Multidisciplinary Reviews

REVIEW ARTICLE
Published Online: March 11, 2024
https://doi.org/10.31893/multirev.2024118

Advancements in UAV image semantic segmentation: A comprehensive literature review

Shouket A. Ahmed[a] | **Hazry Desa**[a] | **Haider Kh. Easa**[b] | **Abadal-Salam T. Hussain**[c] | **Taha A. Taha**[d] | **Sinan Q. Salih**[e] | **Raed Abdulkareem Hasan**[d] | **Omer K. Ahmed**[d] | **Poh Soon Joseph Ng**[f]

[a]Centre of Excellence for Unmanned Aerial Systems (COEUAS), Universiti Malaysia Perlis, Jalan Kangar-Alor Setar, 01000 Kangar, Perlis, Malaysia.
[b]Department of Computer Engineering Technology, Al-Kitab University, Altun Kupri, Iraq.
[c]Department of Medical Instrumentation Techniques Engineering, Technical Engineering College, Al-Kitab University, Altun Kupri, Kirkuk, Iraq.
[d]Unit of Renewable Energy, Northern Technical University, Kirkuk, Iraq.
[e]Technical College of Engineering, Al-Bayan University, Baghdad, 10011, Iraq.
[f]Faculty of Data Science & Information Technology, INTI International University, Persiaran Perdana BBN, Nilai 71800, Negeri Sembilan, Malaysia.

Abstract Unmanned Aerial Vehicles (UAVs) have revolutionized data acquisition across various domains, presenting immense potential for image processing and semantic segmentation. This literature review encompasses a thorough exploration of advancements, techniques, challenges, and datasets pertaining to UAV image semantic segmentation. It begins by introducing the fundamental concepts of UAVs, highlighting their pivotal role in capturing high-resolution imagery that serves diverse applications. The integration of deep learning algorithms with UAVs is emphasized, unlocking new horizons in autonomous flight, security, and environmental monitoring. Delving into the core principles of semantic segmentation, the review elucidates the critical task of classifying every pixel in an image. Convolutional Neural Networks (CNNs) are presented as the cornerstone technology, tracing their evolution from traditional CNNs to the highly adaptable Fully Convolutional Networks (FCNs). A substantial portion of the review is dedicated to FCNs, underscoring their ability to process images of varying dimensions while maintaining spatial coherence in the output. Their pivotal role in semantic segmentation, encompassing both classification and localization, is articulated. The subsequent sections delve into a comprehensive survey of state-of-the-art models, including SegNet, PSPNet, DeepLabNet, EfficientNet, DenseNet-C, and LinkNet. Each model's unique strengths and applications contribute to the evolving landscape of semantic segmentation tasks. The versatility of the U-Net architecture takes center stage in the latter parts of the review. Its fundamental structure is elucidated, followed by a comprehensive examination of its manifold adaptations—3D-U-Net, ResU-Net, U-Net++, Adversarial U-Net, Cascaded U-Net, and Improved U-Net 3+. These modifications address intrinsic challenges such as limited receptive fields and class imbalances, propelling U-Net to the forefront of image segmentation techniques. The subsequent sections pivot toward the application of U-Net in UAV image segmentation, illustrating its efficacy in diverse tasks, including land cover and crop classification. Nevertheless, persisting challenges, such as the scarcity of annotated datasets and the need for model generalization across varied environmental conditions, remain key areas of concern. The review culminates by underlining the significance of large, authentic datasets and data augmentation techniques. Furthermore, a brief exploration of publicly available UAV image datasets is presented, enhancing our understanding of the resources accessible for training and evaluating models. This comprehensive literature review encapsulates the dynamism of UAV image processing and semantic segmentation, illuminating recent developments and avenues for future research in this burgeoning field.

Keywords: UAV image segmentation, deep learning, semantic segmentation, convolutional neural networks, U-Net architecture, environmental monitoring

1. Introduction

Deep learning incorporates numerous hidden layers and more rooted combinations that average artificial neural networks (ANNs) to produce more refined and better-performing autonomy in learning algorithms (O'Shea and Nash,2015). Deep neural networks (DNNs) are neural networks that are made up of "neurons”; these neurons contain a specific invitation (or activation) and specifications (or parameters) that convert input data, such as UAV imagery, into scenario-based maps while gaining knowledge (Schmidhuber et al. 2015). (See Appendix 1 on page 170 for a review study of the DNN architecture with an explanation of its types, history, applications, and development).

Image segmentation, which has grown into an area of study in image processing and computer vision, is the method of dividing an image into meaningful and nonoverlapping areas. It is a crucial step in natural scene perception. According to

Received: October 2, 2024 | Accepted: December 19, 2024

human visual perception, these regions are significant and nonoverlapping. Two obstacles exist in image segmentation: (1) how to define "meaningful regions" given the ambiguity of visual perception and the range of human comprehension, which makes image segmentation a poorly posed problem; and (2) how to accurately represent the objects in an image. The performance of deep-learning algorithms in image processing has considerably improved in the past decade because of the abundance of samples (labeled examples) and increased computer functionality (Traore et al., 2018). This literature review is mainly based on convolutional neural networks (CNNs) for image segmentation, particularly CNNs for UAV image segmentation, and the use of the U-Net model for this segmentation task.

2. CNN Architecture Overview

CNNs are highly efficient at segmentation, classification, natural language processing, and video processing. According to O'Shea and Nash (2015), CNNs are composed of multiple layers, such as convolution, pooling, and entirely linked layers, which train and extract characteristics based on raw input data. During training, the back propagation method is used to regulate the quantity of weight variation in response to the target. CNNs can extract hierarchical features from low to high levels of abstraction, analogous to the neocortex's deep and layered learning process; therefore, the popularity of CNNs is primarily due to this capability. Researchers have enhanced CNN performance by modulating its architecture, weights, and parameters, as well as by increasing and modifying the training data. The primary layer of a CNN, the convolution layer, consists of neurons (convolution kernels) that divide the input image into smaller fields of reception and convolve them using a predetermined weighting scheme (Sakib et al., 2019), as shown in figure 1. Large amounts of data and high-powered processing resources, such as GPUs, are needed for training CNNs.

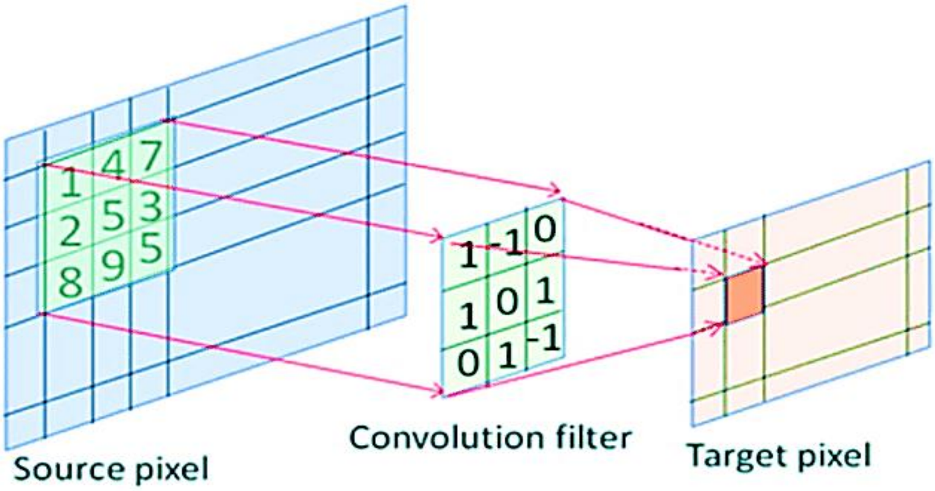


Figure 1 Convolution layer (Sakib et al., 2019).

CNN pooling follows the convolution kernel. This layer’s downsamples, as shown in Figure 2 below, reduce the convolution layer's feature data while retaining the most important information.

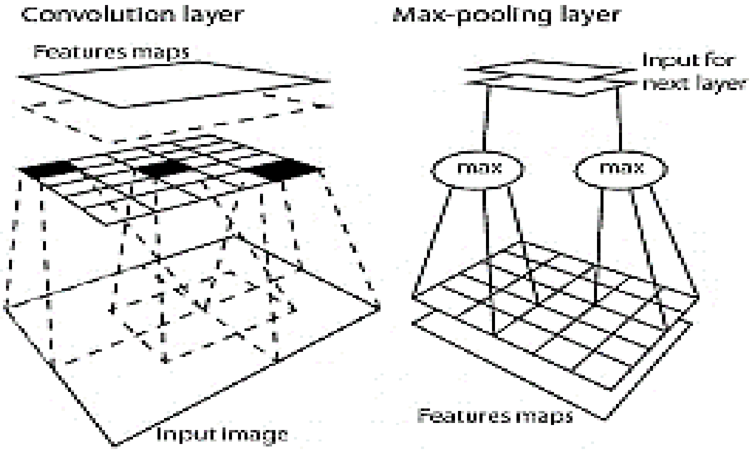


Figure 2 Pooling layer (Sakib et al., 2019).

CNNs, which are excellent image classifiers, require fully connected layers. Figure 3 shows how a few neurons are classified (Pelletier et al., 2019).

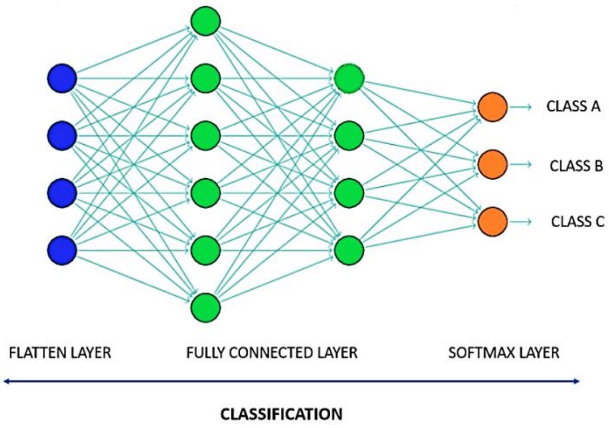


Figure 3 Visualization of the fully connected layer (Pelletier et al., 2019).

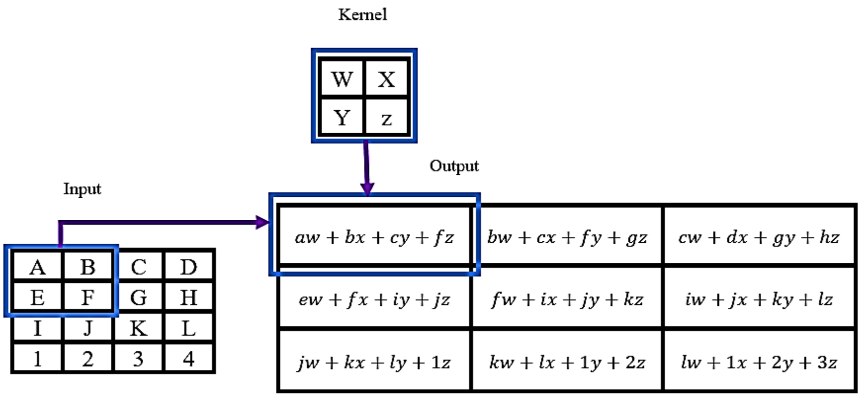
The CNN activation function defines which model data should be communicated and which should not (Nanni et al., 2020). Several functions, including softmax, ReLU, sigmoid, and tanh, are employed to introduce nonlinearity into the network. Function selection is determined by the specific objective, for example, sigmoid for binary classification and softmax for multiclass classification. Using mathematical techniques, the activation function establishes whether the neuron should be engaged based on the input's relevance to the prediction. CNNs can be developed via unsupervised or supervised approaches to machine learning, according to the desired outcome. CNNs are designed with specialized layers, such as convolutional, pooling, fully connected, and dropout layers, to address classification and segmentation challenges.

The process of convolution can be expressed mathematically as shown in equation 1, where s(t) represents the output function at a specific state of time, which is the feature map, x(a) represents the input position function, and w(a) represents the kernel, which means the weights.

$$s(t) = \int x(a)w(t - a)da \; (1)$$

In CNN, the feature map s (t) is produced by this formula. In CNNs, convolutions are performed on 2-D tensors made up of the input image’s height, width, and color channels. These operations take patches from the input and apply transformations to create a feature map with varying depths as an output. Therefore, the convolution is defined by two important aspects: first, the size of the filter applied to the layer, which represents the receptive field, and second, the number of filters, which represents the depth of the output feature map.

The computation of 2D convolutions is performed by moving a square window with a specific dimension across the input feature map. The dimensions of an image are height, width, and color channel. The color image has three RGB channels, which are red, green, and blue; thus, when the feature map of an image is subjected to a 2D convolution process, 2D patches of the surrounding features are produced. Then, by this process, these patches are converted to a 1-D vector. Afterward, the same process is repeated with as many filters as needed. Then, all the 1-D vectors produced by those filters are spatially reconstructed into a 2-D output map that corresponds to all the input map locations. This process is illustrated in Figure 4 below.



W	X
Y	z

A	B	C	D
E	F	G	H
I	J	K	L
1	2	3	4

$aw + bx + cy + fz$	$bw + cx + fy + gz$	$cw + dx + gy + hz$
$ew + fx + iy + jz$	$fw + ix + jy + kz$	$iw + jx + ky + lz$
$jw + kx + ly + 1z$	$kw + lx + 1y + 2z$	$lw + 1x + 2y + 3z$

Figure 4 Visual representation of dimensional convolution.

The stride is the step size by which the kernel moves across the CNN when applying the convolution operation. This step size can produce different output sizes of feature maps; in other words, it can downsample the feature maps. For example, if a stride of 2 implies, then the output feature map's dimensions are downsized by a factor of 2 without applying any padding process, as shown in Figure 5.

