Hindawi
Security and Communication Networks
Volume 2022, Article ID 6010912, 31 pages
https://doi.org/10.1155/2022/6010912

Review Article

Research Contribution and Comprehensive Review towards the Semantic Segmentation of Aerial Images Using Deep Learning Techniques

P. Anilkumar and P. Venugopal

School of Electronics Engineering, Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India

Correspondence should be addressed to P. Venugopal; venugopal.p@vit.ac.in

Received 22 December 2021; Revised 19 January 2022; Accepted 31 January 2022; Published 20 March 2022

Academic Editor: Mamoun Alazab

Semantic segmentation is a significant research topic for decades and has been employed in several applications. In recent years, semantic segmentation has been focused on different deep learning approaches in the area of computer vision, which has aimed for getting superior efficiency while analyzing the aerial and remote-sensing images. The main aim of this review is to provide a clear algorithmic categorization and analysis of the diverse contribution of semantic segmentation of aerial images and expects to give the comprehensive details associated with the recent developments. In addition, the emerged deep learning methods demonstrated much improved performance measures on several public datasets and incredible efforts have been dedicated to advancing pixel-level accuracy. Hence, the analysis on diverse datasets of each contribution is studied, and also, the best performance measures achieved by the existing semantic segmentation models are evaluated. Thus, this survey can facilitate researchers in understanding the development of semantic segmentation in a shorter time, simplify understanding of its latest advancements, research gaps, and challenges to be used as a reference for developing the new semantic image segmentation models in the future.

1. Introduction

Semantic segmentation is an image analysis task, which assigns a label for each pixel in input images for describing the class of its encircled region [1]. Semantic segmentation of aerial images represents the assignment of one land cover category to each pixel, which is a complex task owing to the huge variations in the appearances of ground objects. Several works have been presented in recent years [2]. The state-of-the-art approaches in semantic segmentation are focused on the hand-crafted features, which fail to get the satisfactory performances and are restricted through the depiction ability of features [3]. When compared with object detection and image classification, semantic segmentation is used as the highest level of the image analysis process, which permits complete scene information of the complete input image [4]. In several remote-sensing tasks, semantic segmentation is considered as pixel-wise classification [5]. Semantic segmentation of aerial imagery has been employed in diverse applications such as hazard identification and avoidance, traffic management and evaluation, and urban area planning and monitoring [6]. However, the growth of semantic segmentation techniques was stopped years ago due to the lower accuracy rate of existing image analysis methods focused on the extraction of hand-crafted features [7].

Aerial and satellite imagery have been utilized in different applications such as regional planning, cartography, landscaping, and agriculture [8]. In 2020, Maddikunta et al. [9] have focused on applications, requirements, and challenges of UAV images which were captured from UAV vehicles for smart agriculture system. Multirotor UAVs are usually used for airborne surveillance, photography, and other similar tasks. These are the simplest to produce and the least expensive of all types of UAVs. These images have

different visible colors and other spectra. There is also elevation imagery, which is generally prepared through light detection and ranging (LiDAR) and radar images [10]. Moreover, along with the emergence of satellite and aerial images, remote sensing is also implemented. Remote-sensing images are gathered from the remote object through a device, which cannot be physically contacted the object [11]. In recent years, the data analysis and interpretation are still performed by human experts. Although, semantic segmentation offers superior abilities in object detection, it suffers from implementing it into the real use cases [12]. In 2020, Ch et al. [13] have suggested the security and privacy of UAV data using blockchain technology. The value of virtual circuit (VC)-based devices—UAVs, drones, and similar other IoT-based devices—has grown tremendously in recent years. These gadgets are mostly utilized for aerial surveying in sensitive and isolated locations. The object detection in aerial images is complex due to the bird's-eye view of aerial images, which have huge variations in orientation, high nonuniform object densities, large aspect ratios, and scale variations of objects. Moreover, several challenges are presented in the detection of objects using aerial images, which are low GPU memory capacity, downsampling a large image, and lack of inference on large images [14]. In aerial images, several sensor and resolution are considered as the factors for producing the dataset biases [15]. The standard dataset is prepared by collecting the images from different platforms and sensors through several resolutions including aerial images, satellite images, Gaofen-2 (GF-2) Satellite, and Google Earth [16].

Currently, many DL applications are being used all over the world. Healthcare, social network analysis, audio and speech processing (such as recognition and enhancement), visual data processing methods (such as multimedia data analysis and computer vision), and NLP (translation and sentence classification) are examples of these applications. These applications are divided into five groups: classification, localization, detection, segmentation, and registration. Although each of these jobs has its own aim, as seen in Figure 1, there is significant overlap in the pipeline implementation of these applications.

The semantic segmentation is adopted by deep learning approaches in recent years, which has attained high efficiency in diverse conventional computer vision applications and consists of detection and classification of objects and semantic segmentation [15]. These approaches have automatically derived features, which are customized for classification tasks that create these approaches to offer suitable options for managing complex cases [17]. The huge achievement in other fields makes the extension and adoption of deep learning approaches for solving the challenges in remote-sensing fields. Although, deep learning offers noteworthy performance, it suffers from allocating significant labels to the components of remote-sensing image [18]. Due to the large number and enormous quantity of modalities of the remote-sensing data, the deep neural network has been facilitated for feature extraction [19]. It has also offered great benefits to practitioners and researchers, which require less programming intensive tools for high-level data analysis and are understandable in geosciences [20]. In 2021, Kumar et al. [21] have given a secured privacy preserving framework for smart agriculture unmanned aerial vehicles for both blockchain and nonblockchain frameworks. Balamurugan et al. [22] have given a direction-of-arrival (DOA) tracking for seamless connectivity in beamformed IoT-based drones, and their communication and beamformed performances were increased.

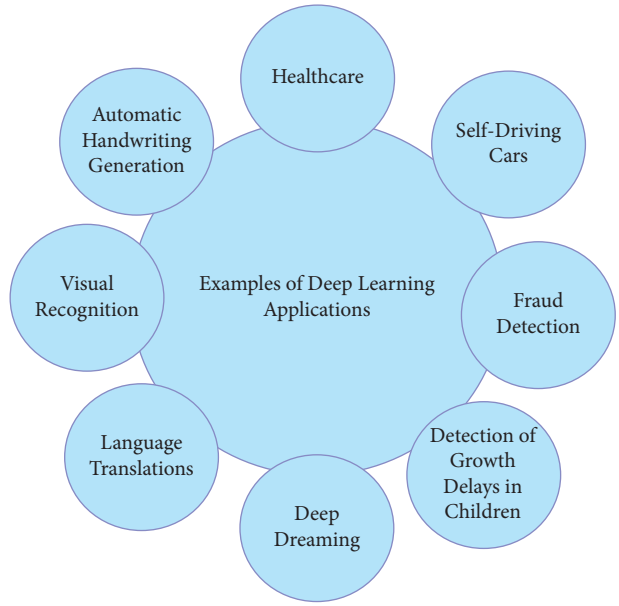

Figure 1: Examples of DL applications.

The primary and significant deep learning approach consists of restricted Boltzmann machines, autoencoders, and convolutional neural networks (CNNs) which have focused on understanding the satellite imagery or aerial imagery [23]. Hence, this study has reviewed several semantic segmentation models with diverse deep learning algorithms for future works.

The major contribution of this survey is (i) to design a detailed survey on existing semantic segmentation models on diverse imaging modalities in recent years by gathering the noteworthy information from each and every semantic segmentation model along with diverse algorithms on machine learning and deep learning, (ii) to present a comprehensive study about datasets, simulation platforms, chronological review, performance metrics, features, and challenges of the conventional semantic segmentation models and their algorithms focused, and (iii) to give the appropriate research gap with the limitations present in existing semantic segmentation systems for motivating the researchers to design a new semantic segmentation model.

The remaining sections of this survey are depicted here. Section 2 discusses the literature review on state-of-the-art semantic segmentation models. Section 3 presents the algorithmic categorization and features and challenges of existing semantic segmentation models. Section 4 describes the simulation platforms and dataset description for conventional semantic segmentation models. Section 5 demonstrates the performance measures and best accuracy rate

attained by the conventional semantic segmentation models. Section 6 gives the research gaps and challenges. Section 7 concludes this survey.

2. Literature Review on State-of-the-Art Semantic Segmentation Models

2.1. Literature Survey. In 2015, Saito et al. [24] have utilized CNN for training the pixel labeling to get the extracted building areas for determining the semantic segmentation of aerial images. Then, they have used Dijkstra's algorithm for discovering the optimal seam line to get shortest path on the map. In 2016, Marmanis et al. [25] have described the semantic segmentation model using high-resolution aerial images and using ENSEMBLE OF CNNS named FCN and modified CNNs to show the superior efficiency on standard dataset. In 2017, Holliday et al. [26] have addressed the semantic segmentation model by applying the model compression techniques for getting the superior segmentation accuracy, which has also used ConvNet to determine the significance of segmentation.

In 2018, Chen et al. [27] have suggested shuffling CNNs for realizing the aerial images for semantic segmentation in a periodic way, which has also proposed a field-of-view improvement for improving the predictions. This model has attained effective and promising results for two datasets. In 2018, Yu et al. [28] have designed an end-to-end scheme for semantically segmenting the high-resolution aerial images by considering the CNN structure with pyramid pooling phase for extracting the feature maps at diverse scales. In 2018, Chen et al. [29] have presented the digital surface models (DSMS). They have presented the deeply supervised shuffling convolutional neural network (DSCNN) for efficient upsampling of feature maps, and furthermore, the multiscale features were attained. In 2018, Volpia and Tuia [30] have suggested a semantic segmentation model using aerial images for leaning the shallow-to-deep visual features, semantic boundaries across classes, and semantic class likelihoods through a multitask CNN. Here, the top-down and bottom-up information were combined and encoded with a conditional random field model. In 2018, Sun et al. [31] have implemented a new semantic segmentation model from LIDAR data and high-resolution aerial images through a multifilter CNN for offering multiresolution segmentation. It has also delineated the object boundaries to reduce the salt and pepper artifacts. In 2018, Kemker et al. [32] have designed a semantic segmentation method using DCNNs from multispectral remote-sensing images for getting the efficient performance on RIT-18 dataset. In 2018, Marmanis et al. [33] have designed a semantic segmentation model from high-resolution aerial images by applying DCNN for representing and extracting the boundaries among the regions of diverse semantic classes. In 2018, Vo and Woong [34] have designed a semantic segmentation method through investigating the effects of deep network and cascaded framework of dilated convolutions, which has improved the localization efficiency. This model has trained efficiently.

