 2170. [CrossRef]
36. Walsh, J.; Othmani, A.; Jain, M.; Dev, S. Using U-Net network for efficient brain tumor segmentation in MRI images. *Healthc. Anal.* **2022**, *2*, 100098. [CrossRef]
37. Shinde, R. Glaucoma detection in retinal fundus images using U-Net and supervised machine learning algorithms. *Intell. -Based Med.* **2021**, *5*, 100038. [CrossRef]
38. Buslaev, A.; Seferbekov, S.; Iglovikov, V.; Shvets, A. Fully convolutional network for automatic road extraction from satellite imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, Salt Lake City, UT, USA, 18–22 June 2018; pp. 207–210.
39. Wang, X.; Zhang, S.; Huang, L. Aircraft segmentation in remote sensing images based on multi-scale residual U-Net with attention. *Multimed. Tools Appl.* **2023**, *38*, 17855–17872. [CrossRef]
40. Ahmed, I.; Ahmad, M.; Jeon, G. A real-time efficient object segmentation system based on U-Net using aerial drone images. *J. Real-Time Image Process.* **2021**, *18*, 1745–1758. [CrossRef]
41. Alzoubi, Y.I.; Topcu, A.E.; Ozdemir, E. Enhancing document image retrieval in education: Leveraging ensemble-based document image retrieval systems for improved precision. *Appl. Sci.* **2024**, *14*, 751. [CrossRef]
42. Chollet, F. GitHub Repository. 2015. Available online: https://github.com/keras-team/keras (accessed on 15 March 2023).

43. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.
44. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.