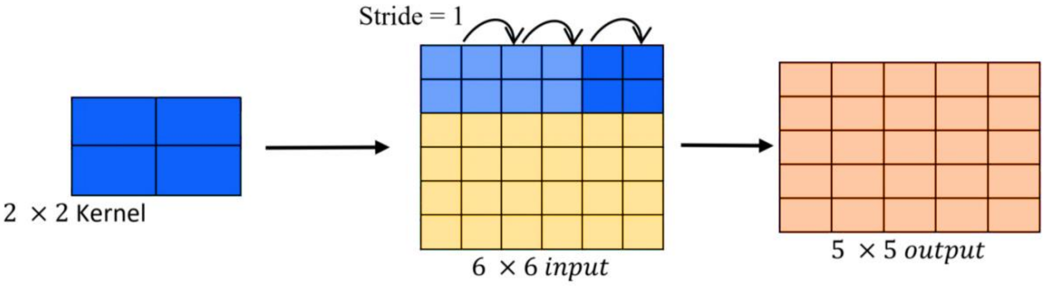


Figure 5 Strides.

Convolutional layers are used to produce a larger receptive filter, which allows the model to recognize more input image features. Examining all of the pixels in an image is a simple way to process it; however, this can take a very long training time, so to minimize this training time, some of the input image pixels are meaningless and have no significant feature of interest; therefore, when a filter is applied in the convolutional layer, it will reduce the image to the specific features of interest.

In the convolutional layer, many and different sizes of filters are applied, each filter has its weights to be learned, and each filter wipes out all pixels of the input image. Then, matrix multiplication is performed between the input image pixels and the applied filter weights. This calculation is repeated at each stride step, and the results of these calculations are fed into a specific activation function. The outputs of all activation maps are stacked together to produce the final volume of the convolutional layer's output.

Image compression is performed by pooling layer application. There are three types of pooling operations. The first type is max-pooling, in which the maximum number is picked up as a result of the sliding window. The second type is average pooling, in which the sliding window result is computed using the average statistic. The third pooling type is min-pooling, in which the minimum value is computed as a result of the window result. Therefore, using a pooling layer can incredibly reduce the amount of information delivered by the image, and this in return leads to minimizing the training time and lowering the memory use.

The pooling function is used mostly in problems that involve multiclass classification because of its desirable performance in real number value compression into a value range between 0 and 1 Banerjee et al. (2020) and ensuring that the sum of the whole probabilities of the output equals 1. The mostly sigmoid function is used as an activation function to compress all input values into this range. However, input values > zero produce results > 0.5, while those < zero produce results < 0.5. Finally, any input with a 0 value produces a result equal to 0.5. Therefore, this sigmoid function is mostly used as the output activation function for binary classification, and the softmax function is used as the output activation function for multiclass classification problems.

3. Regularization Techniques

The first technique of regularization is called dropout. The dropout process is a simple one; the addition of a dropout procedure after a neural network layer either destroys or removes a random number of neurons (Setiawan et al., 2022). The deleting rate or dropout rate is the rate by which the number of neurons is put off per layer, for example, if the dropout rate is set to 0.5, that means half of the neurons are deactivated randomly per layer in the network before feeding to the next layer. The higher the number of deactivated neurons picked, the greater the regularization impact is. These deactivated neurons are temporary during the training phase, and their updated weight is not fed to the backward path neurons, meaning that the greater the number of deactivated neurons is, the fewer the number of training samples for the subsequent layer, which can cause the underfitting problem. Therefore, in practice, the hidden layer should have a typical dropout rate p in the range of 0.2 and 0.5, while that of the input layer is 0.2.

The second regularization technique is batch normalization (Luo et al., 2018). One of the biggest challenges during training is because the input from the previous layer is changed when tweaking the weights throughout the backpropagation process, so the batch normalization technique comes to be applied either before or after activations to reduce the generalization error and speed up the training process by a considerable number of epochs in some cases, reducing the epochs to a half or even more. In addition, batch normalization provides great regularization to the model, so there is no need for another type of regularization technique to be applied.

4. Model Training

In general, the training process of DNNs is challenging for several reasons. One of these reasons is that the values of the parameters of every layer are changed at each iteration during the training process, and the training process of any neural network depends on some important components, such as layers, input data, loss function, and optimizer. The layers are the core component of the neural network (NN); each single NN layer receives the input data from the previous layer and processes these data and then produces output data in data containers known as tensors. In each convolutional layer, there are values known as weights. NNs are trained to identify the patterns in the data, while the goal is to optimize the cost function and use an optimization algorithm to optimize the cost function to obtain the optimal weight.

Furthermore, NNs are trained to identify the optimal weight values that will specify the transformations to be performed on the incoming data. Figure 6 depicts the network layers' parameterization process. The setup of these weights is challenging to optimize for varied tasks because there can be millions of parameters that are interdependent. A loss function is utilized to generate a weight configuration; it measures the extent of performance of the layer by comparing the actual values with the predicted values. The first stage of the backpropagation (BP) algorithm is the determination of the final loss value, followed by the computation of each parameter's contribution in the computed loss value from the input layer through all hidden layers down to the output layer. The stochastic gradient descent is one of the methods that is used to determine the loss function derivatives at a particular point of the training process to tweak the weights` values in the direction of reducing the loss function to the minimum and hence optimizing the model.

Figure 7 shows the representation of an NN training process. The weights are randomly initialized (using a random initializer) in this training loop, which results in a high loss score because the weights at first are mostly unsuitable for determining the patterns. The network, on the other hand, adjusts its weights with each training batch and gradually improves until it has a small loss.

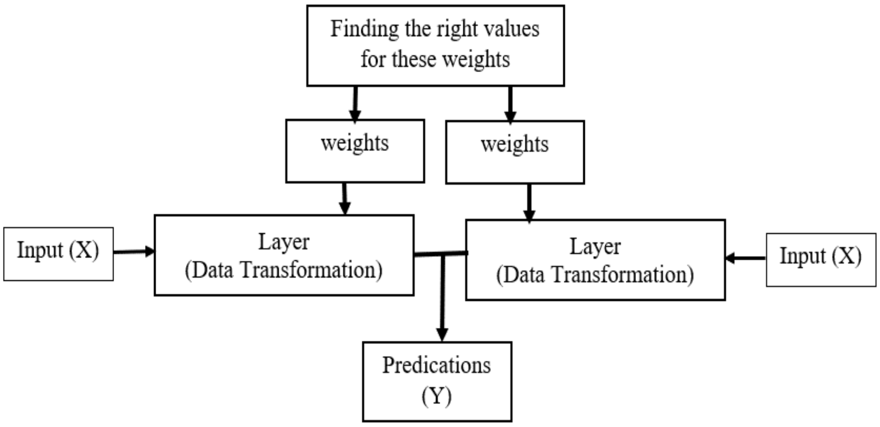

Finding the right values for these weights
weights
weights
Input (X)
Layer (Data Transformation)
Layer (Data Transformation)
Input (X)
Predications (Y)

Figure 6 Weight value-based parameterization of DL layers.

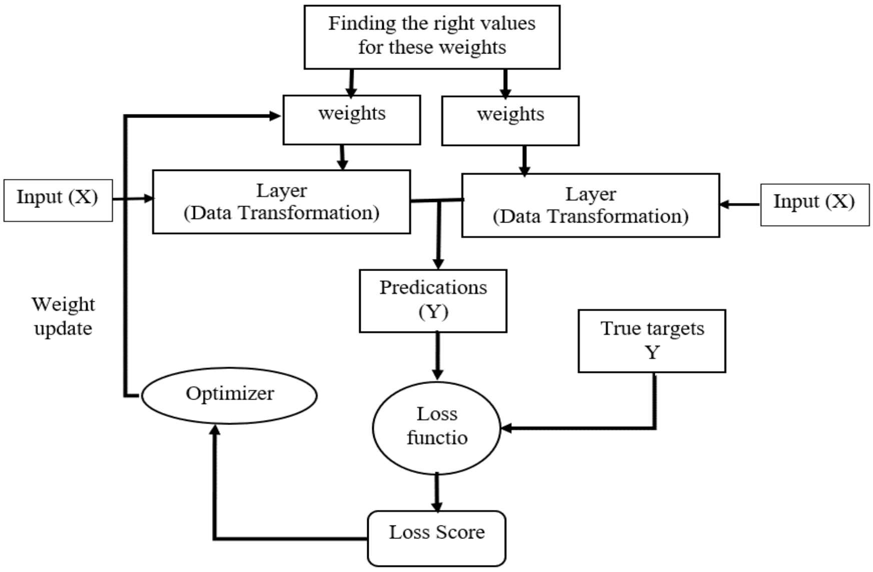


Figure 7 Flow chart of the training process for an NN.

5. Model Architecture

The model architecture is composed of layers arranged in stacks; in these layers, each stage learns a useful pattern by successively filtering the input data. These layers are shown in Figure 8, which depicts a basic 4-layer NN for the classification of handwriting from the MNIST dataset.

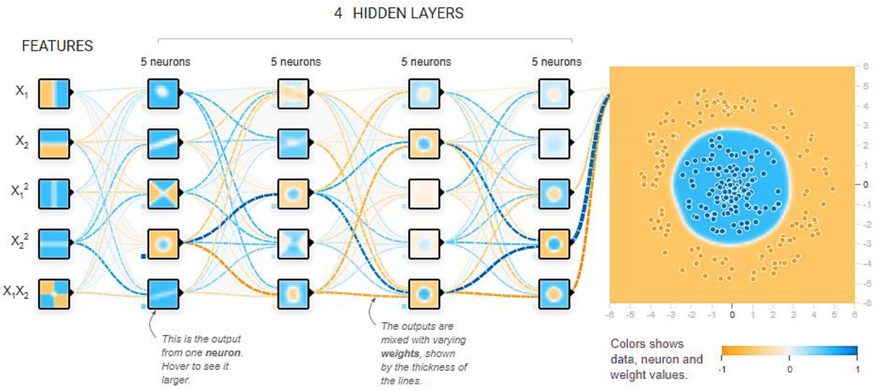

4 HIDDEN LAYERS
FEATURES
5 neurons
5 neurons
5 neurons
5 neurons
X_1
X_2
X_1^2
X_2^2
X_1X_2
This is the output from one neuron. Hover to see it larger.
The outputs are mixed with varying weights, shown by the thickness of the lines.
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
6 5 4 3 2 1 0 -1 -2 -3 -4 -5 -6
Colors shows data, neuron and weight values.
-1
0
1

Figure 8 Four-layer neural network.

Traditionally, densely linked layers, also known as completely connected layers, have been used for datasets other than images, in which all the neurons of one layer are connected to all the neurons of the past layer. On the other hand, dense layers can only learn patterns that fall within their input feature space, while each convolution filter is capable of learning patterns locally (kernel space or region of interest). This implies the possibility of fragmenting input images into edges and textures that are easily learned and are more beneficial for classification than for global patterns.

6. Gradient descent algorithm (GDA)

This is a search strategy for finding the local minimum of a function that can be differentiated, which is called a loss function. GDA is mostly used in ML to determine the parameters (weights) of a function that optimally reduces a cost function; the algorithm can be expressed mathematically as shown in equation (3.2). The following steps are repeated until convergence is reached:

1. Calculate the parametric changes as a function of the learning rate based on the gradient.
2. Use the updated parameter value to recalculate the new gradient.
3. Check for the termination criterion; otherwise, revert to step one.

The GDA is represented by the θ parameter for the optimization algorithm in equation (2.2). (α) is the learning rate – a tuning parameter in the optimization process. It decides the length of the steps. (α) is the gradient of the cost function (J) earnings to θ.

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j}\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad (2.2)$$

for (j=1 and j=0).

The leering rate is a configurable parameter; its value is in the range of 0.0 to 1.0. It is used to train NNs (note that logistic regression is an NN with just one neuron). Hence, the learning rate can determine how much the model should adjust to adapt to a particular situation. In other words, making some modifications in the gradient descent algorithm is workable for powerful deep neural networks. There are a few drawbacks of GDA; the prominent drawback is the needed number of computations per iteration of the algorithm. For instance, assume a case of 20,000 data points with 20 features; here, the sum of squared residuals will contain the same number of terms as the data points (that is, 20,000 terms in this case). Hence, the derivative of this function must be computed based on each of the features. This will require the following computation: 20000 x 20 = 400,000 computations per iteration. If we consider 1000 iterations, it will be arriving at 400,000 x 1000 = 400000000 computations to fully implement the GDA, which is much overhead; hence, GDA is not usually fast on enormous datasets, leading to the development of stochastic gradient descent (SGD). The term “stochastic” here means “random”, which comes to play when data points are detected at each step for the calculation of the derivatives. The SGD picks one data point randomly from the entire dataset per iteration to minimize the computational requirements.

In stochastic gradient descent, only one data point is fed at each iteration; thus, it is a slow process, and to make it faster, batches have to be created in such a way that a batch consists of 16, 32, 48, 64 data samples, and training on batches makes the training process fast and reduces the computation. The Adam optimizer is a hybrid of gradient descent techniques with momentum; this technique is participating in speeding up the GDA in consideration of the exponentially weighted average of the gradients. The algorithm tends to converge rapidly toward minima because it uses the averages. Furthermore, RMS prop, another algorithm, was proposed by Geoffrey Hinton as a gradient-based NN training approach. The gradients of complex functions, such as NNs, tend to either vanish or explode as the data propagate through such a function. RMS prop was created as a stochastic mini-batch learning algorithm that solves the problem by normalizing the gradient with a moving average of squared gradients. These processes of normalization equalize the step size and either lower it for high gradients to prevent explosion or raise it for minor gradients to avoid vanishing. Simply expressed, RMS prop treats the learning rate as an adaptive parameter rather than a hyperparameter. This indicates that the rate of learning fluctuates with time.