In 2019, Peng et al. [35] have presented a new architecture by combining the "dense connection and fully convolutional networks (FCN)" for providing the fine-grained semantic segmented maps for remote-sensing images. The suggested model has achieved the traditional efficiency on two datasets without any postprocessing and pretraining. In 2019, Luo et al. [36] have proposed a new deep FCN with channel attention mechanism (CAM-DFCN) for semantic segmentation using high-resolution aerial images, which has included encoder-decoder architecture. The integration of multilevel feature maps has also facilitated. It has also offered accurate segmentation for offering spatial location information and weight semantic information. In 2019, Li et al. [37] have designed a road segmentation system with the combination of "adversarial networks with multiscale context aggregation." This study has focused on extracting the road by utilising the UAV remote-sensing images. This model has used morphological techniques for getting the results with the elimination of small independent patches. In 2019, Azimi et al. [38] have designed a symmetric FCN improved with wavelet transform for doing the segmentation of lane marking from aerial imagery. This model has used a customized loss function for improving the accuracy of pixel-wise localization. In 2019, Wang et al. [39] have designed a semantic segmentation from UAV-taken images for generating the defect detection outcomes through applying matrix operations with segment connection technique for connecting the segment features of objects. It has also used an artificial contour segment feature generator with a background filter which was used for line accessory detection that has enhanced the detection efficiency. In 2019, Cao et al. [40] have suggested a digital surface fusion models (DSMF) for improving the semantic segmentation results along with four end-to-end networks named DSMFNets to get the overall accuracy on segmenting the high-resolution aerial images. In 2019, Nguyen et al. [41] have suggested a MAVNet for semantic segmentation with the use of deep neural network on microaerial vehicles (MAVs). It has demonstrated the superior efficiency on standard datasets. In 2019, Guo et al. [42] have integrated the super-resolution approaches for improving the segmentation efficiency using "efficient subpixel convolutional neural network (ESPCN) and UNet" using remote-sensing imagery. It has significantly attained more precise and high accurate segmentation results. In 2019, Igonina and Tiumentseva [43] have focused on identifying the known neuroarchitectures to solve the problems persists in remote sensing of Earth's surface, which has also focused on semantic segmentation of UAV images. In 2019, Wu et al. [44] have studied attention dilation-linknet (AD-linknet) neural network by adopting the encoder-decoder framework along with pretrained encoder, channel-wise attention scheme, and serial-parallel integrated dilated convolution for semantic segmentation of high-resolution satellite images. In 2019, Masouleh and Shah-Hosseini [45] have presented a Gaussian–Bernoulli restricted Boltzmann machine (GB-RBM) for the semantic segmentation of UAV-based thermal infrared images, which has evaluated the efficiency on average processing time and average precision concerning with

the extraction of ground vehicles in road. In 2019, Audebert et al. [46] have introduced a regression-based semantic segmentation regularization model through a distance transform, in which the FCN was trained for both continuous and discrete spaces through learning the distance regression and joint classification. In 2019, Mohammadi et al. [47] have implemented a semantic segmentation model from polarimetric synthetic aperture radar images using FCN architecture, which has extracted the discriminative polarimetric features for finding the wetland on complex land cover ecosystem. In 2019, Hua et al. [48] have presented a CNN for processing the extracted features for enhancing the efficiency of semantic segmentation of aerial images, which has used two modules such as patch attention module and attention embedding module for getting the significant information of low level features. In 2019, Panboonyuen et al. [49] have designed a global convolutional network (GCN) for semantic segmentation of remotely sensed images for extracting the multiscale features from diverse phases of the network.

In 2020, Liu et al. [50] have proposed a semantic segmentation model for high-resolution remote-sensing images using a multichannel segmentation network termed DAPN that has completely extracted the multiscale features of the images and retained the spatial features of the object. In 2020, Mou et al. [51] have considered two efficient networks called channel and spatial relation module for learning and reasoning about the global correlations among the feature maps or positions. The suggested model was termed as relation module-equipped FCN. In 2020, Wang et al. [52] have designed a "context and semantic enhanced high-resolution network (CSE-HRNet)" with two comprehensive processes for tackling the intraclass heterogeneity problem and for enhancing the representational ability of multiscale contexts. In 2020, Martinez-Soltero et al. [53] have utilized CNN for terrain detection using aerial images, which has aimed for solving the navigation tasks and robot mapping along with the pixel-level segmentation for generating a high detailed map. In 2020, Jiawe et al. [54] have proposed a real-time semantic segmentation model by designing a new "asymmetric depth-wise separable convolution network (ADSCNet)" for offering the better prediction efficiency. In 2020, Deng et al. [55] have developed a semantic segmentation network from UAV images for real-time weed mapping for reducing the time gap among the herbicide treatment and image collection. This model has focused on implementing a hardware system with combined processes. In 2020, Niu et al. [56] have designed a new "hybrid multiple attention network (HMANET)" for adaptive capturing of global relationships, which has computed the category-based relationship and recalibrated the class level details. This study has introduced an efficient region shuffle attention (RSA) module for enhancing the effectiveness of semantic segmentation. In 2020, Chai et al. [57] have proposed the semantic segmentation model from high-resolution aerial images that has addressed the problem of learning spatial context through Deep CNNs (DCNNs). This model has predicted the distance map rather than the score map for every class that has enhanced the segmentation efficiency. In 2020, Song et al. [58] have offered the sunflower lodging detection method from remote-sensing images by considering the deep semantic segmentation and image fusion from UAV, which has attained by improved SegNet. In 2020, Diakogiannis et al. [59] have suggested a reliable framework with "ResUNet-a" for semantic segmentation of high-resolution aerial images along with dice loss function through UNet encoder-decoder network. In 2020, Ye et al. [60] have introduced Uavid dataset for semantic segmentation of urban scenes through ensemble learning including multispectral dilation with feature space optimization (FSO). In 2020, Bianco et al. [61] have suggested a semantic segmentation model for detecting the road participants and road lane through a multitask instance segmentation neural network. This model has developed an ad-hoc training process for composing the final annotations utilized to train the suggested model by applying the CNN. In 2020, Mi and Chen [62] have introduced "superpixel-enhanced deep neural forest (SDNF)" for improving the classification capability from remote-sensing images along with the semantic segmentation, which has also designed a "superpixel-enhanced region module (SRM)" for reducing the noises and improves the edges of ground objects. In 2020, Zhang et al. [63] have proposed a new fused network with the model-agnostic metalearning (MAML) and FCNN for semantic segmentation of remote sensing based on RGB images along with the optimization algorithm, particle swarm optimization (PSO) algorithm. In 2020, Boonpook et al. [64] have proposed a multifeature semantic segmentation from images of UAV photogrammetry using the deep learning method, in which the accuracy of building extraction has improved with help of SegNet. In 2020, Yang et al. [65] have focused on understanding the pixel-level information from high-spatial resolution remote-sensing images using end-to-end network called residual network (ResNet), which has also considered several additional losses for enhancing the suggested model with optimization of multilevel features. In 2020, Mehra et al. [66] have suggested a semantic segmentation method for classifying the land cover through "six deep learning architectures such as pyramid scene parsing, UNet, and deeplabv3, path aggregation network, encoder-decoder network, and feature pyramid network," which has attained superior results. In 2020, Tasar et al. [67] had proposed a semantic segmentation method by using color mapping GAN named ColorMAPGAN, which has also used element-wise matrix manipulation to learn the transformation of colors in the training data to the colors of the test data. In 2020, Venugopal [68] has suggested "a feature learning method named deep lab dilated CNN (DL-DCNN)" for automatic semantic segmentation for determining the correlation among two images, which has shown the superior efficiency over existing methods.

In 2021, Girisha et al. [69] have an improved encoder-decoder-based CNN architecture termed Uvid-Net for semantic segmentation from UAV video frames. This architecture was used to incorporate the temporal smoothness, which has captured the correlation among the sequence of frames using multibranch CNNs. In 2021, Huang et al. [70] have suggested an attention-guided label refinement

network (ALRNet) to enhance the semantic labeling of very high-resolution remote-sensing images with the encoder-decoder framework. Here, attention-guided feature fusion (AGFF) module was significantly developed for declining the semantic gap among diverse levels of features. In 2021, Abdollahi et al. [71] have suggested a GAN for segmenting the roads from high-resolution aerial imagery. This model has also used a modified UNet model (MUNet) for attaining the suitable results. In 2021, Alam et al. [72] have suggested an integrated framework using CNN with enhanced UNet and "encoder-decoder CNN structure SegNet with index pooling" for semantic segmentation of remote-sensing images, which has attained appropriate segmentation results on multitargets. In 2021, Anagnostis et al. [73] have suggested a semantic segmentation approach for obtaining the orchard trees from aerial images, which has used UNet for improving the efficient performance in terms of accuracy. This designed model has focused on automatic localization and detection of the canopy of orchard tress on different constraints. In 2021, Li et al. [74] have proposed a semantic segmentation model for analyzing the properties of photovoltaic, which has also enhanced the recommendations of segmenting the PV. It has revealed the high nonconcentrated and class imbalance distribution of photovoltaic panel image data through hard sampling and soft sampling. In 2021, Wang et al. [75] have designed a real-time semantic segmentation of high-resolution aerial images named an aerial bilateral segmentation network (Aerial-BiseNet) for offering superior accuracy. This suggested model has used two modules termed "feature attention module (FAM) and channel attention-based feature fusion module (CAFFM)" for analyzing the features. In 2021, Vasquez-Espinoza et al. [76] have suggested a semantic segmentation scheme using indoor imagery through the exploitation of details offered with the metadata utilized in the training stage of UNet. In 2021, Chen et al. [77] have considered different existing approaches such as "deeplabv3, generative adversarial network Pix2Pix, and UNet" for semantic segmentation of partially occluded apple trees, which has provided more details on branch paths, where the recovery of finer details from occlusions was offered. In 2021, Tasar et al. [78] have suggested a coined DAugNet for the semantic segmentation of satellite images, including a data augmentor and classifier, which have performed on life-long, multitarget, multisource, single-source, and single-target problems. In 2021, Li et al. [79] have recommended a "dual attention deep fusion semantic segmentation network of large-scale satellite remote-sensing images (DASSN_RSI)" for getting the significant results which have also analyzed the challenges of conventional semantic segmentation approaches using remote-sensing images. In 2021, Jiang [80] has suggested a semantic segmentation model using high-resolution remote-sensing images through CNN and mask generation, in which the NN architecture was intended for obtaining a precise mask. In 2021, Liu et al. [81] have designed a new semantic segmentation model using remote-sensing images using Inceptionv-4 network for getting the enhanced classified information. This model has introduced the fusion of features for solving the classification of edge of objects. In 2021, Zheng et al. [82] have implemented an "end-to-end CNN network named GAMNet" for balancing the controversies among the local and global information, which has also realized the boundary recovery and multiscale feature extraction. In 2021, Ouyang and Li [83] have offered a new DSSN called attention residual U-shaped network (AttResUNet) for encoding the feature maps and refining of features through attention module, which has also used GCN for classification.

2.2. Chronological Review. The chronological review on semantic segmentation models through deep learning approaches in the past years is given in Figure 2. The semantic segmentation is emerged as a major research area after 2015, and thus, this survey is prepared by gathering a set of research works from the year of 2015 to 2021. In the years of 2015, 2016, and 2017, the total number contributions is taken as 1.67% for each. Similarly, at 2018, 13.3% of the research works are gathered for analysis. In the year of 2020, 31.6% of the contributions are considered for evaluation. Likewise, while considering the 2019 and 2021, the number of research works is taken as 25%, respectively.

2.3. Security and Privacy Issues in Deep Learning. Many applications of deep learning in everyday life are self-driving cars, biometric security, health prediction, speech processing, financial technology, and retail [84]. Depending on the nature of the data and the user's intent, each application has its own set of requirements. Many models were offered by the researchers to fit the application needs, users, and features of each sort of application, including LeNet, VGG, GoogleNet, Inception, and ResNet. Despite the fact that many studies on both attacking and safeguarding users' privacy and security measures have been published, they remain fragmented. Tramèr evaluated different attack strategies based on FGSM and GAN before proposing the R-FGSM algorithm [85]. Xiaoyong Yuan also discusses security vulnerabilities in the deep learning approach. [86]. The preceding research has solely focused on the security of the deep learning model and does not provide an overview of preserving privacy in the deep learning model [87, 88].

In this work, we cover current studies on model security and data privacy that have led to the development of a secure and private artificial intelligence (SPAI). To address the demand for strong artificial intelligence (AI) systems, we compiled fragmented results and methodologies with the goal of delivering insights important to future study.

To conclude, we examine current research on privacy and security problems related to DL in the areas listed below.

(1) DL model attacks: the two primary forms of DL attacks are evasion and poisoning attacks, with evasion attacks involving the inference phase and poisoning attacks involving the training phase

(2) Defense of DL models: the different defense mechanisms presented may be divided into two broad categories based on the kind of attack, evasion and poisoning; tactics applied against evasion assaults

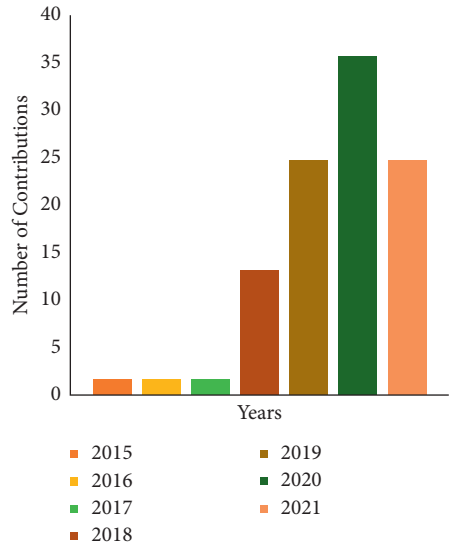

Figure 2: Chronological review of the semantic segmentation model using deep learning.

can be further divided into empirical (e.g., gradient masking, robustness, and detection) and certified approaches

(3) Privacy attacks on AI systems: the potential privacy threats to DL-based systems arising from service providers, information silos and users

(4) Defense against a privacy breach: the most modern cryptographic protection approaches, such as homomorphic encryption, safe multiparty computing, and differential privacy

According to training and testing stages in deep learning model security, attack techniques are categorized. This research emphasises on threats at the testing. Furthermore, the categorization is based on the attacker's expertise as well as the attacker's pattern of assaulting black boxes and white boxes. Attack strategies are classed in order to safeguard user privacy based on the system design and the attacker's knowledge. Attack strategies are divided into two categories in system architecture: centralized and distributed. According to the information, the attacker is also split into white box and black box attacks. Based on the stages of the deep learning model, defensive techniques are classified.