7. Overfitting Issue

If the model performs very well during the training phase but performs poorly in the testing phase when it is fed with unseen data, this model is said to be overfitted, and this happens mostly when the model makes predictions very close to the actual data point. In contrast, underfitting occurs when the model predicts too far from the actual data point. In some cases, overfitting occurs due to the small dataset for training. Therefore, the model must be balanced between overfitting and underfitting and must be regularized by using a regularization technique to avoid these issues because both of them hurt the model.

8. Pretrained Models

Keras applications come with several pretrained segmentation models that can be used. ResNet is one of the most widely used of these accessible models. All of these models have been pretrained on the ImageNet dataset, which contains a variety of images, to detect over 1000 classes. These models and U-Net are tested together to see how well they can learn semantic segmentation from UAV data. Trained ResNet50 as the backbone for the U-Net model is used at the beginning of this research in an attempt to solve the UAV image segmentation problem, but unfortunately, it was poorly performed, which might be because ResNet models are mainly used for image classification problems rather than image segmentation and because of the complexity of the resultant model, which has a high computational cost.

9. Model Evaluation

Metric, in general, measures the success of the model; in other words, it tells how well or bad the algorithm is performed. In general, each deep-learning model is built to implement a specific task. Therefore, for each type of problem, different metrics are developed; therefore, for the classification task, the most commonly used metrics are accuracy, precision, f1 score, ROC curve, and confusion matrix. For the object detection task, the most commonly used metrics are the IOU, dice coefficient, and mean average precision. Moreover, the metrics most considered for image segmentation tasks are the accuracy, IOU, and Dice coefficient. Segmentation means pixel-level classification; i.e., Each pixel needs to be classified into one class of many classes; hence, the considered metrics for the segmentation problems are accuracy, intersection over union, dice coefficient, and loss. These metrics will be used in this project to evaluate the success of the constructed U-Net models. Accuracy is a measure of the percentage of correctly predicted values out of all the considered values. It is calculated as shown in equation (2.3).

$$\text{Accuracy} = \frac{True\ positive + True\ negative}{(True\ postive + False\ positive + True\ negative + False\ negative)} \quad (2.3)$$

Intersection over union is the intersection area divided by the area of the union of the actual and ground truth masks; this metric is extensively used in object detection and segmentation, as shown in Figure 9.

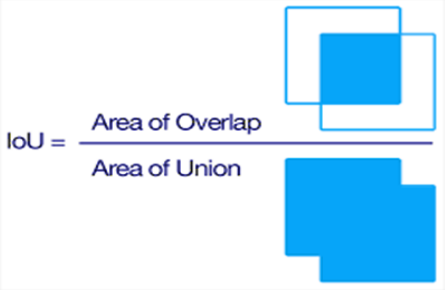


Figure 9 Intersection over union.

The dice coefficient is the evaluation metric by which the similarity of two sets of data is measured. It is mostly used in object segmentation and object detection. The dice coefficient is simply calculated as two times the intersection area divided by the area of the sum of the actual and ground truth mask, as shown in Figure 10.

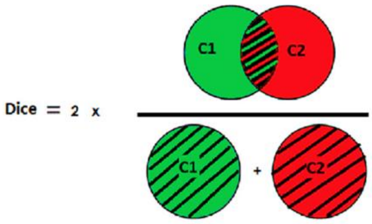


Figure 10 Dice coefficient.

10. Revolution of CNN for Image Classification

The CNN is optimal for image recognition and speech recognition. Its convolutional layer decreases the image dimension without modifying the data. Pattern recognition in CNNs requires the translation of images into numerical data. Due to their pixel composition, images are converted to a numerical format before being input into the network (Traore et al., 2018). Table 1 summarizes the historical development line for the state-of-art CNN models that were developed for image classification tasks.

Table 1 Review of CNN Image-Classification Algorithms.

Algorithm	Summary
LeNet-5 (1998)	Lenet-5 is presented in 1998. The architecture is simple to design (conv-pool-conv-pool-FC-FC).it is utilized to identify both handwritten and machine-printed text.
AlexNet (2012)	AlexNet is developed to enhance ImageNet performance and obtained a precision of 84.7%. Using convolutional layers and reception fields to investigate spatial correlation, it becomes one of the first deep convolutional networks to attain high accuracy (Alom, 2018). Properties of AlexNet include the use of the activation function of ReLU for all layers, simultaneous training on two GPUs, and data augmentation and dropout strategies to combat overfitting. Its design is regarded as the basis for CNN architectures, via subsequent models expanding upon its fundamental framework. AlexNet has the greatest record of citations among CNNs.
ZFNet (2013)	ZFNet is a refinement of AlexNet that improves accuracy. ZFNet employed smaller, 7x7-sized filters, whereas AlexNet used larger, 11x11-sized filters. It is trained with batch stochastic gradient descent
ResNet 18, 34, 50, 101, and 152 (2015-2016)	ResNet is feasible to train with hundreds or even thousands of layers while still achieving a compelling level of performance due to its flexibility. ResNet variants are 18, 34, 50, 101, and 152. It is a deeper network rather than wider which means more layers with less computation, it is used a skip connection technique between blocks to avoid the vanishing gradient problem.
GoogleNet Inception Module V1, V2, V3 (2014-2015**)**	GoogleNet topped the ImageNet competition in 2014 with its initial module, which employs skipping links to create a small module that is repeated throughout the network multiple times. It uses average pooling to decrease the number of layers and parameters with complete links. GoogleNet, with its nine inception modules, is a wider network that prioritizes efficiency. It depicts multiscale objects in an image by employing various filter sizes, 1x1 convolutions, and scaling convolutions, alongside filters for minimizing parameters and computation (Soundty et al., 2021).
VGG16,19 (2015)	Oxford's Visual Geometry Group designed the VGG16 and VGG19 deep architecture neural networks. The convolutional layer parameters are represented as "cone filter size> - number of channels>." To capture up, down, left, and right, VGG employs extremely small receptive fields of 33 with stride and pad = 1. For maintaining the same resolution, it also employs 1x1 convolution filters with stride=1, followed by a ReLU function, alongside max-pooling size 2 with stride 2 for all CNN layers. VGG possesses fewer parameters than AlexNet.
ResNext (2017)	Res-Next, on the contrary hand, is an enhanced variant of ResNet that employs group convolutions and parallel pathways.
SENet (2018)	Squeeze-and-Excitation Networks, or SENets, are a new building element for convolutional neural networks (CNNs) that improves channel interdependencies while incurring essentially no additional computational cost. They were put to use in this 2015's ImageNet competition, where they contributed to a 25% improvement in performance compared to the previous year (Hu et al., 2018). In addition to this enormous improvement in performance, they can be simply integrated into already existing designs.
Mobile Net's V1, V2, V3, V4 (2018-2019)	MobileNet is a convolutional neural network optimized to use in real-time, embedded vision systems on mobile devices. They are built on a simplified design that makes use of depthwise separable convolutions to construct lightweight deep neural networks that can have minimal latency for mobile and embedded devices.
ShuffleNet V1, V2 (2018-2019)	ShuffleNet is a convolutional neural network developed specifically to use on mobile devices with minimal processing capacity. The design used two novel operations—pointwise group convolution and channel shuffle—to lessen the burden of computing without sacrificing precision.
DenseNet (2020)	DenseNet is designed expressly to reverse the reduction in accuracy that high-level neural networks experience as a result of the diminishing gradient. This indicates that the information is lost because the journey from the input layer to the output layer is so much longer than the length it would take to reach its destination. Dense Net used dense blocks in which each layer is connected to each other layer in forward fashion and this technique can alleviate vanishing gradient, strengthen feature propagation and encourage feature reuse (Huang et al., 2017).

V1= Version 1, V2=Version2, V3=Version3, V4=Version4.

11. Potential Use Cases

By identifying defective electronic components, automated failure detection technology increases manufacturing output and decreases waste and expenses (Sultana et al., 2018). In medical care, image classification facilitates the detection of bone fractures, cancer, and tissue abnormalities, thereby enhancing the accuracy of MRI (Lin et al., 2015). In the agricultural sector, image classification reduces the need for human intervention by identifying plant maladies and water-deficient crops (Lin et al., 2015). In the manufacturing process of circuit boards (Lin et al., 2015), defective boards can accrue substantial costs; image classification reduces the reliance on human operators to identify defective boards. The various uses of image classification showcase its potential to increase productivity, decrease expenses, and improve accuracy across a variety of industries.

12. CNN for Image Segmentation

Automated image segmentation is an essential component of computer vision (Sharp et al., 2014), as it simplifies analyzing images by dividing an image into segments or collections of pixels assigned to distinct classifications. It removes the need for pixel-by-pixel analysis and continuously evolves, alongside new models arising frequently. Figure 11 depicts image segmentation, which can be implemented for object detection or image classification. This approach begins by identifying "seeds," which are tiny sections used to determine the tile layout. Segmentation aids in recognizing image regions and is essential for duties such as identifying plant illnesses, broken bones, and tissue abnormalities in medical care.

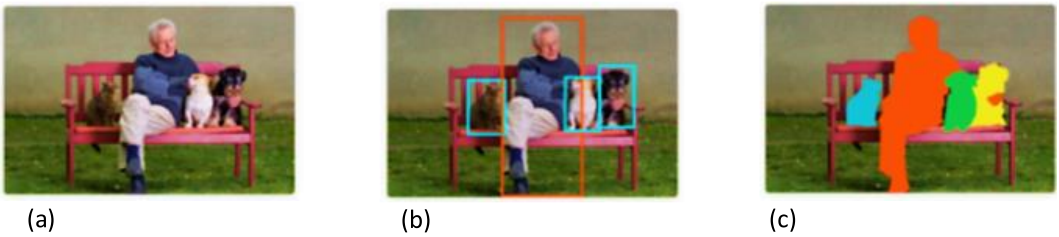


Figure 11 (a) Object classification, (b) Object detection, and (c) Object segmentation (O'Shea, & Nash, 2015).

Using a CNN in a procedure known as region proposal and annotation, semantic segmentation classifies image pixels into semantically interpretable classes. Candidate object patches (COMPs) are tiny clusters of pixels that are most likely associated with the same object. Instance segmentation identifies each instance of an item, whereas semantic segmentation combines all instances of the same class. Semantic segmentation does not designate each pixel in an image, whereas instance segmentation distinguishes each item (Minaee et al., 2021).

13. Revolution of CNN for Image Segmentation

13.1. Fully Convolutional Network (FCN) for Segmentation

The advent of a variant of the CNN is known as the fully convolutional network (FCN). This FCN represents a large advancement in the field of image segmentation problems. The difference between FCN and the traditional CNN is that the completely linked layer at the end of the CNN is replaced by convolution layers, and as a result of this replacement, the completed network can be fed with any size of input image and produce the same spatial dimension outputs. Hence, the classification network can generate a heatmap of the selected item class. Figure 12 depicts an example of this transformation.

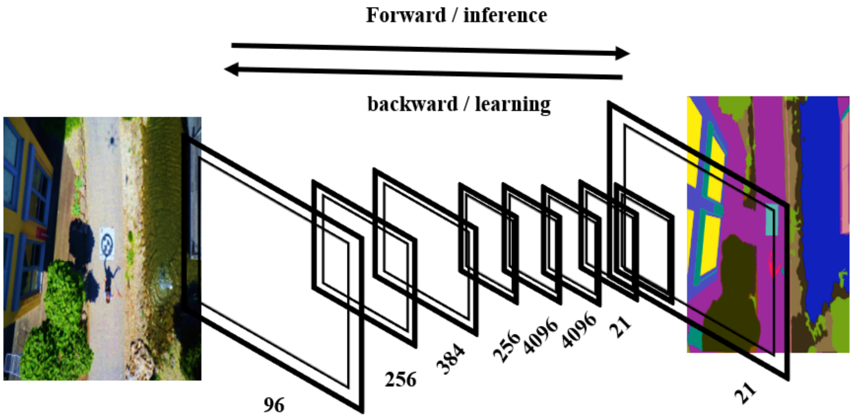


Figure 12 FCN for semantic segmentation.

FCNs are not only more flexible because they can take a variety of input image sizes but also more efficient for learning dense predictions thanks to in-network upsampling. The FCN may also keep track of the input's spatial information, which is important for semantic segmentation because it requires both classification and localization. Although FCNs can accept any size of the input image, nonpadded convolutions can be used to decrease the output resolution; these were created to keep the sizes of filters modest and hence reduce the computational cost. The outcome is a coarse output with a size reduction equal to the pixel stride of the output units' receptive field. Day times. U-Net networks, as fully convolutional networks, can be utilized for this task (Minaee et al., 2021). However, one of the major drawbacks of fully convolutional networks is that they require considerable data for accurate learning of image segmentation and localizations, whereas only a few datasets are publicly available online. Additionally, the technique of manually labeling each pixel with multiple classes is a tiresome and time-consuming operation and mostly comes out with a lack of precision. Hence, FCNs and U-Nets are commonly employed in UAV imagery segmentation, as they require datasets of lesser size if they are compared with the other algorithms.

13.2. Seg Net

SegNet is a segmentation model comprised of an encoder network, a decoder network, and a pixel-by-pixel classification layer. Like VGG16, which consists of thirteen convolutional layers, the encoder network applies low-resolution characteristics to high-resolution data maps for pixelwise classification. The decoder utilizes nonlinear upsampling with pooling indices generated during the encoder's max-pooling stage. SegNet is distinguished by its innovation.