The assumptions for implementing certain threats in deep learning security are based on situations. The threat models are classified depending on the adversary's knowledge, the goal of the attacker, and the frequency of attacks.

2.3.1. The Adversary's Knowledge. A black box attack occurs when the attacker lacks knowledge of the system, in case of which the attacker submits input and receives output without understanding the system parameters. In contrast, in the event of a white box attack, the attacker has access to all system information, including the model's structure and parameter values.

2.3.2. Attacker's Target. Targeted attacks detect certain data or object types that misclassify this data collection. These types of attacks are common when categorization systems are used. In face recognition or authentication systems, for example, an attacker selects a certain face, one of which is misclassified among hostile samples. Nontargeted attacks, on the contrary, choose arbitrary data and are simpler to execute than targeted attacks.

2.3.3. Frequency of Attacks. One-time attacks require only one hostile example to be created. Otherwise, repeated attacks build adversarial instances through multiple updates. Iterative attacks outperform one-time attacks every time, but they need more queries to the deep learning system and take longer.

Deep learning security threats are classified into two types: adversarial and poisoning. We will concentrate on adversarial assaults in this research. During a system query, an adversarial attack introduces noise to the usual data. When the attacker receives the reported results, he or she utilizes this information to generate adversarial instances. This type of assault may be found in image processing, audio processing, and virus detection. It can trick deep learning machines, but not humans, particularly in the field of image processing. The gap between the source data and the adversarial example is represented by the noise value.

2.4. Limitations and Alternate Solutions of Deep Learning. Several challenges are frequently taken into account when adopting DL. Those that are more difficult are mentioned next, with various viable solutions supplied.

2.4.1. Training Data. Because it also requires representation learning, DL is tremendously data-hungry. To produce a well-behaved performance model, DL necessitates a massive quantity of data, i.e., as the data accumulates, an even more well-behaved performance model may be achieved. Most of the time, the supplied data are adequate to generate a solid performance model. However, there are situations when there is insufficient data to use DL directly. There are three proposed techniques for dealing with this issue. The first entails using the transfer-learning idea after collecting data from similar activities. While the transmitted data will not directly enhance the real data, it will aid in improving both the original input data representation and its mapping function. The model's performance is improved as a result. Another method is to use a well-trained model from a comparable assignment and fine-tune the end of two layers, or even one layer, depending on the limited original data. The second option involves data augmentation. Because picture translation, mirroring, and rotation frequently do not modify the image label, this activity is extremely useful for supplementing image data. In contrast, it is critical to exercise caution while using this approach in some circumstances, such as with bioinformatics data. When mirroring an enzyme sequence, for example, the resulting data may not represent the real enzyme sequence. In the third

way, simulated data may be used to increase the size of the training set. If the problem is sufficiently understood, it is sometimes possible to construct simulators based on the physical process. As a result, the end product will comprise the simulation of as much data as is required.

2.4.2. Transfer Learning. Deep CNNs, which provide ground-breaking help for solving numerous classification issues, have been widely used in recent research. Deep CNN models, in general, need a large amount of data in order to function well. The most prevalent problem with employing such models is a lack of training data. Gathering a big number of data is a demanding task, and no viable solution is currently available. As a result, the undersized dataset problem is now being addressed utilising the TL approach, which is very efficient in handling the lack of training data issue. The TL technique entails training the CNN model with vast amounts of data. The model is then fine-tuned for training on a small request dataset.

The student-teacher interaction is an effective method for explaining TL. The first step is to learn everything there is to know about the subject. The teacher then gives a "course" by imparting the material over time through a "lecture series." Simply put that the instructor transmits information to the pupil. More specifically, the expert (teacher) imparts knowledge (information) to the learner (student). Similarly, the DL network is trained using a large amount of data and learns the bias and weights during training. These weights are then transmitted to several networks in order to retrain or test a comparable unique model. As a result, the innovative approach can pretrain weights rather than requiring training from beginning.

2.4.3. Data Augmentation Techniques. Data augmentation techniques are one viable answer if the aim is to expand the quantity of accessible data while avoiding overfitting. These strategies are data-space solutions to any problem with little data. Data augmentation refers to a set of approaches for improving the properties and quantity of training datasets. As a result, when these strategies are used, DL networks perform better. Following that, we will go through some other data augmentation solutions.

(i) Flipping*:* vertical axis flipping is a less prevalent procedure than horizontal axis flipping. On datasets such as ImageNet and CIFAR10, flipping has been shown to be beneficial. Furthermore, it is really simple to implement. Furthermore, it is not a label conserving transformation on datasets involving text recognition (such as SVHN and MNIST).

(ii) Color space*:* as a dimension tensor, encoding digital picture data is often utilized (height *x* width *x* colour channels). Performing enhancements in the colour space of the channels is an alternate method that is particularly practical for implementation. Color augmentation is as simple as isolating a channel of a certain colour, such as red, green, or blue. By dividing that matrix and introducing extra double zeros from the remaining two colour channels, you may quickly transform a picture utilising a single-color channel. Furthermore, the picture brightness may be increased or decreased by utilising simple matrix operations to modify the RGB values. Additional better colour augmentations can be acquired by generating a colour histogram that represents the image. Lighting changes can also be done by altering the intensity values in histograms similar to those used in photo-editing software.

(iii) Cropping*:* cropping a prominent region of every single image is a technique used as a specialised processing step for image data with combined dimensions of height and width. Furthermore, random cropping can be used to achieve the same effect as translations. The distinction between translations and random cropping is that translations preserve the image's spatial dimensions, but random cropping decreases the input size. The label-preserving transformation may not be addressed because to the cropping reduction threshold that was chosen.

(iv) Rotation*:* rotation augmentations are created by rotating a picture left or right from 0 to 360° around the axis. The rotation degree parameter has a significant impact on the applicability of rotation augmentations. Small rotations (from 0 to 20°) are quite useful in digit identification tasks. When the rotation degree rises, however, the data label cannot be kept post-transformation.

(v) Translation*:* shifting the picture up, down, left, or right is a highly important transformation for avoiding positional bias in image data. For example, it is typical for all of the photos in a dataset to be centred; also, the tested dataset should be fully composed of centred images in order to test the model. It is worth noting that, after translating the starting pictures in a certain direction, the remaining space should be filled with Gaussian or random noise, or a constant value such as 255 s or 0 s. Using this padding, the spatial dimensions of the picture after augmentation are kept.

(vi) Noise injection: this method entails introducing a matrix of arbitrary values. A Gaussian distribution is typically used to generate such a matrix. Injecting noise into photos allows the CNN to learn more robust features.

2.4.4. Interpretability of Data. DL approaches are occasionally studied to serve as a black box. They can, in fact, be interpreted. Many areas, such as bioinformatics, have a requirement for a way of interpreting DL, which is utilized to acquire the valuable motifs and patterns detected by the network. It is necessary not only to understand just the illness diagnosis or prediction findings of a trained DL model but also how to improve the certainty of the prediction outcomes, as the model bases its choices on these verifications. To do this, each section of the specific example

can be assigned a weighted score. Backpropagation-based techniques or perturbation-based approaches are employed in this solution. A fraction of the input is altered in the perturbation-based techniques, and the effect of this modification on the model output is monitored. This notion has a high computational complexity, yet it is easy to grasp. With contrast, in backpropagation-based approaches, the signal from the output propagates back to the input layer to verify the score of the relevance of distinct input sections.

2.4.5. Overfitting. Because of the large number of parameters involved, which are complexly interrelated, DL models have an extremely high risk of resulting in data overfitting during the training stage. Such circumstances limit the model's capacity to perform well on the tested data. This issue is not just restricted to a single field, but also encompasses a variety of duties. As a result, while proposing DL approaches, this issue should be thoroughly examined and handled correctly. According to current research, the inherent bias of the training process helps the model to overcome critical overfitting concerns in DL. Nonetheless, strategies for dealing with the overfitting problem must be developed. An examination of the various DL algorithms for easing the overfitting problem may be divided into three categories. The first class contains the most well-known methods, such as weight decay, batch normalisation, and dropout, and it operates on both the model architecture and model parameters. Weight decay is the default approach in DL, and it is used widely as a universal regularizer in practically all ML algorithms. The second class is concerned with model inputs such as data corruption and data augmentation. One cause of overfitting is a paucity of training data, which causes the learnt distribution to differ from the true distribution. Data augmentation increases the size of the training data. In contrast, marginalised data corruption improves the solution solely through data augmentation. The last class is concerned with the model's output. For regularising the model, a recently developed method penalises overconfident outputs. This approach has been shown to be capable of regularising RNNs and CNNs.

2.4.6. Vanishing Gradient Problem. In general, when utilising backpropagation- and gradient-based learning approaches with ANNs, an issue known as the vanishing gradient problem emerges, particularly, during the training stage. In further detail, during each training iteration, each weight of the neural network is updated depending on the current weight and is proportionately relevant to the partial derivative of the error function. However, owing to a vanishingly tiny gradient, this weight update may not occur in some situations, implying that no more training is feasible and the neural network would cease entirely. In contrast, the sigmoid function, such as other activation functions, compresses a huge input space to a compact input region. As a result of the huge fluctuation at the input resulting in a little variation at the output, the derivative of the sigmoid function will be small. Only a few layers in a shallow network employ these activations, which is not a big deal. While having additional layers causes the gradient to become very tiny during the training stage, the network operates effectively in this scenario. The gradients of neural networks are determined using the backpropagation approach. Initially, this approach identifies the network derivatives of each layer in reverse order, beginning with the most recent layer and moving back to the first. The next step is to multiply the derivatives of each layer along the network in the same way that the previous step was done. When there are N hidden layers, for example, multiplying N small derivatives together requires an activation function such as the sigmoid function. As a result, the gradient decreases exponentially as it propagates back to the first layer. Because the gradient is modest, the biases and weights of the initial layers cannot be updated efficiently during the training stage. Furthermore, because these early layers are typically vital in detecting the main aspects of the input data, this circumstance reduces total network accuracy. However, by using activation functions, such an issue may be avoided. These functions lack the squishing attribute, which allows them to squish the input space to a tiny space. The ReLU is the most preferred choice for mapping X to max since it does not provide a modest derivative that is useful in the field. Another option is to use the batch normalisation layer. As previously stated, the difficulty arises when a huge input space is squeezed into a tiny space, resulting in vanishing the derivative. Using batch normalisation mitigates this problem by simply normalising the input, i.e., the expression $|x|$ does not achieve the sigmoid function's outside borders. The normalisation procedure causes the majority of it to fall into the green region, ensuring that the derivative is large enough for future activities. Furthermore, faster hardware, such as that supplied by GPUs, can address the above issue. In comparison to the time necessary to notice the vanishing gradient problem, this enables normal backpropagation over many deeper levels of the network.

2.4.7. Exploding Gradient Problem. The gradient problem is the inverse of the vanishing problem. Specifically, during backpropagation, huge error-gradients accrue. The latter will result in extraordinarily big modifications to the network's weights, causing the system to become shaky. As a result, the model's capacity to learn successfully will deteriorate. Moving backward in the network during backpropagation causes the gradient to expand exponentially by repeatedly compounding gradients. As a result, the weight values may get extremely big and may overflow to produce a not-a-number (NaN) value. Some potential solutions include

(1) Using different weight regularization techniques
(2) Redesigning the architecture of the network model

2.4.8. Underspecification. In 2020, a Google team of computer scientists found a new difficulty known as underspecification. When evaluated in real-world applications such as computer vision, medical imaging, natural language processing, and medical genomics, machine learning models, particularly, deep learning models, frequently

exhibit startlingly low performance. Underspecification is to blame for the poor performance. It has been demonstrated that modest changes may push a model to an entirely new solution and result in different predictions in deployment domains. There are several methods for dealing with the issue of underspecification. One of them is to create "stress tests" to see how well a model performs on real-world data and to identify potential problems. Nonetheless, this necessitates a solid grasp of the process, as the model can perform incorrectly. "Designing stress tests that are well-matched to application criteria and that give adequate "covering" of probable failure modes is a huge problem," the researchers concluded. Underspecification severely limits the trustworthiness of ML predictions and may necessitate some reconsideration of some applications. Because ML is tied to humans through applications such as medical imaging and self-driving automobiles, it will necessitate careful consideration of this issue.

2.5. Computational Approaches and Comparison between Different Aspects Related to Devices. Complex ML and DL algorithms have quickly emerged as the most significant techniques for computationally exhausting applications, and they are widely applied in a variety of domains. The creation and refinement of algorithms, together with the capabilities of well-behaved computational performance and massive datasets, allow for the successful execution of various applications that were previously either impossible or difficult to conceive.

2.5.1. CPU-Based Approach. The CPU nodes' well-behaved performance frequently aids robust network connectivity, storage capabilities, and huge memory. Although CPU nodes are more general purpose than FPGA or GPU nodes, they lack the ability to compete in raw compute facilities since this demands improved network capability and a bigger memory capacity.

2.5.2. GPU-Based Approach. GPUs are exceptionally effective for various fundamental DL primitives, including highly parallel-computing operations such as activation functions, matrix multiplication, and convolutions. Incorporating HBM-stacked memory onto modern GPU models dramatically improves bandwidth. This enhancement enables a wide range of primitives to make efficient use of all available computational resources on GPUs. In the case of dense linear algebra computations, the boost in GPU performance over CPU performance is typically 10–20 : 1.

2.5.3. FPGA-Based Approach. FPGA is widely used in a variety of functions, including deep learning. FPGA is widely used to create inference accelerators. The FPGA can be effectively configured to reduce the number of unnecessary or overhead functions in GPU systems. The FPGA, in comparison to the GPU, is limited to both poor-behaved floating-point performance and integer inference. The key FPGA feature is the ability to dynamically modify the array characteristics (at run-time), as well as to configure the array using effective design with little or no overhead. Table 1 [89] represents the comparison between different aspects related to the devices.

3. Algorithmic Categorization and Features and Challenges of Existing Semantic Segmentation Models

3.1. Algorithmic Classification. This section presents different deep learning approaches utilized for developing a semantic segmentation model as given in Figure 3.

The semantic segmentation models mostly use deep learning algorithms for getting superior accuracy with better quality. The techniques have been categorized into two sections, namely, deep learning and miscellaneous approaches. In deep learning, CNN architectures play a major role for semantic segmentation, which is extended by adopting different convolutional layers or other frameworks.

Supervised learning: in this model, training data consist of both input and desired results. These supervised learning algorithms are often accurate and fast. It has the ability of generalization that gives the precise results while processing new data without knowing a priori about the target.

CNN [24, 48, 53, 61, 80] inspires the researchers because of the superior efficiency in the area of computer vision, which has been adopted in diverse applications such as object detection, image recognition, and other fields. Figure 4 represents the architecture of convolutional neural network This architecture enhances accuracy of prediction or classification due to the large number of training samples along with building neural networks with several layers. CNN is a hierarchical system, which takes the input data as raw data through stacking a set of operations such as mapping of nonlinear activation functions, convolution, and pooling operations. This procedure is named as "feedforward operation." Due to this effective operation of CNN, it has attained superior results in the data mining and natural processing tasks when compared with the deep neural networks. Owing to the efficiency of CNN architectures, multiple CNN-based approaches are designed by integrating many ideas or integration of FCN architecture. This adoption of several networks into one framework is named as ensemble learning [60, 66, 77], which has attained superior results compared to single architecture because of the utilization of multiple layers. Ensemble of CNNs [25] is adopted by utilising several layers of CNN architecture to reduce the computational cost and avoids aliasing problem. It provides promising performance when compared to the existing models. DP-DCN [35] focuses on extracting the significant features from DSM data and spectral channels for fusing them through an encoder-decoder framework. The extended version of CNN consists of Shuffling CNNs [27], DSMFNets [28], UVid-Net [69], ESPCN [42], neuro-architectures [72], ensemble of CNNs [25], ADSCNet [54], DSCNN [29], DCNN [32, 33, 57], multitask CNN [30], multifilter CNN [31], ConvNet [26], GAMNet [82], DL-DCNN [68], and GCN [49, 83]. This modified or integrated

Table 1: A comparison between different aspects related to the devices.

Feature	Assessment	Leader
Development	CPU is the easiest to program, then GPU, and then FPGA	CPU
Size	Both FPGA and CPU have smaller volume solutions due to their lower power consumption	FPGA-CPU
Customization	Broader flexibility is provided by FPGA	FPGA
Ease of change	Easier way to vary application functionality is provided by GPU and CPU	GPU-CPU
Backward compatibility	Transferring RTL to novel FPGA requires additional work; furthermore, GPU has a less stable architecture than CPU	CPU
Interfaces	Several varieties of interfaces can be implemented using FPGA	FPGA
Processing/$	FPGA configurability assists utilization in wider acceleration space; due to the considerable processing abilities, GPU wins	FPGA-GPU
Processing/watt	Customized designs can be optimized	FPGA
Timing latency	Implemented FPGA algorithm offers deterministic timing, which is in turn much faster than GPU	FPGA
Large data analysis	FPGA performs well for inline processing, while CPU supports storage capabilities and the largest memory	FPGA-GPU
DCNN interference	FPGA has lower latency and can be customized	FPGA
DCNN training	Greater float-point capabilities provided by GPU	GPU

concept of CNN is designed for efficient semantic segmentation.

FCN (see [38, 46, 47, 51]): the basic idea of FCN includes processes such as "multilayer convolution, deconvolution, and fusion," where the convolutional layers are replaced with the fully connected layers. The image score is computed by using pixel-wise convolution. UNet [42, 73, 76] is a type of FCN that is efficient for small training dataset, which includes convolution and deconvolution layers with filters along with ReLU activation function. The modified versions of FCN are given here as integrated algorithm [71], ResU-Net-a [59], CAM-DFCN [36], relation module-equipped FCN [52], FCN-Alexnet model [55], and AD-LinkNet [44]. Improved SegNet [58] and SegNet [64] follow a FCN structure with encoder and decoder network. SegNet saves the element index in the upsampling process of the decoder network for solving the ambiguous spatial information in the resultant of deeper layers. Figure 5 depicts the architecture of fully connected Network.

FCN introduces many significant ideas: (i) end-to-end learning of the upsampling algorithm via an encoder/decoder structure that first downsamples the size of the activations and then upsamples it again, (ii) using fully convolutional architecture allows the network to take images of arbitrary size as input since there is no fully connected layer at the end that requires a specific size of the activations, and (iii) introducing skip connections as a way of fusing information from different depths in the network for multiscale inference.

GAN (see [37, 71]): generative adversarial network (GAN) model considers a softmax layer, in which the discriminator of the GAN produces label types for efficient classification of unlabeled samples and labeled examples. The architecture of generative adversarial network is shown in Figure 6. The modified version of ColorMapGAN [67] has aimed at minimizing the computational complexity and improving the accuracy.

DNN: deep neural networks (DNN) focus on semantic segmentation of high-resolution images which consist of several parameters that need a large number of labeled examples for training. A general scheme for constructing a deep network to process a rich dataset is complex. The improved DNN models are modified inceptionV-4 network [50], NDRB [52], ResNet101-v2 [39], ALRNet [70], HMANET [56], ResNet [65], inceptionV-4 network [81], MAVNet [41], and SDNF [61], which are aimed to enhance the superior accuracy on segmentation.

Unsupervised learning*:* this model is not offered with the precise results during training, which can be employed for clustering the input data in classes through statistical properties.

DAugNet [78]**:** DAugNet generates the precise maps and has provided life-long adaptation settings for giving the superior semantic segmentation results. GB-RBM [73] is introduced for enhancing the segmentation results and improving the speed and accuracy. Figure 7 gives the training procedure of data augmentation network.

3.2. Features and Challenges. The features and challenges of the conventional semantic segmentation model using deep learning techniques are listed in Table 2. This description provides the researchers for focusing on a new semantic segmentation model on aerial images for solving the existing challenges through adopting deep learning techniques.

4. Simulation Platforms and Dataset Description for Conventional Semantic Segmentation Models

4.1. Simulation Platforms. The simulation environments used for implementing a semantic segmentation model with different imaging modality is presented in Figure 8. Here, some of the tools such as CUDA version 8.0, Edge Detection and Image Segmentation (EDISON) library, MXNet, TensorRT, and two-fold validation tool are used in 1.7% of the contributions, respectively. MATLAB and Tesla use 3.3% of the contributions for implementation and Pascal and Keras utilize 8.3% of the research works with the TensorFlow as a platform, respectively. TensorFlow is used as the simulation

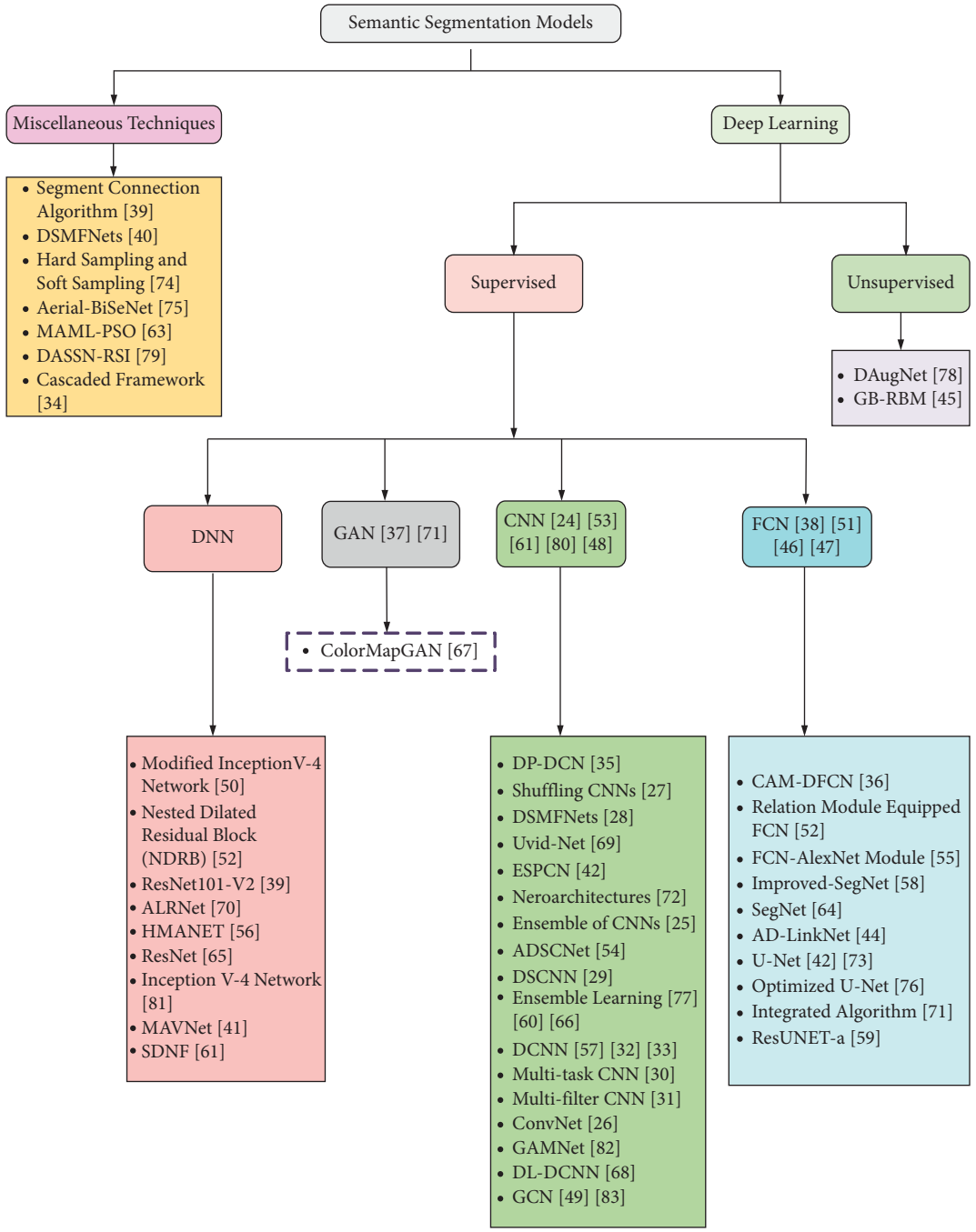

Figure 3: Algorithmic categorization of existing semantic segmentation models.

environment for 18.3% of the works and NVIDIA is considered in 5% of the contributions. Finally, the python tool is used in 6.6% of the research works and other platform environments are taken in 20% of the contributions.