13.3. PSP Net

PSP Net, a semantic segmentation model, employs a pyramid pooling module (PPM) to address the narrow receptive field of convolutional networks such as ResNet, which cannot adequately represent the global context. The PPM pools kernel from multiple subregions and sizes for a robust global representation. It quickly gathers features from the entire image, half of it, and smaller portions, creating a universal prior. The PPM then joins before the initial feature map, solving the global context problem accurately and efficiently (Zhao et al., 2017).

13.4. DeepLab Net

DeepLab Net employs atrous convolution, which extends the field of view of the filters to incorporate more context and controls the field of vision. It offers a mechanism for finding the optimal equilibrium across precise localization and context integration. Through atrous convolution, inputs are sampled alternately, resulting in a larger output feature map. (Chen et al., 2017) The architecture of DeepLab Net employs atrous convolution to accomplish enhanced accuracy and resolution in semantic image segmentation, as shown in Figure 13.

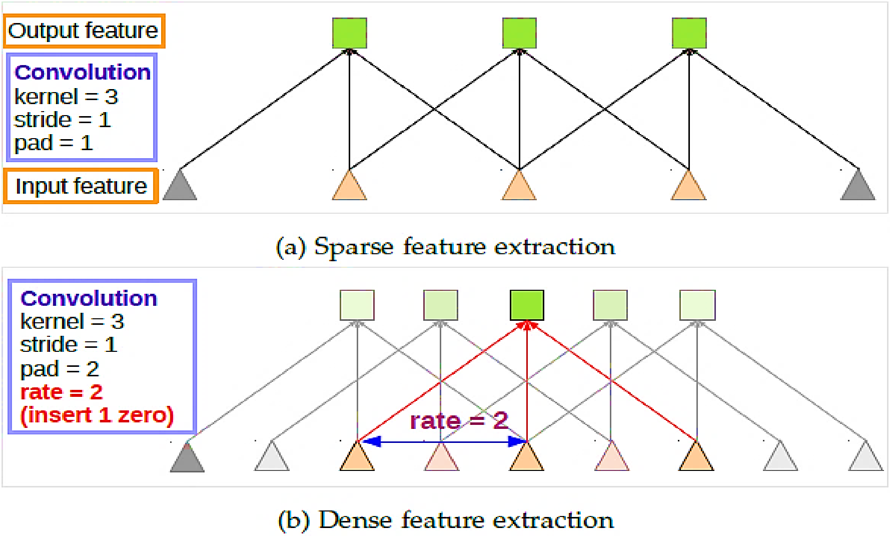


Figure 13 Atrous convolution (Chen et al., 2017).

13.5. Efficient Net

Following conducting a neural architecture search within the AutoML MNAS framework, Efficient Nets were developed by increasing the initial network size, which optimizes accuracy and efficiency for optimal performance. The new baseline network employs mobile inverted bottleneck convolution (MBConv) and has a substantially larger FLOP budget (Lee et al., 2020), as shown in Figure 14.

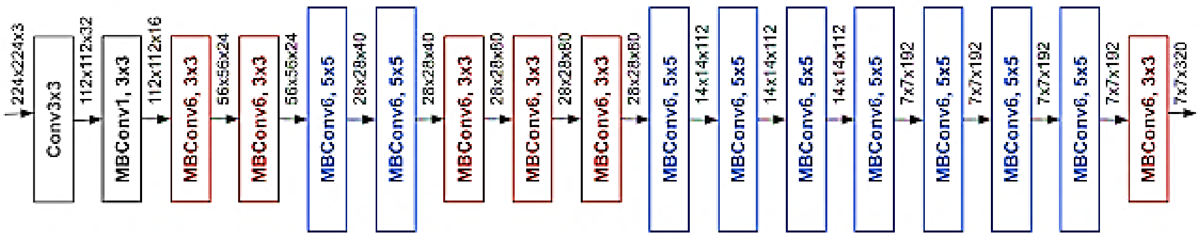


Figure 14 Efficient Net (Lee et al., 2020).

13.6. Dense Net-C

Dense Nets-C is a variant of Dense Nets-B that uses a compression factor (theta) to minimize the output quantity of feature maps. When theta equals 1, it operates as DenseNet-B. When theta is not 1, dense Nets-C alongside theta*m feature maps at a layer are generated. Figure 15 depicts the architecture of Dense-Nets (AbdelMaksoud et al., 2020).

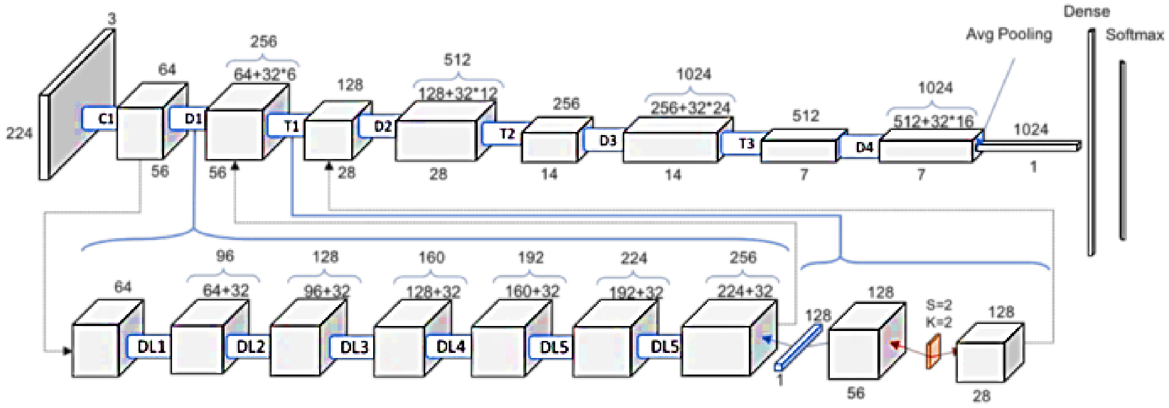
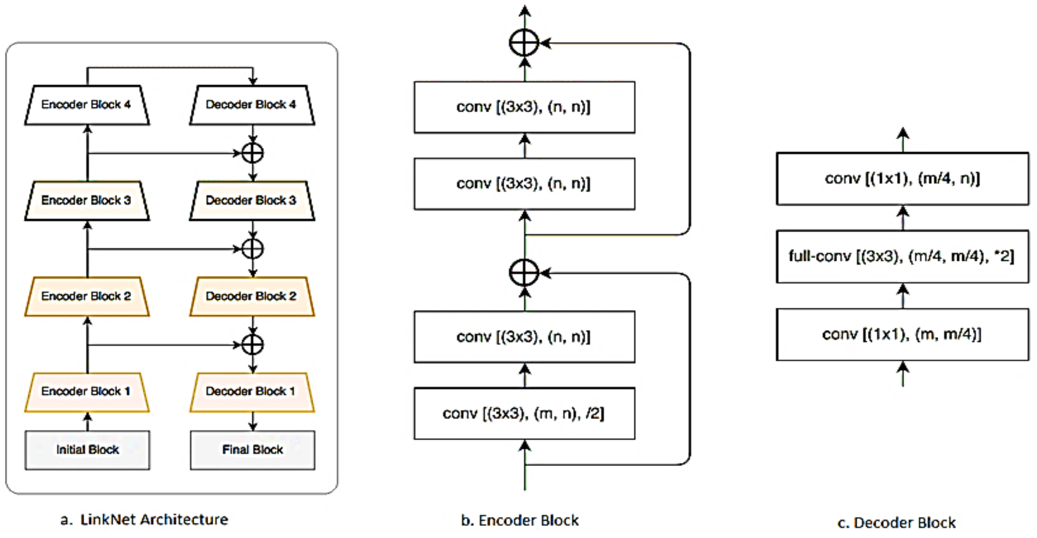


Figure 15 Dense Net architecture (AbdelMaksoud et al., 2020).

13.7. Link Net

LinkNet is a lightweight neural network for semantic segmentation that can operate in real time on GPUs and embedded devices. It utilizes an encoder-decoder design with feature forwarding to transmit information between layers, enhancing precision and lowering the number of parameters. The initial block comprises a 7x7 convolution layer and a max-pool layer, whereas the final block concludes the convolution with a 2D-convolution layer. Full convolution with a 2x2 kernel is used as a classifier. Figure 16 illustrates the architectural design. (Chaurasia & Culurciello 2017).

Table 2 summarizes the concepts of all the abovementioned state-of-the-art models that are used for semantic segmentation tasks.



a. LinkNet Architecture

b. Encoder Block

c. Decoder Block

Figure 16 LinkNet architecture (Chaurasia & Culurciello 2017).

Table 2 Summary of CNN image-segmentation algorithms.

CNN for Image segmentation	Summary
FCN (V1 Nov 2014), (V2 Mar.2015)	Adapts various classification nets, converts FC layer to 1x1 Convolution layer, only one transpose conv for upsampling
SegNet (Nov 2015)	Segmentation models. Encoder, decoder, and pixelwise classification layers make up this basic trainable segmentation architecture.
PSPNet (V1 Dec.2016), (V2 Apr.2017)	It is a bully encoder-decoder architecture, it uses pooling indices to provide a prior global representation that is both efficient and accurate
DeepLabNet (V1 Dec.2014), (V2Feb.2015), (V3 Apr.2015), (V4 Jun.2016)	It introduces (dilated) Conv and uses Atrous Spatial Pyramid Pooling (ASPP) module as well as Conditional Random Field (CRF)
DenseNets(V1 Nov.2016, V2 Dec.2016,V3 Oct.2017)	In denseness each layer receives additional inputs from all preceding layers and passes its feature maps to all subsequent layers, it features high computational and memory efficiency, strong gradient flow, more diversified features, concatenate the output feature maps of the layer with the input feature maps instead of adding them together
E-Net (Jun.2016)	Compounding scale gives EfficiencNet fast, accurate inference. The baseline network strongly influences it. AutoML MNAS neural architecture search optimizes accuracy and efficiency for FLOPS. The design uses MBConv like MobileNetV2 and MnasNet.
Link Net (Jun. 2017)	Link Net recovers spatial information by passing each encoder's input to its decoder's output. For autonomous cars, augmented reality, and other applications, it is a lightweight deep neural network architecture for semantic segmentation.

V1= Version 1, V2=Version2, V3=Version3, V4=Version4.

14. Review of UAV Image Segmentation

The growing use of UAVs has given researchers a new opportunity for semantic segmentation of UAV images (Su et al., 2022). Zhang et al. (2022) presented a new approach for semantically segmenting UAV imagery through a hybrid CNN architecture. To enhance segmentation accuracy, the proposed architecture integrates U-Net and ResNet, two prominent CNN architectures. ResNet served as the neural network's fundamental architecture and underwent pretraining on ImageNet to obtain basic features. The ResNet output is subjected to the attention mechanism to prioritize relevant features for segmentation. Yi et al. (2022) utilized a squeeze-and-excitation (SE) block in the attention mechanism to acquire channelwise attention weights. These weights are subsequently applied to the feature maps before transmitting them to the decoder. The decoder employs upsampling and convolutional layers to enhance the spatial resolution of the feature maps and reduce the channel count. The final result is a segmentation map that maintains the spatial resolution of the original image. The proposed method was evaluated using the Potsdam and Vaihingen image datasets. The study's findings demonstrate that the hybrid approach surpasses the conventional FCN and CNN models, exhibiting superior F1 scores and accuracy. The authors, Su et al, Zhang et al, and Yi et al, compared their results with various state-of-the-art models and achieved comparable outcomes. Moreover, Yu, Yang, and Chen (2018) proposed the study of a hybrid CNN that integrates multispectral and RGB data to perform semantic segmentation of UAV imagery. Furthermore, Li et al. (2023) employed a parallel CNN architecture that processed each data type separately and combined the final outputs. This approach outperformed conventional single-modality CNNs, resulting in more precise segmentation outcomes.

A study suggested a CNN architecture that merged a preexisting network with a CNN for feature extraction (Liu et al. 2021). The preexisting network was optimized for semantic segmentation through unmanned aerial vehicle (UAV) imagery. The method proposed by Liu et al. (2021) achieved higher segmentation accuracy and reduced training time compared to conventional CNNs using pretrained weights. Gebrehiwot et al. (2019) employed image segmentation techniques to distinguish flooded water from buildings, vegetation, and roadways in UAV images. Ichim et al. (2020) utilized decision fusion and foliage in UAV imagery to segment flooded areas. Zhang et al. (2020) employed residual U-Net modules to segment plants in UAV images.

15. U-Net for Image Segmentation

U-Net is an advanced fully convolutional network for the precise segmentation of pixel-based images using a small number of training images. Each phase of downsampling doubles the number of feature channels, while each step along the expanding route consists of an upsampling of the feature map, a 2x2 convolution that halves the number of feature channels, a combination with the trimmed feature map from the contracting route, and two 3x3 convolutions. Every convolution results in border pixel loss, necessitating cropping. In the last layer, a 1x1 convolution allocates the appropriate number of classes to each 64-component feature vector. As shown in Figure 17, the network's 23 convolutional layers make it suitable for segmenting annotated Orth mosaic tiles.

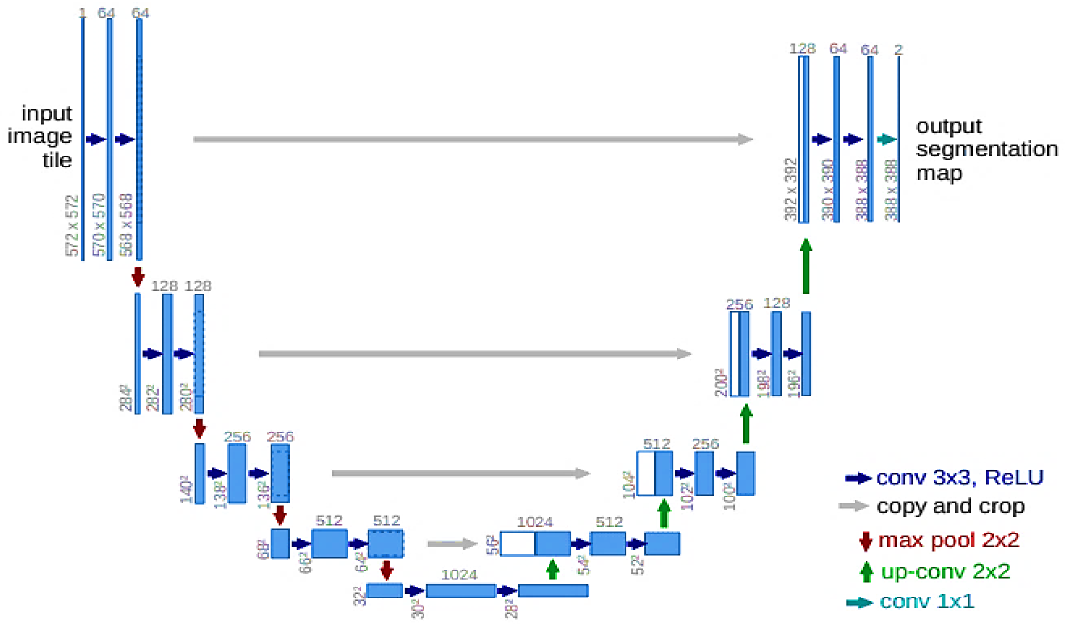


Figure 17 U-Net architecture (Cicek et al., 2016).

16. Modification of U-Net

U-Net became available in 2015 for biomedical segmentation, but it can also be used to address a variety of other issues. It has been transformed into various structures, such as residual U-Net, 3D-U-Net, U-Net++, improved U-Net3+, cascaded U-Net, adversarial U-Net R2U-Net, and attention U-Net. These modifications are based on the architecture of the original U-Net. Some of these modified structures are explained as follows.

16.1. 3D-UNet

3D-UNet comprises a contractive and an expanding route that, using a mix of convolution and pooling processes, tries to create a bottleneck in its central portion. Following this bottleneck, the picture is recreated using convolutions and upsampling. Adding skip connections is intended to facilitate the backward flow of gradients to enhance training.

3D-UNet consists of a route that is both contracting (left) and growing (right). It employs unpadded convolutions followed by maximum pooling for downsampling (Cicek et al., 2016). Every step along the expanding route involves upsampling the feature maps and concatenating them with the proportionally cropped feature map from the contractive path, as shown in Figure 18.

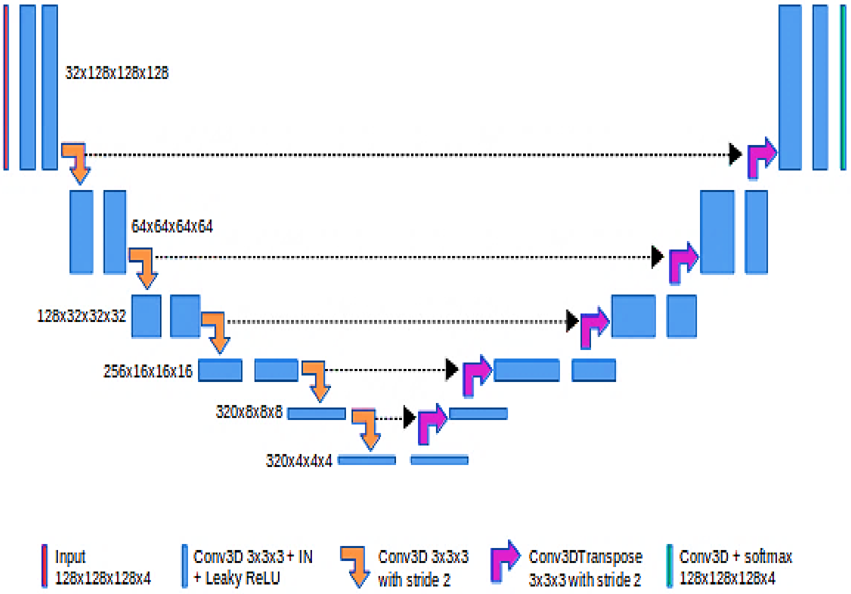


Figure 18 3D-UNet architecture (Cicek et al, 2016).

16.2. Residual U-Net (ResU-Net)

ResU-Net employs residual units as its fundamental building block rather than a simple convolutional block. Residual units include:

- Two convolutional 3x3 bloc
- An identity mapping
- Identity mapping links the residual unit's input and output.
- The convolutional block includes one layer of batch normalization, one layer of ReLU activation, and one convolutional layer, as shown in Figure 19.

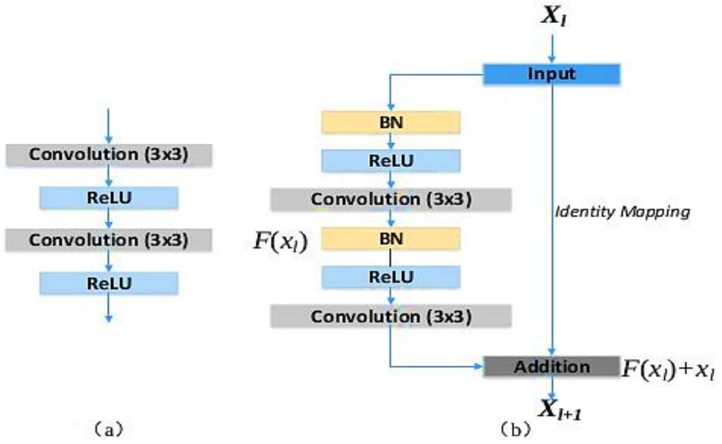


Figure 19 Residual U-Net (Vega Arellano, 2022).

The three components of ResU-Net are encoding, bridge, and decoding. It substitutes pooling in the first convolution block with a stride of 2. Before each decoding unit, lower-level feature maps are upsampled and concatenated with the corresponding encoding route. The segmentation map is generated by a 1x1 convolution with sigmoid activation. As shown in Figure 19.

16.3. U-Net++

U-Net++ comprises an encoder and decoder coupled by a succession of dense convolutional blocks layered within one another. The primary goal of U-Net++ is to bridge the semantic gap between the encoder and decoder feature maps before fusion (Zhou et al., 2018). In U-Net, the decoder receives the encoder's feature maps directly; however, in U-Net++, they are subjected to a dense convolution block whose number of convolution layers is dependent on the pyramid level.

16.4. Adversarial U-Net

Adversarial U-Net is composed of a U-Net for segmentation (generator) and a CNN for feature vectors (discriminator) that encode spatial interactions. In contrast to conventional GANs, the discriminator generates multilevel features as opposed to binary labels. The architecture is depicted schematically in Figure 20 (Sriker et al., 2021).

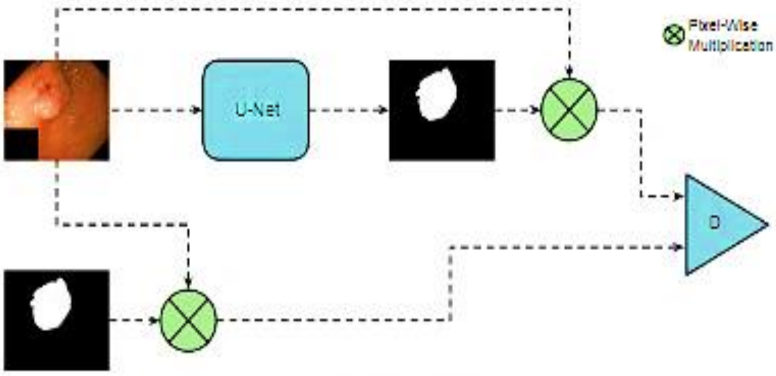


Figure 20 Adversarial U-Net ((Sriker et al., 2021).

The discriminator takes a pixel-by-pixel multiplication of the input image by either the ground truth or the projected segmentation map and outputs a feature vector for each input, while the generator generates a segmentation map. Discriminator blocks are convolutions. Each block's output is flattened and merged to create a hierarchical vector.

16.5. Cascaded U-Net

Cascaded U-Net is an enhanced cascaded network in which the encoder of the consecutive U-Net is placed between the encoder and decoder of the previous U-Net. The inputs that are part of each convolution block of the latter U-Net are obtained from three sources: the previous layer's output, the previous U-Net's output at the same level, and the matching upsampled output of the lower convolution block of the previous U-Net. This structure enhances segmentation performance by enabling the two U-Nets to communicate their learned features, thereby enhancing feature selection and combination, as shown in Figure 21.

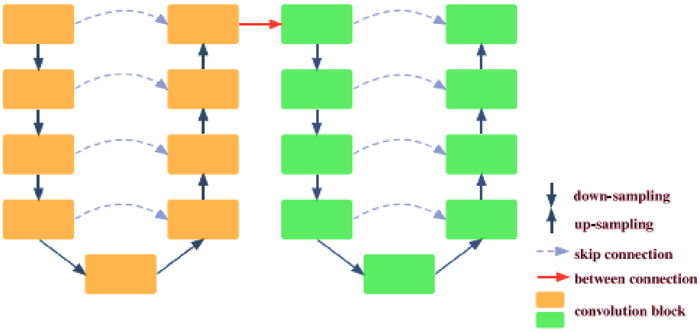


Figure 21 Cascaded U-Net (Jiang et al., 2019).

16.6. Improved U-Net 3+

U-Net3+ employs full-scale skip connections to mix high-level and low-level semantics from multiple scales to increase segmentation accuracy, and its network parameters are less than those of U-Net and U-Net++ (Lefkovits et al.,2022). The encoder component of U-Net3+ is identical to that of U-Net and U-Net++. Each encoder comprises two convolutional layers with a kernel size of 3. These fundamental networks can enhance performance further. The encoder can extract the image's abstract characteristics. The degree of abstraction of characteristics varies between scales. Features with a higher level of abstraction are essential for the subsequent network. Table 3 summarizes the concepts of U-Net variants and the differences between them.

Table 3 Summary of U-Net Variants.

MODIFIED U-Net	Summary
3D UNet (Jun. 2016)	3D-UNet comprises a contractive and an expanding route that using a mix of convolution and pooling processes, tries to create a bottleneck in its central portion. Adding skip connections is intended to facilitate the backward flow of gradients to enhance training (Cicek et al, 2016).
Residual U-Net (Nov.2017)	Instead of using a standard convolutional block, ResUNet makes use of Residual Units as its primary building element.
U-Net++ (Jul 2018)	U-Net++ begins with an encoder subnetwork or backbone, which is then followed by a decoder subnetwork. The redesigned skip paths that connect the two subnetworks and the use of deep supervision distinguish U-Net++ from U-Net. Additionally, it uses dense block concepts to improve U-Net (Zhou et al., 2018).
Adversarial U-Net (May 2019)	The Adversarial U-Net consists of the previously described U-Net to produce segmentation maps (the generator) and a CNN to generate multilevel feature vectors (the discriminator) (Li et al,2019).
Cascaded U-Net (MAY 2020)	A cascaded U-Net is an improved cascaded network in which the encoder of the next U-Net is positioned between the encoder and decoder of the previous U-Net, rather than at the end of the previous U-Net.
Improved U-Net3+ (Jan.2022)	To improve its segmentation accuracy, U-Net3+ uses full-scale skip connections to combine high-level and low-level semantics across various scales. Its network parameters are also less than those of U-Net and U-Net++. U-Net3+ shares an encoder with U-Net and U-Net++(Lefkovits et al.,2022).

17. U-Net for UAV Image Segmentation

Utilizing deep neural network-based strategies in environmental methodologies empowers the execution of diverse tasks, particularly notable in image segmentation tasks involving data such as UAV-acquired images. These operations employ various types of sensors depending on the nature of the task. To map natural habitats and their properties, reviews concentrate on techniques and systems explicitly about the objectives; as such, no globally applicable model could be developed or inferred. Even though deep learning-related strategies have not arrived at this kind of globally applicable model, they are effectively clearing doubts by being effectively executed in the most exceptional situations. However, although UAV-related techniques provide a few limitations to object classification and regression operations, deep neural network strategies are being increasingly perceived as more suitable for performing such undertakings. In any case, there is still much to be explored.

The most widely recognized applications in the context of ecological image segmentation tasks are land cover, land use, and different formats of landscape examination. A new report by Giang et al. (2020) applied semantic-segmentation models to evaluate land use over a mining extraction region. Additionally, Al-Najjar et al. (2019) consolidated data from a digital surface model (DSM) with UAV-related RGB format pictures and implemented a variant of feature combination as input for a convolutional neural network architecture. To plan seaside areas, a methodology by Buscombe et al. (2018), with RGB-format image data enlisted with different scales, utilized a convolutional neural network with a graphical technique referred to as conditional random field (CRF). Furthermore, an additional study by Park et al. (2020), with hyperspectral image data between a mixture of 2-dimensional and 3-dimensional convolution layers, was created to deduce the variation in land cover within the allocated land classification of cadastral-map packages. Owing to advancements in image-capturing systems, it is possible to obtain hyperspectral images through UAVs. An incredible number of spectral bands are contained in these images.

The task of multiclass segmentation poses a significant challenge in unmanned aerial vehicle (UAV) image segmentation, as it necessitates partitioning an image into more than two distinct classes (Ling et al., 2022). In recent years, U-Net has been utilized for the multiclass segmentation of UAV images and has demonstrated promising outcomes. Zhang et al. (2020) proposed a U-Net architecture for multiclass crop segmentation on unmanned aerial vehicle (UAV) images. The study yielded a cumulative precision rate of 94.6% across six distinct crop varieties. The researchers employed data augmentation methodologies, including rotation and magnification, to augment the variety of data employed for training and improve the model's generalizability.