4.2. Dataset Description and Imaging Modalities Focused. The dataset used for implementing the semantic segmentation model along with different imaging modalities is given in tabular forms (Tables 3–7). Most of the contributions are considered aerial images for semantic segmentation, which is used in 23.3% of the work, *s*, and high-resolution aerial imagery is taken in 16.6% of the contributions. Similarly, the remote-sensing and high-resolution remote-sensing images are taken in 25% of the research works. Unoccupied aerial vehicles' (UAVs) images are gathered in 11.7% of the contributions.

Multiscale and multispatial resolution images are included in 1.7% of the research papers, respectively, and satellite images are taken in 5% of the contributions. Other high-resolution images are taken in 13.4% of the research works.

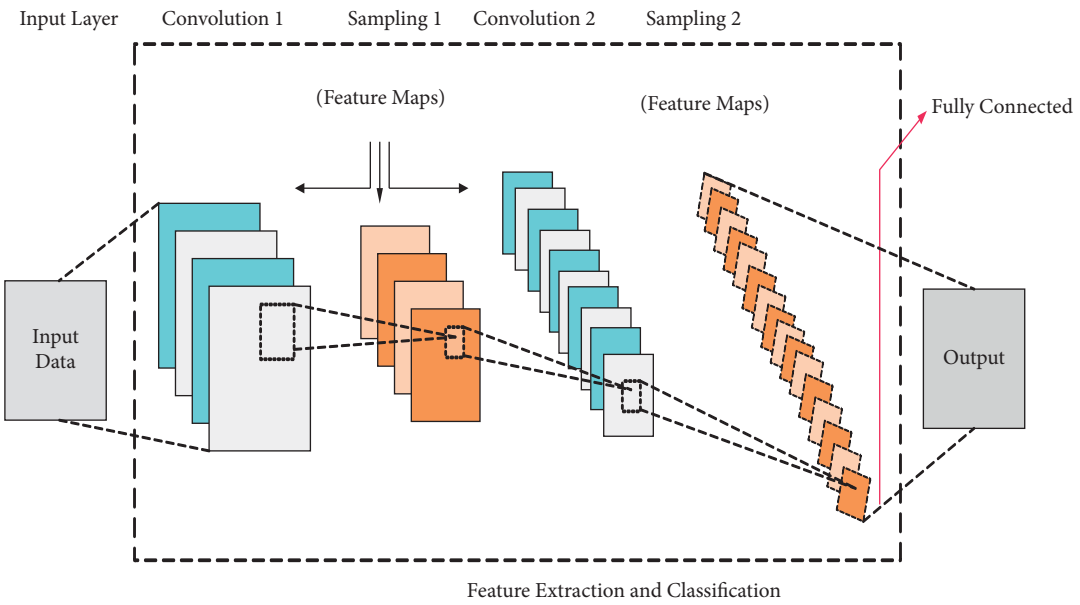

Figure 4: CNN architecture.

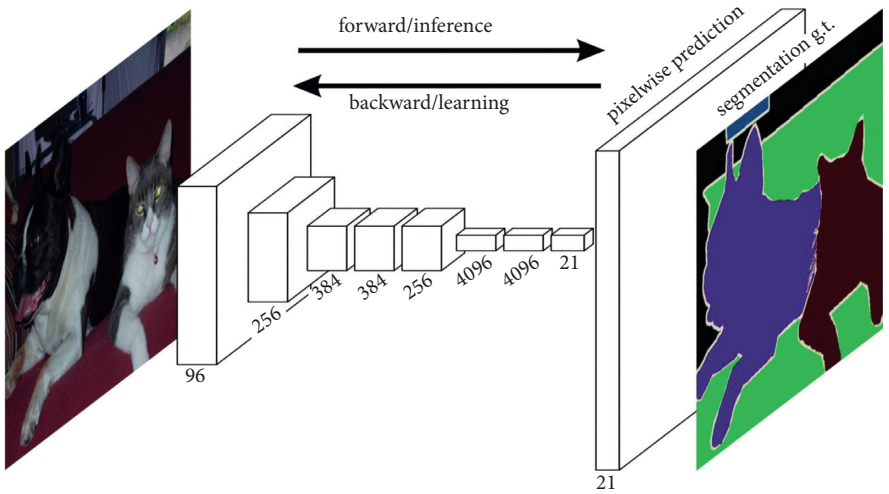

Figure 5: FCN architecture [90].

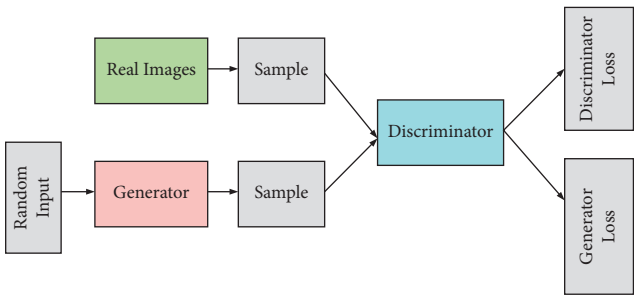

Figure 6: GAN architecture.

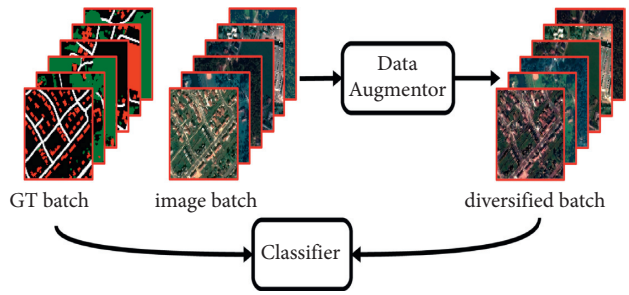

Figure 7: The training procedure of DAugNet that comprises a data augmentor and a classifier. In each training iteration, the classifier learns from the diversified batch generated by the data augmentor [78].

Table 2: Merits and demerits of existing semantic segmentation model using deep learning approaches.

Citation number	Methodology	Features	Challenges
[24]	CNN	It accurately extracts the round objects using CNN	It requires more cost for getting pixel intensities on shadow regions
[25]	Ensemble of CNNs	It reduces the computational cost and avoids aliasing problems; it provides promising performance when compared to the existing models	Conversely, multicore parallelization over diverse scenes is complex
[26]	ConvNet	It gets efficient segmentation performance with better sophistication	This model is not suitable for unlabeled data
[27]	Shuffling CNNs	This model is limited to use ensemble approaches	This model is limited to use ensemble approaches
[28]	ResNet101-v2 and a pyramid pooling module	It has offered an effective network framework with superior performance	However, the segmentation accuracy is limited while considering the large spectral similarities among imperious surfaces and buildings
[29]	DSCNN	It provides enhanced and smoother identifications for different objects	It does not offer superior numerical outcomes
[30]	Multitask CNN	It has offered a principled and flexible structure for providing the efficient segmentation results	It does not preserve the geometrical features and complex for segmentation
[31]	Multifilter CNN	It has achieved the highest overall accuracy and removed the noise	This paper does not investigate how diverse data sources from other sensors are integrated in deep CNN
[32]	DCNN	This model gets superior efficiency with discriminative frameworks; it avoids overfitting problems	However, this model does not fully remove the salt and pepper noise
[33]	DCNN	The DCNN achieves superior effectiveness on a standard dataset	However, it is a tedious and small problem that affects the segmentation quality
[34]	Cascaded framework	It has improved the prediction with object boundaries and removed the isolated false positives	It has high computational costs
[35]	Dual-path densely convolutional networks (DP-DCN)	It avoids the vanishing gradient problem and strengthens the information flow among the layers by a dense connection	However, it requires less test time and training time
[36]	CAM-DFCN	This model has attained mainstream performance; it also promotes the segmentation results with efficient feature selection	The performance of the suggested CAM-DFCN was not improved significantly while comparing to the CNN + RF + CRF
[37]	GANs with multi-scale context aggregation	This network has improved the accuracy of road extraction and offered superior visual effects	It is a computationally inefficient one, which has to enhance the segmentation precision

Table 2: Continued.

Citation number	Methodology	Features	Challenges
[38]	FCN	This model has recovered the lost data to get high robustness and accuracy	This model is not applicable for processing the shadow areas
[39]	Segment connection algorithm	It enhances the detection efficiency that enhances the applicability of the framework	The precision rate of lost vibration damper identification is less which gives a lower F1 score
[40]	DSMFNets	It shows superior fusion results using DSMFNets with efficient performance	However, the effectiveness can be affected due to the restriction on the feature extraction module
[41]	MAVNet	It has shown a better tradeoff between performance and inference time	This model is not applicable to apply modestly sized networks
[42]	ESPCN and UNet	It enhances the segmentation and improves the robustness	It suffers from insufficient training samples
[43]	Neuroarchitectures including (a) MultiNet, (b) SegNet, and (c) UNet	This model has improved the quality of object segmentation	However, the implementation is restricted
[44]	AD-LinkNet	The suggested AD-LinkNet boosts the efficiency on segmentation	The designed model does not show the better performance on different road interruptions
[45]	GB-RBM	It enhances the segmentation results; it has improved the speed and accuracy	This model is restricted on high spatial resolution thermal infrared images
[46]	FCN	The segmentation efficiency is improved while comparing with the conventional approaches; it gets less overhead	The multitask degrades the efficiency of segmentation
[47]	FCN	It efficiently discriminates the nonwetland classes from wetland classes; It enhances the accuracy of semantic segmentation	However, processing the restricted availability of ground truth data in large-scale remote-sensing applications is challenging
[48]	CNN	This model has enriched the semantic information, which has focused on attaining the representative extracted features	It lacks in performance due to the processing of high-level features
[49]	GCN	It has shown superior performance with capturing of complex features; it solves the scarcity problem	It has to enhance the accuracy by adopting different approaches such as optimization and semantic labeling
[50]	Modified InceptionV-4 network called DAPN	This technique has robust generalization ability	Although the potsdam dataset has offered consistent performance, there is a considerable reduction in the vaihingen IR-R-G dataset concerning IOU scores
[51]	Relation module-equipped FCN	The performance of semantic segmentation is enhanced with the use of a network using aerial scenes	However, the suggested relation modules regarding segmentation are basic one, and thus, it does not offer superior efficiency
[52]	Nested dilated residual block (NDRB)	It offers precise object boundaries and labeling for complex scenes	The per-class accuracy is not evaluated which does not estimate the efficiency
[53]	CNN	This model has offered the best tradeoff with the fewer number of parameters along with less memory utilization; it gives the suitable mapping of terrain	The considered images do not have a fixed shape or resolution, and thus, the training may be affected
[54]	ADSCNet	This model has reduced the network complexity because of the depth-wise convolution; it improves the performance along with better information flow	This model does not evaluate the actual inference speed
[55]	FCN-AlexNet model	This model has maintained reasonable accuracy and inference speed	The limited dataset is used for validation, which has to be rectified

Table 2: Continued.

Citation number	Methodology	Features	Challenges
[56]	HMANet	This model captures the global contextual details for efficient segmentation; it enhances the efficiency of the self-attention scheme and reduces feature redundancy	It takes huge consumption of memory
[57]	DCNN	It has shown better smoothing effects; it has extracted the multilevel features	It does not consider the complementary and orthogonal technical progressions
[58]	Improved SegNet	It increases the accuracy and speed of sunflower lodging; it efficiently monitors the lodging in equivalent low canopy density crops	The complexity of the identification is increased due to the growth and status of sunflow which is varied through spatial distribution changes
[59]	ResUNet-a	It has offered better convergence properties; the superior F1 score is observed	It shows slow operation due to the GPU synchronization that makes it impractical for future processes
[60]	Ensemble learning	It extracts multiscale features; the manual dataset offers superior performance with temporal consistency	However, this dataset has different challenges such as number of types in scenes, dataset size and large-scale differentiation for several objects
[61]	CNN	It reduces the computational constraints; it provides real-time performance	However, the weak labeling stage is observed that affects the performance
[62]	Superpixel-enhanced deep neural forest (SDNF)	It shows superior classification ability with reduced noises; it gives robust results	However, for some of the classes, the accuracy is reduced
[63]	MAML-PSO	The misclassification of objects with specific height variance can be effectively minimized by introducing LIDAR data; it increases the testing accuracy	In this model, the overall accuracy is not very good
[64]	SegNet	It shows the superior building extraction for medium- and high-sized buildings; it also enhances the classification accuracy	However, the small size buildings are complex for identification
[65]	ResNet	It has efficiently extracted the global and local deep features that offer better semantic segmentation results	This study does not consider the digital surface models on both datasets
[66]	Ensemble learning	The semantic segmentation is improved due to the extracted features	The suggested model is limited on dataset size
[67]	ColorMapGAN	The suggested model has minimized computational complexity and improved accuracy	Though, the results' quality for nonlearning-based approaches is inefficient
[68]	DL-DCNN	It has achieved better convergence rate and accelerated network training; it obtains enhanced results with efficient identification of changes	Conversely, it gets overfitting and low accuracy rate
[69]	UVid-net	It reduces the computational complexity and provides superior segmentation results on aerial videos	However, it is a laborious and time-consuming task
[70]	ALRNet	This model chooses the most nonredundant and representative features to offer outstanding efficiency	ALRNet has higher computational inefficiency
[71]	GAN	This model efficiently preserves the edge information and gets a better accuracy rate on segmenting the maps	This model lacks accuracy and also it suffers from extracting the continuous road parts or complex regions
[72]	Integrated algorithm (encoder-decoder CNN structures SegNet with index pooling and UNet)	This integrated technique has offered superior features of both CNN and UNet to offer better semantic segmentation of images	For some classes, the suggested integrated model has attained less performance than other algorithms

Table 2: Continued.

Citation number	Methodology	Features	Challenges
[73]	UNet	This approach has the capability of precise segmentation of tree canopies; it also solves complex problems in environments such as agricultural production	Although the designed model shows superior performance on detection, it does not solve the issue of densely merged and located false positives
[74]	Hard sampling and soft sampling	This model explores the heterogeneous colour feature and texture feature of the PV panel	On the contrary, the uncertainties have remained
[75]	Aerial-BiSeNet	A superior balance among the speed and accuracy is offered; it has shown better efficiency and accuracy on both datasets	It suffers from weak representation ability and high model complexity
[76]	Optimized UNet	It solves the computational complexity; it improves the overall performance	Conversely, some of the images attain the worst results because of the estimation problem
[77]	Ensemble learning	The accuracy and practical implementation is superior to other existing approaches	The efficiency can be affected due to the noise present in images
[78]	DAugNet	The precise maps are generated and have provided life-long adaptation settings	This model does not apply on sentinel and aerial images
[79]	DASSN_RSI	It has reduced the training loss and enhanced the convergence rate; it verifies the advancement and efficiency of the suggested method	It lacks in robustness, which does not focus on low-shot learning methods
[80]	CNN	The segmentation and overall training time have been reduced; it also improves the overall precision	This model does not consider low-resolution images
[81]	InceptionV-4 network	It has attained superior segmentation efficiency and training efficiency	It shows poor generalization ability
[82]	GAMNet	The efficiency of the integration module is improved; the accurate results has attained with precise boundaries even for small objects	The confusing problem is occurred and suffered from misclassification problem in shaded areas
[83]	GCN	It restores the boundaries of ground objects and reduced the pixel-level noises	However, it does not utilize the spatial correlation details for interpreting remote-sensing images

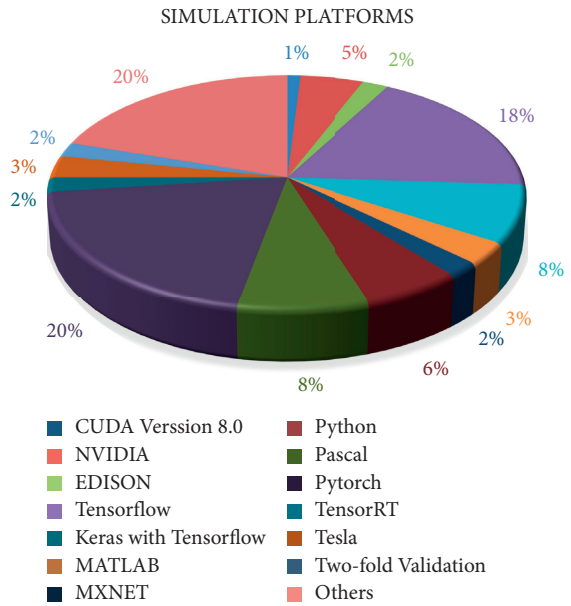

Figure 8: Simulation platforms used for implementing the semantic segmentation models.

Table 3: Dataset description based on modality of aerial imagery.

Citation number	Dataset description
[38]	Aeriallanes18 dataset
[52]	The potsdam dataset and the vaihingen dataset
[27]	ISPRS vaihingen and potsdam datasets
[24]	Manual dataset that includes 127 aerial images
[41]	Mavnet
[73]	Manual dataset
[53]	The dataset can be accessed at https://github.com/gabrielmtzsoltero/ssegfor_aerial_mapping/
[56]	ISPRS 2D semantic labeling challenging for vaihingen and potsdam
[29]	Use the vaihingen dataset
[77]	Commercial apple orchard in northeastern melbourne
[61]	Vaihingen dataset and potsdam dataset
[46]	ISPRS 2D semantic labeling and data fusion contest 2015
[30]	Dataset is composed of 33 orthorectified image tiles acquired by a near infrared (NIR)-green (*G*)-red (*R*) aerial camera, over the town of vaihingen (Germany)
[48]	ISPRS Benchmarks1, deepglobe contest2, and spacenet competition3

Table 4: Dataset description based on modality of high-resolution aerial images.

Citation number	Dataset description
[51]	ISPRS vaihingen and potsdam
[70]	Potsdam dataset, vaihingen dataset, and whu dataset
[71]	Massachusetts dataset
[25]	Vaihingen dataset
[75]	Potsdam and vaihingen datasets
[59]	ISPRS 2D potsdam dataset
[31]	ISPRS 2D semantic labeling contest of potsdam and an area of guangzhou in China
[33]	ISPRS vaihingen 2D semantic labeling challenge
[82]	ISPRS 2D semantic labeling datasets

Table 5: Dataset description based on remote-sensing images.

Citation number	Dataset description
[35]	Vaihingen and potsdam
[50]	International society for photogrammetry and remote-sensing (ISPRS) 2D semantic labeling contest potsdam and inria aerial image labeling dataset
[28]	ISPRS
[42]	Manual dataset on Tokyo
[72]	Big data and computing intelligence contest (BDCI)
[54]	Cityscapes
[58]	The remote-sensing data collected from field 1
[62]	ISPRS 2D semantic labeling benchmark dataset
[63]	2015 igrss data fusion competition
[32]	Rit-18
[79]	Gaofen image dataset (GID) datasets
[80]	Potsdam and vaihingen datasets
[67]	Luxcarta dataset
[81]	ISPRS 2D semantic labeling contest vaihingen dataset and Massachusetts building dataset
[68]	Ottawa dataset, stone gate dataset, sardinia dataset, yellow river estuary dataset, barbara dataset, and USGS dataset
[49]	Landsat-8 satellite and ISPRS vaihingen challenge dataset
[83]	UCM dataset and the deepglobe dataset

4.3. Datasets for Image Segmentation. In this section, we give a synopsis of a portion of the most generally utilized datasets for image segmentation. We combine these datasets into 3 classifications is 2-dimensional images, 2.5-dimensional RGB-D (complexity + colour) images, and 3-dimensional images and give subtle ties with regards to the attributes of each dataset. The recorded datasets have pixel-wise marks, which can be utilized for assessing model execution.

It is worth focusing on that a portion of these works, use augmentation of data to expand the quantity of marked

Table 6: Dataset description based on modality OF UN-OCCUPIED aerial vehicles (UAVS) images.

Citation number	Dataset description
[37]	UAV images of three regions (Baoxing, Jiaxing, and Chengyang)
[39]	China southern power grid company
[69]	Manual Uavid dataset and cityscape dataset
[43]	Worldview-3
[55]	Rice field located in southern China
[60]	Manual Uavid dataset
[45]	UAV-based thermal infrared imagery named NPU_CS_UAV_IR_DATA that was collected from some streets of China by using FLIR TAU2
[64]	RGB-D UAV dataset

Table 7: Dataset description based on multiscale, multispatial resolution, satellite images, and other high-resolution images.

Citation number	Dataset description
[76]	Lsun dataset
[44]	CVPR2018 deepglobe challenge
[78]	Dataset consists of pleiades images collected over five cities in Austria
[42]	Polarimetric RADARSAT-2
[66]	High-resolution images from LANDSAT-8 datasets of Google Earth engine
[34]	Challenging PASCAL VOC2012 database

samples, uncommonly the ones which manage little datasets such as in the medical domain. Augmentation of data serves to expand the quantity of preparing tests by applying a set of changes either in the information space, or element space, or now and again both to the images, i.e., both the input image and the segmentation map. Some normal changes incorporate interpretation, reflection, pivot, twisting, scaling, colour space shifting, trimming, and projections onto principal components. Augmentation of data has demonstrated to work on the presentation of the models, particularly when gaining from restricted datasets, like those in medical image investigation.

The common image segmentation research has concentrated on 2-dimensional images. From Figure 9 [91], pink, green, and yellow blocks mention semantic occurrence and panoptic segmentation algorithms, respectively. Therefore, several 2-dimensional image segmentation datasets are existing, and they are PASCAL Visual Object Classes (VOC) [92], PASCAL Context [93], Microsoft Common Objects in Context (MS COCO) [94], Cityscapes [95], ADE20 K/MIT Scene Parsing (SceneParse150) [96], SiftFlow [97], Stanford background [98], Berkeley Segmentation dataset [99], Youtube-Objects [100], KITTI [101], Semantic Boundaries Dataset (SBD) [102], PASCAL Part [103], SYNTHIA [104], Dobe's Portrait Segmentation [105], etc., With the obtainability of reasonable range scanners, RGB-D images have become standard in both research and industrial applications. Some of the most standard 2.5-dimensional RGB-D datasets are NYU-D V2 [106], SUN-3D [107], SUN RGB-D[108], UW RGB-D Object Dataset [109], ScanNet [110], etc., Three-dimensional image datasets are standard in robotic, medical image analysis, 3D scene analysis, and construction applications. Three-dimensional images are generally provided via meshes or other volumetric illustrations, such as point clouds. Some of the standard 3-dimensional datasets are Stanford 2D-3D [111], ShapeNet Core [112], Sydney Urban Objects Dataset [113], etc.

4.4. Frameworks and Benchmark Datasets Employed for Different DL Tasks. Several deep learning frameworks and datasets have been developed in the last few years. Various frameworks and libraries have also been used in order to expedite the work with good results. Through their use, the training process has become easier. Tables 8 and 9 [89] list the most utilized frameworks and libraries and Benchmark datasets.

4.5. Algorithms Comparison Based on Different Datasets. Comparison of different algorithmic features and their results obtained based on clustering methods, conditional random field, PASCAL VOC2012 dataset, CamVid dataset, and MS COCO dataset are tabulated (Tables 10–14).