Huang et al. (2020) proposed a modified U-Net model to address the task of multiclass land cover segmentation in UAV images. The research's overall precision for eight distinct land cover categories was 89.9%. The authors enhanced the U-Net architecture by integrating residual connections and spatial pyramid aggregating, increasing the model's efficacy. Zhang et al. (2022) conducted a study wherein a U-Net model was utilized for multiclass crop segmentation in UAV images. The study yielded a combined precision rate of 94.2% for four distinct varieties of crops. The researchers employed a hybrid loss function that integrated the Dice loss and cross-entropy loss to mitigate the class imbalance issue and augment the model's efficacy. Limitations arise when employing the U-Net architecture for semantic segmentation tasks involving over 20 classes.

The U-Net model features a symmetrical encoder-decoder design incorporating skip connections and a limited receptive field. The receptive field of a network is limited as its depth increases. The restricted receptive field can hinder the capture of contextual information and long-range dependencies in complex scenes with numerous classes, leading to imprecise segmentation. Class imbalance is more likely with over 20 classes. Imbalanced training data may occur due to variations in the number of pixels across different classes. The model's capacity to segment minority classes may be compromised due to an imbalance, as the network may prioritize training on majority classes. The computational challenge of the segmentation task increases with the number of classes. The U-Net architecture's parameter counts increase with the number of classes, resulting in higher memory and computational needs. Training a U-Net model with numerous classes can be computationally expensive and necessitate significant hardware resources.

The U-Net architecture utilizes an encoder-decoder structure that involves downsampling in the encoder and upsampling in the decoder, decreasing spatial resolution. The resolution of the segmented output may decrease with an increase in the number of classes due to repeated pooling and upsampling operations. Reduced spatial resolution may lead to fuzzy segmentation and challenges in capturing intricate or small structures. Annotating training data for many classes is a time-consuming and error-prone process. Acquiring a precise and well-labeled dataset for training the U-Net model can be challenging for uncommon or intricate classes.

Nevertheless, there are still research gaps in the literature on applying U-Net for multiclass segmentation in unmanned aerial vehicle (UAV) scenarios. One of the main challenges in the multiclass segmentation of UAV images is the scarcity of annotated datasets (Ling et al., 2022). The creation of annotated datasets can be a resource-intensive and time-consuming process, which can limit the availability of datasets suitable for training and evaluating U-Net models. Furthermore, the generalization of U-Net models to diverse environmental conditions and unmanned aerial vehicle (UAV) platforms poses a challenge (Majidizadeh, Hasani, and Jafari, 2023). The quality and reliability of segmentation results can be affected by the variation in resolution, perspective, and illumination of images obtained through UAVs (Wang, 2022). To surmount these challenges, scholars may explore diverse methodologies to enhance the efficacy and versatility of U-Net for the multiclass segmentation of unmanned aerial vehicle (UAV) images. The techniques employed in this context may encompass transfer learning, amalgamating numerous U-Net models, and data augmentation (Tian, Zhong, and Chen, 2021).

The U-Net architecture is a powerful tool for performing multiclass segmentation of images captured by unmanned aerial vehicles (UAVs). The efficacy of this approach has been demonstrated in various applications, such as the classification of crops and land cover. Nonetheless, research gaps require attention, including the constrained accessibility of annotated datasets and the necessity for U-Net models to achieve generalization (Guo, Zhao, and Wu, 2020).

18. Data collection and available UAV image datasets

In general, the availability of large, real datasets is essential for successful training and testing of the model. Owing to the large number of parameters in U-Net, the result cannot be assured unless the model is trained very well by a large set of samples before deploying this model to the real world. Studies have previously proven that data augmentation is effective in the training of network models, so samples for network training can be added by applying different types of data augmentation techniques based on the real scene.

A review of publicly available UAV image datasets is conducted to obtain an appropriate dataset for the proposed case study, which is UAV image multiclass semantic segmentation. Thus, a brief exploration of the available datasets online is presented along with their advantages, disadvantages, and challenges. See Appendix 2.

19. Summary

In summary, the concepts listed in this chapter cover the up-and-coming subjects in the field of computer vision, and incorporating an aggregation of these topics can provide incredible advancements and assist in the development of modern methods for the application of UAV image processing. With an examination of various domains, the development and enhancement of more modern and efficient methods can be accomplished. Additional research within the field of UAV image processing may find these advancements advantageous, as they can be applied to different tasks and operations.

UAV images have a high resolution. The automation of map creation and the semantic segmentation of UAV images are difficult tasks in semantic segmentation. Due to some challenges in UAV images, the semantic segmentation procedure cannot provide accurate information on UAV images. Therefore, the U-Net architecture technique is proposed to resolve this issue. It is divided into two categories. The compression path (also known as the encoder) is the initial path and is responsible for capturing the image's context. A convolutional and maximum pooling layer stack constitutes the encoder. The second method utilized to provide precise localization via transposed convolutions is the symmetric expanding path (also known as the decoder). This job is frequently referred to as dense prediction, which involves neurons that are entirely interconnected with one another and with the neurons that gave birth to dense layers. Therefore, it is an end-to-end fully convolutional network (FCN), meaning that it has only convolutional layers and no dense layers, allowing it to accept images of any size.

Semantic segmentation has significantly progressed with deep learning algorithms and hyperparameter optimization strategies. This study aims to improve the semantic segmentation performance of UAV images by utilizing convolutional neural networks (CNNs) and selecting the U-Net model as the optimal algorithm for this task. Recognizing and classifying objects is a crucial aspect of autonomous flight. Therefore, by integrating deep learning algorithms such as U-Net into the internal control unit of UAVs, their autonomous piloting capabilities and practical security can be greatly enhanced. The study has yielded valuable findings on the difficulties and possible remedies for improving the semantic segmentation of UAV images and its potential implications for flight autonomy. Figures 22 and 23 below show the role of the proposed U-Net model in improving the process of UAV image semantic segmentation for this project, and it will be presented and discussed thoroughly and profoundly.

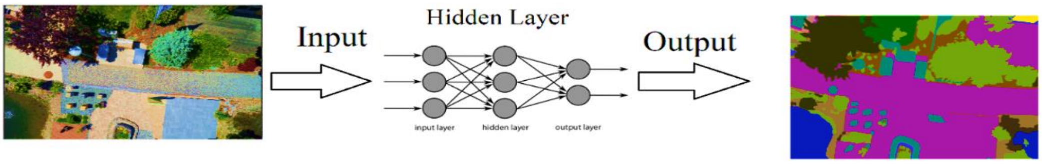


Figure 22 Schematic diagram of UAV image semantic segmentation using a deep learning algorithm based on U-Net architecture.

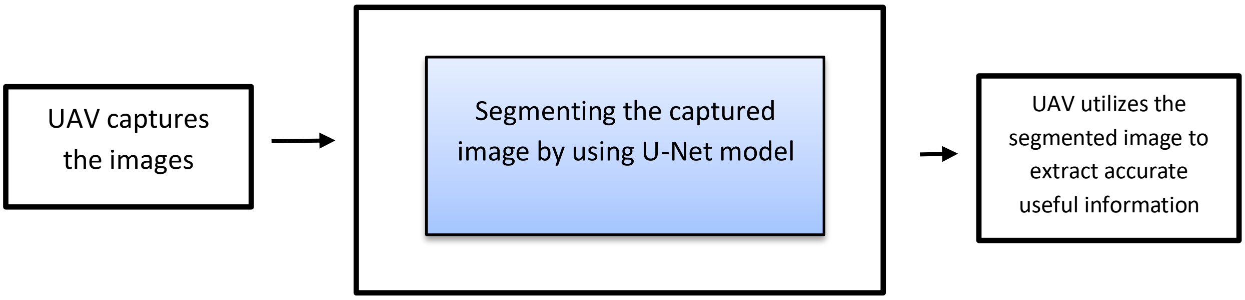

UAV captures the images
Segmenting the captured image by using U-Net model
UAV utilizes the segmented image to extract accurate useful information

Figure 23 Process of UAV image semantic segmentation using the improved U-Net architecture.

Ethical considerations

Not applicable.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support.

References

AbdelMaksoud, E., Barakat, S., & Elmogy, M. (2020). Diabetic retinopathy grading based on a hybrid deep learning model. In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy* (ICDABI) (pp. 1-6). IEEE.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon, 4*(11), e00938.

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., & Sousa, J. J. (2017). Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing*, *9*(11), 1110.

Adayel, R., Bazi, Y., Alhichri, H., & Alajlan, N. (2020). Deep open-set domain adaptation for cross-scene classification based on adversarial learning and pareto ranking. *Remote sensing, 12*(11), 1716.

Al-Najjar, H. A., Kalantar, B., Pradhan, B., Saeidi, V., Halin, A. A., Ueda, N., & Mansor, S. (2019). Land cover classification from fused DSM and UAV images using convolutional neural networks. *Remote Sensing*, *11*(12), 1461.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164.

Ammour, N., Alhichri, H., Bazi, Y., Benjdira, B., Alajlan, N., & Zuair, M. (2017). Deep learning approach for car detection in UAV imagery. *Remote Sensing, 9*(4), 312.

Ampatzidis, Y., & Partel, V. (2019). UAV-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence. *Remote Sensing, 11*(4), 410.

Apolo-Apolo, O. E., Martínez-Guanter, J., Egea, G., Raja, P., & Pérez-Ruiz, M. (2020). Deep learning techniques for estimation of the yield and size of citrus fruits using a UAV. *European Journal of Agronomy, 115*, 126030.

Audebert, N., Le Saux, B., & Lefèvre, S. (2019). Deep learning for classification of hyperspectral data: A comparative review. *IEEE geoscience and remote sensing magazine, 7*(2), 159-173.

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence, 39*(12), 2481-2495.

Bah, M. D., Hafiane, A., & Canals, R. (2018). Deep learning with unsupervised data labeling for weed detection in line crops in UAV images. *Remote sensing, 10*(11), 1690.

Baldi, P., & Vershynin, R. (2019). The capacity of feedforward neural networks. *Neural networks, 116*, 288-311.

Ball, J. E., Anderson, D. T., & Chan Sr, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing, 11*(4), 042609.

Banerjee, K., Gupta, R. R., Vyas, K., & Mishra, B. (2020). Exploring Alternatives to Softmax Function. arXiv preprint arXiv:2011.11538.

Barbedo, J. G. A., Koenigkan, L. V., Santos, P. M., & Ribeiro, A. R. B. (2020). Counting cattle in uav images—dealing with clustered animals and animal/background contrast changes. *Sensors, 20*(7), 2126.

Barbedo, J. G. A., Koenigkan, L. V., Santos, T. T., & Santos, P. M. (2019). A study on the detection of cattle in UAV images using deep learning. *Sensors, 19*(24), 5436.

Bell, T. W., Nidzieko, N. J., Siegel, D. A., Miller, R. J., Cavanaugh, K. C., Nelson, N. B., ... & Griffith, M. (2020). The utility of satellites and autonomous remote sensing platforms for monitoring offshore aquaculture farms: A case study for canopy forming kelps. *Frontiers in Marine Science, 7*, 1083.

Benardos, P. G., & Vosniakos, G. C. (2007). Optimizing feedforward artificial neural network architecture. *Engineering applications of artificial intelligence, 20*(3), 365-382.

Bendale, A., & Boult, T. E. (2016). Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 1563-1572).

Benjdira, B., Bazi, Y., Koubaa, A., & Ouni, K. (2019). Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing, 11*(11), 1369.

Bhatia, Y., Rai, R., Gupta, V., Aggarwal, N., & Akula, A. (2019). Convolutional neural networks based potholes detection using thermal imaging. *Journal of King Saud University-Computer and Information Sciences*.

Bhowmick, S., Nagarajaiah, S., & Veeraraghavan, A. (2020). Vision and deep learning-based algorithms to detect and quantify cracks on concrete surfaces from uav videos. *Sensors, 20*(21), 6299.

Biffi, L. J., Mitishita, E., Liesenberg, V., Santos, A. A. D., Gonçalves, D. N., Estrabis, N. V., ... & Gonçalves, W. N. (2021). ATSS deep learning-based approach to detect apple fruits. *Remote Sensing, 13*(1), 54.

Bithas, P. S., Michailidis, E. T., Nomikos, N., Vouyioukas, D., & Kanatas, A. G. (2019). A survey on machine-learning techniques for UAV-based communications. *Sensors, 19*(23), 5170.

Boonpook, W., Tan, Y., & Xu, B. (2021). Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry. *International Journal of Remote Sensing, 42*(1), 1-19.

Bui, D. T., Tsangaratos, P., Nguyen, V. T., Van Liem, N., & Trinh, P. T. (2020). Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. *Catena, 188*, 104426.

Buscombe, D., & Ritchie, A. C. (2018). Landscape classification with deep neural networks. *Geosciences, 8*(7), 244.

Cai, Z., & Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6154-6162).

Cai, Z., & Vasconcelos, N. (2019). Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, Y., Chen, K., Loy, C. C., & Lin, D. (2020). Prime sample attention in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11583-11591).

Carbonneau, P. E., Dugdale, S. J., Breckon, T. P., Dietrich, J. T., Fonstad, M. A., Miyamoto, H., & Woodget, A. S. (2020). Adopting deep learning methods for airborne RGB fluvial scene classification. *Remote Sensing of Environment, 251*, 112107.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European Conference on Computer Vision,* 213-229. Springer, Cham.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV),* 132-149.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882.