5. Performance Measures and Best Accuracy Rate Attained by the Conventional Semantic Segmentation Models

5.1. Performance Metrics. An exemplary ought to preferably remain assessed in an assortment of ways, including quantitative precision, speed, and capacity necessities. The majority of previous research has concentrated on parameters for assessing model accuracy. The most commonly used parametric for evaluating the accuracy of segmentation algorithms is summarized below [91, 136]. On benchmarks, to analyze various models, quantitative measurements are

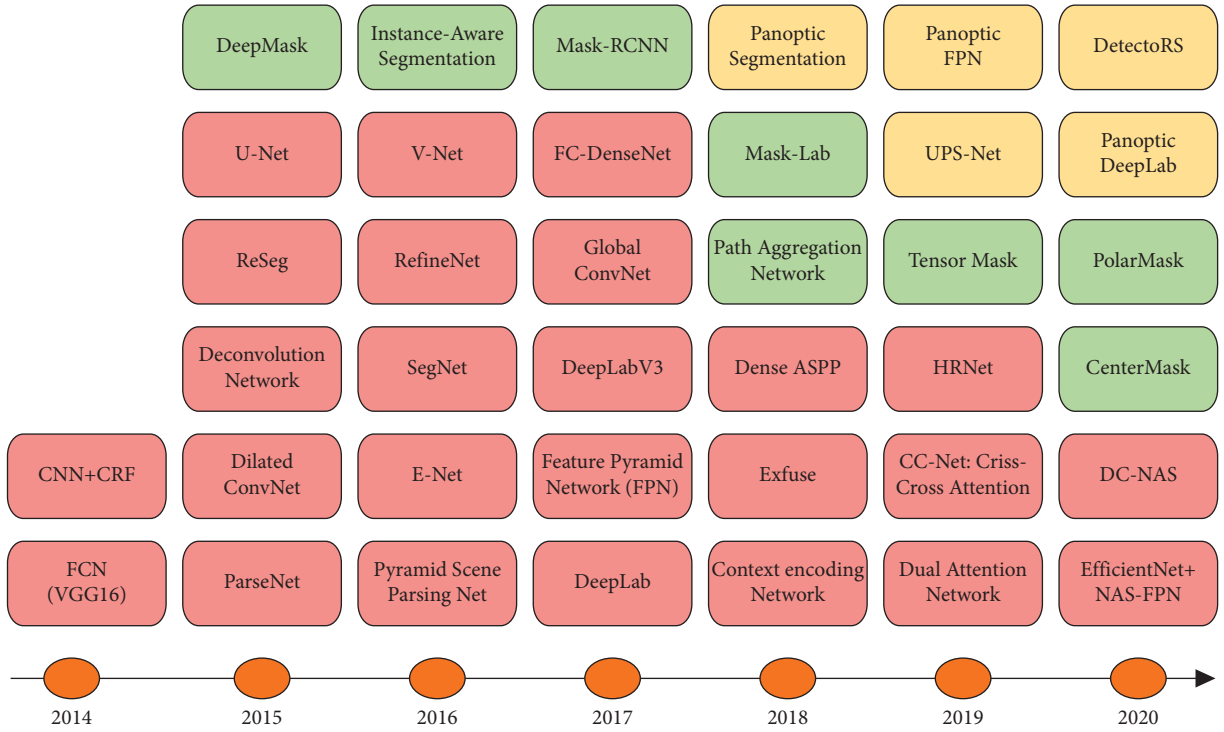

Figure 9: The evolution of deep learning-based segmentation algorithms for 2-dimensional images from 2014 to 2020 [91].

Table 8: LIST of most common frameworks and libraries.

Framework	License	Core language	Year of release	Homepages
TensorFlow	Apache 2.0	C++ and python	2015	https://www.tensorflow.org/
Keras	MIT	Python	2015	https://keras.io/
Caffe	BSD	C++	2015	http://caffe.berkelyvision.org/
MatConvNet	Oxford	MATLAB	2014	http:/www.vlfeat.org/matconvnet/
MXNet	Apache 2.0	C++	2015	https://github.com/dmic/mxnet
CNTK	MIT	C++	2016	https://github.com/Microsoft/CNTK
Theano	BSD	Python	2008	http://deeplearning.net/software/theano/
Torch	BSD	C and lua	2002	http://torch.ch/
DL4j	Apache 2.0	Java	2014	https://deeplearning4j.org/
Gluon	AWS microsoft	C++	2017	https://github.com/gluon-api/gluon-api/
OpenDeep	MIT	Python	2017	http://www.opendeep.org/

utilized, and the visual nature of the model yields significance in figuring out.

(i) Pixel accuracy (PA): basically, pixel accuracy states the ratio of correctly classified pixels to the total quantity of pixels. Pixel accuracy is known for $N+1$ classes as

$$PA = \frac{\sum_{i=0}^{n} a_{ii}}{\sum_{i=0}^{n} \sum_{j=0}^{n} aij}, \tag{1}$$

where a_{ij} is the quantity of pixels of class I predicted as belonging to class j.

(ii) Average/mean pixel accuracy (MPA): mean pixel accuracy has marginally further developed, in which the ratio of correct pixels is computed in a per-class basis and then averaged over the total number of classes:

$$MPA = \frac{1}{N+1} \sum_{i=0}^{n} \frac{aii}{\sum_{j=0}^{n} aij}. \tag{2}$$

(iii) Intersection over union (IoU): this is quite possibly the most generally utilized measurement in semantic segmentation. It is determined as the area of intersection of the predicted division map and the ground truth divided by the area of the union of the predicted segmentation map and the ground truth:

Table 9: Benchmark datasets.

Dataset	No. of classes	Applications	Link to dataset
ImageNet	1000	Image classification, object localization, object detection, etc.	http://www.image-net.org/
CIFAR10/100	10/100	Image classification	https://www.cs.toronto.edu/~kriz/cifar.html
MNIST	10	Classification of handwritten digits	http://yann.lecun.com/exdb/mnist/
Pascal VOC	20	Image classification, segmentation, and object detection	http://host.robots.ox.ac.uk/pascal/VOC/voc2012/
Microsoft COCO	80	Object detection and semantic segmentation	https://cocodataset.org/#home
YFCC100 M	8M	Video and image understanding	http://projects.dfki.unikl.de/yfcc100m/
YouTube-8M	4716	Video classification	https://research.google.com/youtube8m/
UCF-101	101	Human action detection	https://www.crcv.ucfedu/data/UCF101.php
Kinetics	400	Human action detection	https://deepmind.com/research/open-source/kinetics
Google open images	350	Image classification, segmentation, and object detection	https://storage.googleapis.com/openimages/web/index.html
CalTech101	101	Classification	http://www.vision.caltech.edu/Image_Datasets/Caltech101/
Labeled faces in the wild	–	Face recognition	http://vis-www.cs.umass.edu/lfw/
MIT-67 scene dataset	67	Indoor scene recognition	http://web.mit.edu/torralba/www/indoor.htm

Table 10: Comparison of algorithms based on clustering methods (%).

Citations	Algorithm features	Datasets	Segmentation results
[114]	Weak supervision, spectral clustering, and discriminative clustering	MSRC-21	70 (mA)
[115]	Weak supervision and double-end clustering	MSRC-21	52.9 (mIoU)
		LABLEME	26 (mA)
[116]	FCM algorithm and grouping algorithm	Self-built dataset	2.2 (mError)

Table 11: Comparison of algorithms based on conditional random field (%).

Citations	Algorithm features	Datasets	Segmentation results
[117]	CRF, dense features, and high-order potential energy	MSRC-21	75.8 (mA)
[118]	CRF and joint-boosting algorithm	MSRC-21	71.6 (mA)
[119]	CRF and interactive	Self-built dataset	95.3 (mA)
[120]	CRF and high-order energy items	MSRC-21	72.2 (PA)
[121]	CRF and maximum flow-minimum cut	MSRC-21	0.7 s (time)

Table 12: Comparison of algorithms based on PASCAL VOC2012 dataset (%).

Citations	Algorithm features	Datasets	Segmentation results
[122]	Convolution and deconvolution neural networks	PASCAL VOC2012	63.6 (mIoU)
[123]	Deconvolution networks	PASCAL VOC2012	72.5 (mIoU)
[124]	PSPNet	PASCAL VOC2012	82.6 (mIoU)
[125]	RefineNet	PASCAL VOC2012	83.4 (mIoU)
[126]	Decoupled deep neural networks	PASCAL VOC2012	66.6 (mIoU)

Table 13: Comparison of algorithms based on CAMVID dataset (%).

Citations	Algorithm features	Datasets	Segmentation results
[127]	SegNet	CamVid	60.1 (mIoU)
[128]	Densely connected convolutional networks	CamVid	66.9 (mIoU)
[129]	ENet	CamVid	51.3 (mIoU)
[130]	Gated feedback refinement networks	CamVid	68.0 (mIoU)
[131]	Generative adversarial networks	CamVid	58.2 (mIoU)

Table 14: Comparison of algorithms based on MS COCO dataset (%).

Citations	Algorithm features	Datasets	Segmentation results
[132]	Mask R-CNN	MS COCO	37.1 (PA)
[133]	FCIS	MS COCO	59.9 (PA)
[134]	Multitask network cascades	MS COCO	51.5 (PA)
[135]	Residual networks	MS COCO	48.4 (PA)

$$IoU = J(P,Q) = \frac{|P \cap Q|}{|P \cup Q|}, \tag{3}$$

where P = true segmentation map and Q = predicted segmentation maps.

The value of intersection over union lies between 0 and 1.

(iv) Mean-IoU: mean intersection over union is an alternative standard metric defined by average intersection over union across entire modules. It is commonly used in reporting the performance of contemporary segmentation algorithms [91].

(v) Precision/recall: for numerous classical image segmentation models, precision and recall are the standard metrics for recording. Definition for precision and recall for every class is as follows:

$$\begin{aligned} \text{Precision} &= \frac{TP\ Fraction}{TP\ Fraction + FP\ Fraction}, \\ \text{Recall} &= \frac{TP\ Fraction}{TP\ Fraction + FN\ Fraction}, \end{aligned} \tag{4}$$

where TP = True Positive, FP = False Positive, and FN = False Negative. Usually, we are attentive in a united form of precision and recall rates.

(vi) F1 score: F1 score is also the standard metric and defined by the harmonic mean of precision and recall:

$$F1 - Score = \frac{2\text{Precision } x\ \text{Recall}}{\text{Precision } +\ \text{Recall}}. \tag{5}$$

(vii) Dice coefficient: Dice coefficient is an alternative standard metric used in medical image analysis for image segmentation, defined by "twice the overlap area of predicted and ground truth maps, divided by the total number of pixels in both images. The Dice coefficient is very identical to the IoU" [91]:

$$Dice = \frac{2|P \cap Q|}{|\text{P}| + |\text{Q}|}. \tag{6}$$

While practical to Boolean data, the Dice coefficient is nearly equal to the F1 score:

$$Dice = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = F1score, \tag{7}$$

where TP indicates True Positive Fraction, FP indicates False Positive Fraction, and FN indicates False Negative Fraction.

(viii) Frequency weighted mIoU: over the raw mIoU, frequency weighted mean intersection over union is an improved which weights each class importance depending on their appearance frequency [136]:

$$FWmIoU = \frac{1}{\sum_{i=0}^{K}\sum_{j=0}^{K} aij} \sum_{i=0}^{K} \frac{\sum_{j=0}^{K} aij\, aii}{\sum_{j=0}^{K} aij + \sum_{j=0}^{K} aji - aii}. \tag{8}$$

(ix) Jaccard index: the Jaccard index, commonly known as the Jaccard similarity coefficient, is a statistic used to assess the similarity between sample sets. The measurement stresses similarity between finite sample sets and is officially defined as the intersection size divided by the sample set union size. The mathematical representation of the index is written as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{9}$$

(x) Confusion matrix: a Confusion matrix is an $N\ x\ N$ matrix that is used to assess the effectiveness of a classification model, where N is the number of target classes. Figure 10 represents the confusion matrix. The matrix compares the actual goal values to the machine learning model's predictions. This provides us with a comprehensive picture of how well our classification model is working and the kind of errors it is producing. For a binary classification task, we would have a 2×2 matrix with four values, as illustrated in figure [137].

Let us decode the matrix. The target variable has two values: positive or negative. The columns represent the actual values of the target variable. The rows represent the predicted values of the target variable.