Castro, W., Marcato Junior, J., Polidoro, C., Osco, L. P., Gonçalves, W., Rodrigues, L., ... & Matsubara, E. (2020). Deep learning applied to phenotyping of biomass in forages with UAV-based RGB imagery. *Sensors, 20*(17), 4802.

Chaurasia, A., & Culurciello, E. (2017, December). Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*,1-4. IEEE.

Chen, J., Wu, Q., Liu, D., & Xu, T. (2020, August). Foreground-background imbalance problem in deep object detectors: A review. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*,285-290. IEEE.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., ... & Lin, D. (2019). Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974-4983.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence, 40*(4), 834-848.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, Z., Li, D., Fan, W., Guan, H., Wang, C., & Li, J. (2021). Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images. *Remote Sensing, 13*(13), 2524.

Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE, 105*(10), 1865-1883.

Cheng, G. & Junwei H. (2027). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing, 117*, 11-28.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, 424-432. Springer, Cham.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. arXiv preprint arXiv:2009.09796.

da Silva, C. C., Nogueira, K., Oliveira, H. N., & dos Santos, J. A. (2020, March). Towards open-set semantic segmentation of aerial images. In *2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS)*, 16-21. IEEE.

De Oliveira, D. C., & Wehrmeister, M. A. (2018). Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV. *Sensors, 18*(7), 2244.

Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetr

Di Franco, G., & Santurro, M. (2021). Machine learning, artificial neural networks and social research. *Quality & quantity, 55*(3), 1007-1025.

Ding, L., Tang, H., & Bruzzone, L. (2020). Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 59*(1), 426-435.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., ... & Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 370-386.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6569-6578.

Elshamli, A., Taylor, G. W., Berg, A., & Areibi, S. (2017). Domain adaptation using representation learning for the classification of remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(9), 4198-4209.

Fang, B., Kou, R., Pan, L., & Chen, P. (2019). Category-sensitive domain adaptation for land cover mapping in aerial scenes. *Remote Sensing, 11*(22), 2631.
Fang, W., Yu, Z., Chen, Y., Huang, T., Masquelier, T., & Tian, Y. (2021).

Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems, 34*, 21056-21069. Feng, Q., Yang, J., Liu, Y., Ou, C., Zhu, D., Niu, B., ... & Li, B. (2020). Multi-temporal unmanned aerial vehicle remote sensing for vegetable mapping using an attention-based recurrent convolutional neural network. *Remote Sensing, 12*(10), 1668.

Ferreira, M. P., de Almeida, D. R. A., de Almeida Papa, D., Minervino, J. B. S., Veras, H. F. P., Formighieri, A., ... & Ferreira, E. J. L. (2020). Individual tree detection and species classification of Amazonian palms using UAV images and deep learning. *Forest Ecology and Management, 475*, 118397.

Gao, S., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. H. (2019). Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gebrehiwot, A., Hashemi-Beni, L., Thompson, G., Kordjamshidi, P., & Langan, T. E. (2019). Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data. *Sensors, 19*(7), 1486.

Gevaert, C. M., Persello, C., Nex, F., & Vosselman, G. (2018). A deep learning approach to DTM extraction from imagery using rule-based training labels. *ISPRS Journal of Photogrammetry and Remote Sensing, 142*, 106-123.

Gevaert, C. M., Persello, C., Sliuzas, R., & Vosselman, G. (2020). Monitoring household upgrading in unplanned settlements with unmanned aerial vehicles. *International Journal of Applied Earth Observation and Geoinformation, 90*, 102117.

Ghiasi, G., Lin, T. Y., & Le, Q. V. (2019). Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 7036-7045.

Giang, T. L., Dang, K. B., Le, Q. T., Nguyen, V. G., Tong, S. S., & Pham, V. M. (2020). U-Net convolutional networks for mining land cover classification based on high-resolution UAV imagery. *IEEE Access, 8*, 186257-186273.

Gomes, M., Silva, J., Gonçalves, D., Zamboni, P., Perez, J., Batista, E., ... & Gonçalves, W. (2020). Mapping utility poles in aerial ortho Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*.

Gray, P. C., Bierlich, K. C., Mantell, S. A., Friedlaender, A. S., Goldbogen, J. A., & Johnston, D. W. (2019). Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry. *Methods in Ecology and Evolution, 10*(9), 1490-1500.

Guo, T., Dong, J., Li, H., & Gao, Y. (2017, March). Simple convolutional neural network on image classification. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, 721-724. IEEE.

Guo, Y. (2018). A survey on methods and theories of quantized neural networks. arXiv preprint arXiv*:1808.04752*. Hamdi, Z. M., Brandmeier, M., & Straub, C. (2019). Forest damage assessment using deep learning on high resolution remote sensing data. *Remote Sensing, 11*(17), 1976

Hamdi, Z. M., Brandmeier, M., & Straub, C. (2019). *Forest damage assessment using deep learning on high resolution remote sensing data. Remote Sensing, 11*(17), 1976.

Hamylton, S. M., Morris, R. H., Carvalho, R. C., Roder, N., Barlow, P., Mills, K., & Wang, L. (2020). Evaluating techniques for mapping island vegetation from unmanned aerial vehicle (UAV) images: Pixel classification, visual interpretation and machine learning approaches. *International Journal of Applied Earth Observation and Geoinformation, 89*, 102085.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , 9729-9738.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778).

Hennessy, A., Clarke, K., & Lewis, M. (2020). Hyperspectral classification of plants: a review of waveband selection generalisability. *Remote Sensing, 12*(1), 113.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8).

Horning, N., Fleishman, E., Ersts, P. J., Fogarty, F. A., & Wohlfeil Zillig, M. (2020). Mapping of land cover with open-source software and ultra-high-resolution imagery acquired with unmanned aerial vehicles. *Remote Sensing in Ecology and Conservation, 6*(4), 487-497.

Hossain, M. D., & Chen, D. (2019). Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. *ISPRS Journal of Photogrammetry and Remote Sensing, 150*, 115-134.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process, 5*(2), 1.

Hou, J., He, Y., Yang, H., Connor, T., Gao, J., Wang, Y., ... & Zhou, S. (2020). Identification of animal individuals using deep learning: A case study of giant panda. *Biological Conservation, 242*, 108414.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hu, G., Yin, C., Wan, M., Zhang, Y., & Fang, Y. (2020). Recognition of diseased Pinus trees in UAV images using deep learning and AdaBoost classifier. *Biosystems Engineering, 194*, 138-151.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132-7141).

Hua, Y., Marcos, D., Mou, L., Zhu, X. X., & Tuia, D. (2021). Semantic segmentation of remote sensing images with sparse annotations. *IEEE Geoscience and Remote Sensing Letters*.

Huang, G., Chen, D., Li, T., Wu, F., Van Der Maaten, L., & Weinberger, K. Q. (2017). Multi-scale dense convolutional networks for efficient prediction. arXiv preprint arXiv:1703.09844, 2(2).

Ichim, L., & Popescu, D. (2020). Segmentation of vegetation and flood from aerial images based on decision fusion of neural networks. *Remote Sensing, 12*(15), 2490.

Ienco, D., Gaetano, R., Dupaquier, C., & Maurel, P. (2017). Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters, 14*(10), 1685-1689.

Imran, H. A., Mujahid, U., Wazir, S., Latif, U., & Mehmood, K. (2020). Embedded development boards for edge-ai: A comprehensive report. arXiv preprint arXiv:2009.00803.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1125-1134).

Jakovljevic, G., Govedarica, M., Alvarez-Taboada, F., & Pajic, V. (2019). Accuracy assessment of deep learning based classification of lidar and uav points clouds for dtm creation and flood risk mapping. *Geosciences, 9*(7), 323.

Jia, S., Jiang, S., Lin, Z., Li, N., Xu, M., & Yu, S. (2021). A survey: Deep learning for hyperspectral image classification with few labeled samples. *Neurocomputing, 448*, 179-204.

Jiang, Z., Ding, C., Liu, M., & Tao, D. (2019, October). Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *International MICCAI Brainlesion Workshop* (pp. 231-241). Springer, Cham.

Kang, H., & Chen, C. (2020). Fast implementation of real-time fruit detection in apple orchards using deep learning. *Computers and Electronics in Agriculture, 168*, 105108.

Kang, J., Fernandez-Beltran, R., Duan, P., Liu, S., & Plaza, A. J. (2020). Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Transactions on Geoscience and Remote Sensing, 59*(3), 2598-2610.

Kannojia, S. P., & Jaiswal, G. (2018). Effects of varying resolution on performance of CNN based image classification: An experimental study. *Int. J. Comput. Sci. Eng, 6*(9), 451-456.

Karami, A., Crawford, M., & Delp, E. J. (2020). Automatic plant counting and location based on a few-shot learning technique. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13*, 5872-5886.

Kellenberger, B., Marcos, D., & Tuia, D. (2018). Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote sensing of environment, 216*, 139-153.

Kerkech, M., Hafiane, A., & Canals, R. (2020). Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach. *Computers and Electronics in Agriculture, 174*, 105446.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2019). A survey of the recent architectures of deep convolutional neural networks. *arXiv. Preprint, 10*.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review, 53*(8), 5455-5516.

Khelifi, L., & Mignotte, M. (2020). Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. *IEEE Access, 8*, 126385-126400.

Kim, K., & Lee, H. S. (2020). Probabilistic anchor assignment with iou prediction for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 355-371. Springer International Publishing.

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9404-9413).

Kirillov, A., Wu, Y., He, K., & Girshick, R. (2020). Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9799-9808.

Kitano, B. T., Mendes, C. C., Geus, A. R., Oliveira, H. C., & Souza, J. R. (2019). Corn plant counting using deep learning and UAV images. *IEEE Geoscience and Remote Sensing Letters*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems, 25*, 1097-1105.

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters, 14*(5), 778-782.

Larsen, A., Hanigan, I., Reich, B. J., Qin, Y., Cope, M., Morgan, G., & Rappold, A. G. (2021). A deep learning approach to identify smoke plumes in satellite imagery in near-real time for health risk communication. *Journal of exposure science & environmental epidemiology, 31*(1), 170-176.

Lathuilière, S., Mesejo, P., Alameda-Pineda, X., & Horaud, R. (2019). A comprehensive analysis of deep regression. *IEEE transactions on pattern analysis and machine intelligence, 42*(9), 2065-2081.

Law, H., & Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, 734-750.

Lee, J., Kang, D., & Ha, S. (2020). S3NAS: Fast NPU-aware neural architecture search methodology. arXiv preprint arXiv*:2009.02009*.

Lefkovits, S., Emerich, S., & Lefkovits, L. (2022). Boosting Unsupervised Dorsal Hand Vein Segmentation with U-Net Variants. *Mathematics, 10*(15), 2620.

Li, C., Xu, C., Cui, Z., Wang, D., Zhang, T., & Yang, J. (2019, September). Feature-attentioned object detection in remote sensing imagery. In *2019 IEEE International Conference on Image Processing (ICIP)*, 3886-3890. IEEE.

Li, L., Han, J., Yao, X., Cheng, G., & Guo, L. (2020). DLA-MatchNet for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., & Benediktsson, J. A. (2019). Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing, 57*(9), 6690-6709.

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., ... & Yang, J. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. arXiv preprint arXiv*:2006.04388*.

Li, Y., Cao, Z., Lu, H., & Xu, W. (2020). Unsupervised domain adaptation for in-field cotton boll status identification. *Computers and Electronics in Agriculture, 178*, 105745.

Li, Y., Chen, Y., Wang, N., & Zhang, Z. (2019). Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6054-6063.

Li, Y., Huang, Q., Pei, X., Jiao, L., & Shang, R. (2020). RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sensing, 12*(3), 389.

Li, Y., Peng, B., He, L., Fan, K., Li, Z., & Tong, L. (2019). Road extraction from unmanned aerial vehicle remote sensing images based on improved neural networks. Sensors, 19(19), 4115.

Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(6), e1264.

Licciardi, G., Marpu, P. R., Chanussot, J., & Benediktsson, J. A. (2011). Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles. *IEEE Geoscience and Remote Sensing Letters, 9*(3), 447-451.

Lin, D., Fu, K., Wang, Y., Xu, G., & Sun, X. (2017). MARTA GANs: Unsupervised representation learning for remote sensing image classification. *IEEE Geoscience and Remote Sensing Letters, 14*(11), 2092-2096.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117-2125.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759-8768.

Liu, W., & Qin, R. (2020). A multikernel domain adaptation method for unsupervised transfer learning on cross-source and cross-region remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing, 58*(6), 4279-4289.

Lobo Torres, D., Queiroz Feitosa, R., Nigri Happ, P., Elena Cue La Rosa, L., Marcato Junior, J., Martins, J., ... & Liesenberg, V. (2020). Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution UAV optical imagery. *Sensors, 20*(2), 563.

Lu, X., Li, B., Yue, Y., Li, Q., & Yan, J. (2019). Grid r-cnn plus: Faster and better. arXiv preprint arXiv*:1906.05688*.

Ma, J., Jiang, X., Fan, A., Jiang, J., & Yan, J. (2021). Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision, 129*(1), 23-79.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing, 152*, 166-177.

Makantasis, K., Karantzalos, K., Doulamis, A., & Doulamis, N. (2015, July). Deep supervised learning for hyperspectral data classification through convolutional neural networks. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 4959-4962.

Micheli-Tzanakou, E. (2011). Artificial neural networks: an overview. *Network: Computation in Neural Systems, 22*(1-4), 208-230.

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

Minh, D. H. T., Ienco, D., Gaetano, R., Lalande, N., Ndikumana, E., Osman, F., & Maurel, P. (2018). Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR Sentinel-1. *IEEE Geoscience and Remote Sensing Letters, 15*(3), 464-468.

Mittal, S. (2019). A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. *Journal of Systems Architecture, 97*, 428-442.

Miyoshi, G. T., Arruda, M. D. S., Osco, L. P., Marcato Junior, J., Gonçalves, D. N., Imai, N. N., ... & Gonçalves, W. N. (2020). A novel deep learning method to identify single tree species in UAV-based hyperspectral images. *Remote Sensing, 12*(8), 1294.

Nanni, L., Lumini, A., Ghidoni, S., & Maguolo, G. (2020). Stochastic selection of activation layers for convolutional neural networks. *Sensors, 20*(6), 1626.

Nevavuori, P., Narra, N., Linna, P., & Lipping, T. (2020). Crop yield prediction using multitemporal uav data and spatio-temporal deep learning models. *Remote Sensing, 12*(23), 4000.

Nezami, S., Khoramshahi, E., Nevalainen, O., Pölönen, I., & Honkavaara, E. (2020). Tree species classification of drone hyperspectral and rgb imagery with deep learning convolutional neural networks. *Remote Sensing, 12*(7), 1070.

Nogueira, K., Dalla Mura, M., Chanussot, J., Schwartz, W. R., & Dos Santos, J. A. (2019). Dynamic multicontext segmentation of remote sensing images based on convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing, 57*(10), 7503-7520.

Neupane, B., Horanont, T., & Aryal, J. (2021). Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sensing, 13*(4), 808.

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv*:1811.03378*.

Osco, L. P., de Arruda, M. D. S., Gonçalves, D. N., Dias, A., Batistoti, J., de Souza, M., ... & Gonçalves, W. N. (2021). A CNN approach to simultaneously count plants and detect plantation-rows from UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing, 174*, 1-17.

Osco, L. P., de Arruda, M. D. S., Junior, J. M., da Silva, N. B., Ramos, A. P. M., Moryia, É. A. S., ... & Gonçalves, W. N. (2020). A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing, 160*, 97-106.

Osco, L. P., Nogueira, K., Ramos, A. P. M., Pinheiro, M. M. F., Furuya, D. E. G., Gonçalves, W. N., ... & dos Santos, J. A. (2021). Semantic segmentation of citrus-orchard using deep neural networks and multispectral UAV-based imagery. *Precision Agriculture*, 1-18.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv*:1511.08458*.

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D. (2019). Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 821-830.

Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2019). Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal of Photogrammetry and Remote Sensing, 158*, 279-317.

Park, S., & Song, A. (2020). Discrepancy analysis for detecting candidate parcels requiring update of land category in cadastral map using hyperspectral UAV Images: A case study in Jeonju, South Korea. *Remote Sensing, 12*(3), 354.

Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2013). How to construct deep recurrent neural networks. arXiv preprint arXiv*:1312.6026*.

Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv*:1606.02147*.

Penatti, O. A., Nogueira, K., & Dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 44-51.

Peng, C., Zhang, X., Yu, G., Luo, G., & Sun, J. (2017). Large kernel matters--improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4353-4361.

Petersson, H., Gustafsson, D., & Bergstrom, D. (2016, December). Hyperspectral image analysis using deep learning—A review. In *2016 sixth international conference on image processing theory, tools and applications (IPTA),* pp. 1-6.

Prochaska, J. X., Cornillon, P. C., & Reiman, D. M. (2021). Deep Learning of Sea Surface Temperature Patterns to Identify Ocean Extremes. *Remote Sensing, 13*(4), 744.

Qiao, S., Chen, L. C., & Yuille, A. (2021). Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10213-10224.

Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y., & Sun, J. (2019). ThunderNet: Towards real-time generic object detection on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6718-6727.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10428-10436.

Ramprasath, M., Anand, M. V., & Hariharan, S. (2018). Image classification using convolutional neural networks. *International Journal of Pure and Applied Mathematics, 119*(17), 1307-1319.

Rivas, A., Chamoso, P., González-Briones, A., & Corchado, J. M. (2018). Detection of cattle using drones and convolutional neural networks. *Sensors, 18*(7), 2048.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234-241.

Sakib, S., Ahmed, N., Kabir, A. J., & Ahmed, H. (2019). An overview of convolutional neural network: its architecture and applications.

Santos, A. A. D., Marcato Junior, J., Araújo, M. S., Di Martini, D. R., Tetila, E. C., Siqueira, H. L., ... & Gonçalves, W. N. (2019). Assessment of CNN-based methods for individual tree detection on images captured by RGB cameras attached to UAVs. *Sensors, 19*(16), 3595.

Sarkar, D., Khan, T., & Laskar, R. H. (2020). Multi-parametric ANN modelling for interference rejection in UWB antennas. *International Journal of Electronics, 107*(12), 2068-2083.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Shanmugamani, R. (2018). Deep Learning for Computer Vision: Expert techniques to train advanced neural networks using TensorFlow and Keras. Packt Publishing Ltd.

Sharma, V., & Mir, R. N. (2020). A comprehensive and systematic look up into deep learning based object detection techniques: A review. *Computer Science Review, 38*, 100301.

Sheng, G., Yang, W., Xu, T., & Sun, H. (2012). High-resolution satellite scene classification using a sparse coding based multiple feature combination. *International journal of remote sensing, 33*(8), 2395-2412.

Signoroni, A., Savardi, M., Baronio, A., & Benini, S. (2019). Deep learning meets hyperspectral image analysis: A multidisciplinary review. *Journal of Imaging, 5*(5), 52.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv*:1409.1556*.

Soderholm, J. S., Kumjian, M. R., McCarthy, N., Maldonado, P., & Wang, M. (2020). Quantifying hail size distributions from the sky–application of drone aerial photogrammetry. *Atmospheric Measurement Techniques, 13*(2), 747-754.

Soudy, M., Afify, Y. M., & Badr, N. (2022). GenericConv: A Generic Model for Image Scene Classification Using Few-Shot Learning. *Information, 13*(7), 315.

Souza, R., Bento, M., Nogovitsyn, N., Chung, K. J., Loos, W., Lebel, R. M., & Frayne, R. (2020). Dual-domain cascade of U-nets for multi-channel magnetic resonance image reconstruction. *Magnetic resonance imaging, 71*, 140-153.

Sriker, D., Cohen, D., Cahan, N., & Greenspan, H. (2021, February). Improved segmentation by adversarial U-Net. In *Medical Imaging 2021: Computer-Aided Diagnosis 11597*, 290-295.

Su, Y., Wu, Y., Wang, M., Wang, F., & Cheng, J. (2019, July). Semantic segmentation of high resolution remote sensing image based on batch-attention mechanism. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 3856-3859.

Sultana, F., Sufian, A., & Dutta, P. (2018, November). Advancements in image classification using convolutional neural network. In *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 122-129.

Sundaram, D. M., & Loganathan, A. (2020). FSSCaps-DetCountNet: fuzzy soft sets and CapsNet-based detection and counting network for monitoring animals from aerial images. *Journal of Applied Remote Sensing, 14*(2), 026521.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018, October). A survey on deep transfer learning. In *International conference on artificial neural networks*, 270-279. Springer, Cham.

Tetila, E. C., Machado, B. B., Menezes, G. K., Oliveira, A. D. S., Alvarez, M., Amorim, W. P., ... & Pistori, H. (2019). Automatic recognition of soybean leaf diseases using UAV images and deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters, 17*(5), 903-907.

Thoma, M. (2016). A survey of semantic segmentation. arXiv preprint arXiv*:1602.06541*.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347-10357.

Traore, B. B., Kamsu-Foguem, B., & Tangara, F. (2018). Deep convolution neural network for image recognition. *Ecological Informatics, 48*, 257-268.

Tsagkatakis, G., Aidini, A., Fotiadou, K., Giannopoulos, M., Pentari, A., & Tsakalides, P. (2019). Survey of deep-learning approaches for remote sensing observation enhancement. *Sensors, 19*(18), 3929.

Tuia, D., Persello, C., & Bruzzone, L. (2021). Recent advances in domain adaptation for the classification of remote sensing data. arXiv preprint arXiv*:2104.07778*.

Vaddi, R., & Manoharan, P. (2020). CNN based hyperspectral image classification using unsupervised band selection and structure-preserving spatial features. *Infrared Physics & Technology, 110*, 103457.

Van Dao, D., Jaafari, A., Bayat, M., Mafi-Gholami, D., Qi, C., Moayedi, H., ... & Pham, B. T. (2020). A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *Catena, 188*, 104451.

Vega Arellano, J. P. (2022). Stroke Detection with Deep Learning (Doctoral dissertation, SRH Hochschule Heidelberg).

Wang, J., Chen, K., Yang, S., Loy, C. C., & Lin, D. (2019). Region proposal by guided anchoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2965-2974.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Wang, J., Yu, L. C., Lai, K. R., & Zhang, X. (2016, August). Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 225-230.

Wang, J., Zhang, W., Cao, Y., Chen, K., Pang, J., Gong, T., ... & Lin, D. (2020, August). Side-aware boundary localization for more precise object detection. In *European Conference on Computer Vision*, 403-419.

Wang, J., Zhang, X., Lv, P., Zhou, L., & Wang, H. (2021). EAR-U-Net: EfficientNet and attention-based residual U-Net for automatic liver segmentation in CT. arXiv preprint arXiv*:2110.01014*.

Wang, S., Zhou, J., Lei, T., Wu, H., Zhang, X., Ma, J., & Zhong, H. (2020). Estimating Land Surface Temperature from Satellite Passive Microwave Observations with the Traditional Neural Network, Deep Belief Network, and Convolutional Neural Network. *Remote Sensing, 12*(17), 2691.

Wang, Y., Ding, W., Zhang, R., & Li, H. (2020). Boundary-aware multitask learning for remote sensing imagery. *IEEE Journal of selected topics in applied earth observations and remote sensing, 14*, 951-963.

Wu, T., Tang, S., Zhang, R., Cao, J., & Zhang, Y. (2020). Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing, 30*, 1169-1179.

Wu, X., Sahoo, D., & Hoi, S. C. (2020). Recent advances in deep learning for object detection. *Neurocomputing, 396*, 39-64.

Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., ... & Zhang, L. (2018). DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974-3983.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,1492-1500.

Xu, R., Tao, Y., Lu, Z., & Zhong, Y. (2018). Attention-mechanism-containing neural networks for high-resolution remote sensing image classification. *Remote Sensing, 10*(10), 1602.

Yang, M. D., Tseng, H. H., Hsu, Y. C., & Tsai, H. P. (2020). Semantic segmentation using deep learning with vegetation indices for rice lodging identification in multi-date UAV visible images. *Remote Sensing, 12*(4), 633.

Yao, C., Luo, X., Zhao, Y., Zeng, W., & Chen, X. (2017, December). A review on image classification of remote sensing using deep learning. In *2017 3rd IEEE International Conference on Computer and Communications* (ICCC), 1947-1955.

Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., & Hu, H. (2020, August). Disentangled non-local neural networks. In *European Conference on Computer Vision* (pp. 191-207). Springer, Cham.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., ... & Zhang, L. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment, 241*, 111716.

Yuan, X., Shi, J., & Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications, 169*, 114417.

Zhang, C., Atkinson, P. M., George, C., Wen, Z., Diazgranados, M., & Gerard, F. Identifying and mapping individual plants in a highly diverse high-elevation ecosyst

Zhang, G., Wang, M., & Liu, K. (2019). Forest fire susceptibility modeling using a convolutional neural network for Yunnan province of China. *International Journal of Disaster Risk Science, 10*(3), 386-403.

Zhang, H., Chang, H., Ma, B., Wang, N., & Chen, X. (2020, August). Dynamic R-CNN: Towards high quality object detection via dynamic training. In *European Conference on Computer Vision* , 260-275.

Zhang, H., Liptrott, M., Bessis, N., & Cheng, J. (2019, September). Real-time traffic analysis using deep learning techniques and UAV based video. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance* (AVSS), 1-5.

Zhang, H., Wang, Y., Dayoub, F., & Sunderhauf, N. (2021). Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8514-8523.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., ... & Smola, A. (2020). Resnest: Split-attention networks. arXiv preprint arXiv*:2004.08955*.

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine, 4*(2), 22-40.

Zhang, S., Chi, C., Yao, Y., Lei, Z., & Li, S. Z. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9759-9768.

Zhang, X., Han, L., Han, L., & Zhu, L. (2020). How well do deep learning-based methods for land cover classification and object detection perform on high resolution remote sensing imagery?. *Remote Sensing, 12*(3), 417.

Zhang, X., Jin, J., Lan, Z., Li, C., Fan, M., Wang, Y., ... & Zhang, Y. (2020). ICENET: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features. *Remote Sensing, 12*(2), 221.

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6848-6856.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,2881-2890.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems, 30*(11), 3212-3232.

Zheng, Z., Lei, L., Sun, H., & Kuang, G. (2020, July). A Review of Remote Sensing Image Object Detection Algorithms Based on Deep Learning. In *2020 IEEE 5th International Conference on Image, Vision and Computing* (ICIVC), 34-43.

Zhou, D., Wang, G., He, G., Long, T., Yin, R., Zhang, Z., ... & Luo, B. (2020). Robust Building Extraction for High Spatial Resolution Remote Sensing Images with Self-Attention Network. *Sensors, 20*(24), 7241.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support,* 3-11.

Zhu, C., He, Y., & Savvides, M. (2019). Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 840-849.

Zhu, R., Yan, L., Mo, N., & Liu, Y. (2019). Attention-based deep feature fusion for the scene classification of high-resolution remote sensing images. *Remote Sensing, 11*(17), 1996.

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 8-36.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43-76.

Zou, Q., Ni, L., Zhang, T., & Wang, Q. (2015). Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters, 12*(11), 2321-2325.