(xi) Kappa coefficient: it is used to assess the level of agreement between two human evaluators or raters (for example, psychologists) when assessing topics (patients). The machine learning community then "appropriated" it to quantify categorization performance. The kappa score, also known as Cohen's kappa coefficient [138], is named after Jacob Cohen, an American statistician and psychologist who produced the foundational study on the subject. This measure is also known as Cohen's kappa and the kappa statistic. To compute the kappa score, it is

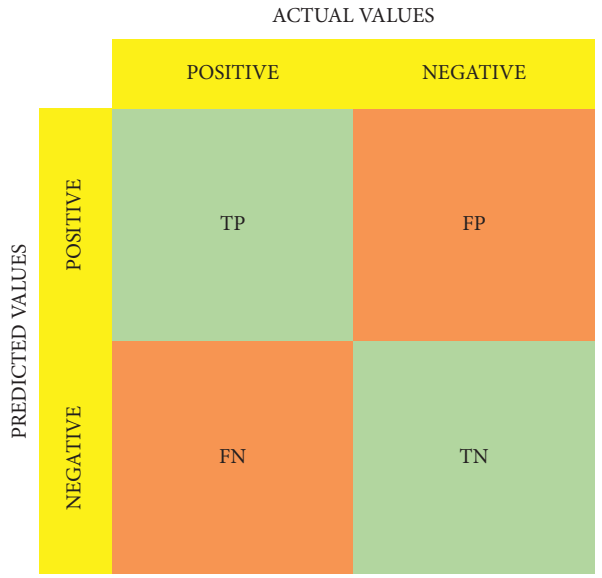

Figure 10: Confusion matrix [137].

		Professor A		
		Accept	WL	Reject
Professor B	Accept	4	6	3
	WL	1	2	0
	Reject	1	2	6

Figure 11: Summarization of ratings for kappa coefficient [138].

Table 15: The best performances were obtained by diverse semantic segmentation models.

Citations	Performance metric	Best performance in percentage
[64]	Overall accuracy	97.00
[68]	F1 score	99.41
[62]	Intersection over union	96.50
[24]	Recall	99.84

convenient to first summarize the ratings in a matrix shown in Figure 11.

The columns show the ratings by professor A. The rows show the ratings by Professor B. The value in each cell is the number of candidates with the corresponding ratings by the two professors.

The performance metrics employed for analyzing the diverse semantic segmentation models through deep learning is given in Table 15. From the set of research works, 63.3% of the works use OA, 48.3% of the contributions use F1 score, and 25% of the works consider recall and precision measures, respectively. mIoU metric is taken in 28.3% of the research works, 5% of the papers use Jaccard index, kappa coefficient, and dice coefficient the performance metric, confusion matrix, and PA are considered in 4% of the research works, respectively, and 23.3% of the contributions consider IoU measure. Furthermore, some of the additional measures are also taken for evaluating the efficiency of semantic segmentation, which are FWIoU, MCC, average accuracy, etc.

5.2. Best Performance Measures. The best performance measures obtained by diverse semantic segmentation models are depicted in Figure 12. From this comprehensive survey, Figure 8(a) represents contributions such as [32, 64] to get 97% as the highest accuracy rate than others. Secondly, the work in [50] obtains 94.49%, and the research works

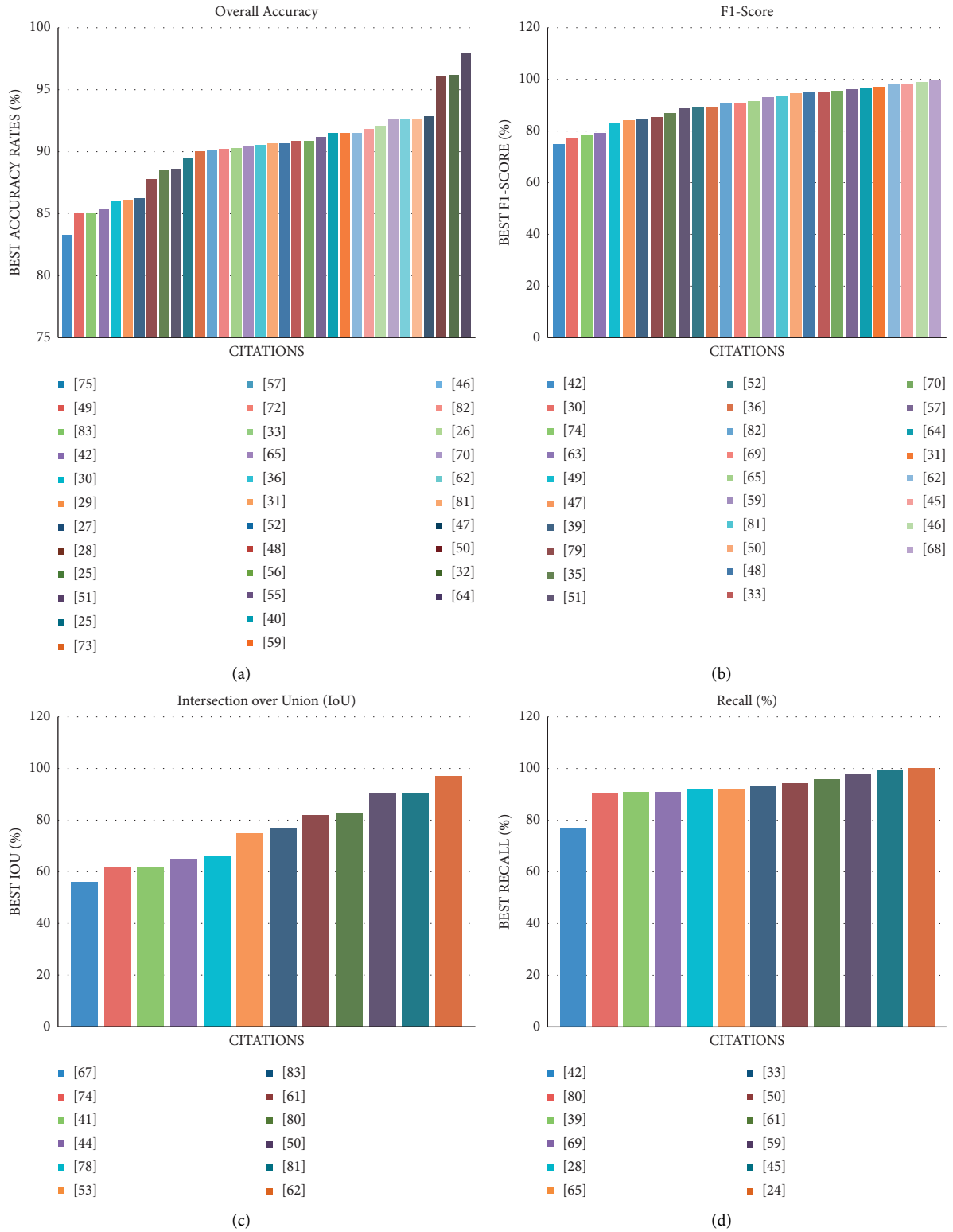

FIGURE 12: Best performance measures obtained by state-of-the-art semantic segmentation models: (a) best accuracy vs. citations; (b) best F1 score vs. citations, (c) best IoU vs. citations, and (d) best recall vs. citations.

Table 16: Performance measures used for analyzing the semantic segmentation models.

Citations	Overall accuracy (OA)	F1 score	Recall	Precision	Mean intersection over union (mIoU)	Jaccard index	Dice coefficient	Confusion matrix	Intersection over union (IoU)	Kappa coefficient	Pixel accuracy (PA)	Miscellaneous techniques
[24]	—	—	✓	✓	—	—	—	—	—	—	—	—
[25]	✓	—	—	—	—	—	—	—	—	—	—	—
[26]	✓	—	—	—	✓	—	—	—	—	—	—	—
[27]	✓	—	—	—	—	—	—	—	—	—	—	—
[28]	✓	—	✓	✓	—	—	—	—	—	—	—	—
[29]	✓	—	—	—	—	—	—	—	—	—	—	—
[30]	✓	✓	—	—	—	—	—	—	—	—	—	Average accuracy and ROC
[31]	✓	✓	—	—	—	—	—	✓	—	—	—	—
[32]	✓	—	—	—	—	—	—	—	—	—	—	Mean-class accuracy (AA)
[33]	✓	✓	✓	✓	—	—	—	✓	—	—	—	—
[34]	—	—	—	—	—	—	—	—	✓	—	—	—
[35]	✓	✓	—	—	—	—	—	—	—	—	✓	—
[36]	✓	✓	—	—	—	—	—	—	—	—	—	Average F1 score
[37]	—	—	—	—	✓	—	—	—	—	—	✓	—
[38]	—	—	✓	✓	✓	—	✓	—	—	—	✓	Mean accuracy and frequency weighted IoU
[39]	—	✓	✓	✓	—	—	—	—	—	—	—	Harmonic average of the precision and recall
[40]	✓	✓	—	—	✓	—	—	—	—	—	—	—
[41]	—	—	✓	✓	—	—	—	—	✓	—	—	—
[42]	✓	✓	✓	✓	—	✓	—	—	—	✓	—	—
[43]	—	—	—	—	—	—	—	—	—	—	—	Error
[44]	✓	—	—	—	✓	—	—	—	—	—	—	—
[45]	—	✓	✓	✓	—	✓	—					MCC, average processing time
[46]	✓	✓	—	—	—	—	—	—	—	—	—	—
[47]	✓	✓	—	—	—	—	—	—	—	✓	—	Average accuracy (AA)
[48]	✓	✓	—	—	—	—	—	—	—	—	—	Average F1 score
[49]	✓	✓	—	—	✓	—	—	—	—	—	—	—
[50]	✓	✓	✓	✓	—	—	—	—	✓	—	—	—
[51]	✓	✓	—	—	—	—	—	—	—	—	—	—
[52]	✓	✓	—	—	—	—	—	—	—	—	—	Mean F1 (mF1)
[53]	✓	—	—	—	—	—	✓	—	✓	—	—	—
[54]	—	—	—	—	✓	—	—	—	—	—	—	Time complexity
[55]	✓	—	—	—	✓	—	—	—	—	—	—	—
[56]	✓	—	—	—	—	—	—	—	✓	—	—	—
[57]	✓	✓	✓	✓	—	—	—	✓	—	—	—	—
[58]	—	—	—	—	—	—	—	—	—	—	✓	Training time amd prediction speed
[59]	✓	✓	✓	✓	—	—	—	—	—	—	—	—

Table 16: Continued.

Citations	Overall accuracy (OA)	F1 score	Recall	Precision	Mean intersection over union (mIoU)	Jaccard index	Dice coefficient	Confusion matrix	Intersection over union (IoU)	Kappa coefficient	Pixel accuracy (PA)	Miscellaneous techniques
[60]	—	—	—	—	✓	—	—	—	—	—	—	—
[61]	—	—	✓	✓	—	—	—	—	✓	—	—	—
[62]	✓	✓	—	—	—	—	—	—	✓	—	—	—
[63]	✓	✓	—	—	✓	—	—	—	—	—	—	—
[64]	✓	✓	—	—	✓	—	—	—	—	—	—	—
[65]	✓	✓	✓	✓	—	—	—	—	—	—	—	—
[66]	—	—	—	—	—	—	—	—	—	—	—	Weighted IoU
[67]	—	—	—	—	—	—	—	—	✓	—	—	Running time and execution time
[68]	—	✓	—	—	—	—	—	—	—	✓	—	Percentage correct classification and overall error rate
[69]	—	✓	✓	✓	✓	—	—	—	—	—	—	—
[70]	✓	✓	—	—	✓	—	—	—	—	—	—	—
[71]	—	—	—	—	✓	—	—	—	—	—	—	Matthews correlation coefficient (MCC)
[72]	✓	—	—	—	—	—	—	—	—	—	—	—
[73]	✓	—	—	—	—	✓	✓	—	—	—	—	—
[74]	—	✓	—	—	—	—	—	—	✓	—	—	—
[75]	✓	—	—	—	✓	—	—	—	—	—	—	—
[76]	—	—	—	—	—	—	—	—	—	—	—	Binary cross entropy (BCE)
[77]	—	—	—	—	✓	—	—	—	—	—	—	Binary accuracy, and boundary F1 score
[78]	✓	—	—	—	—	—	—	—	✓	—	—	Training time
[79]	—	✓	—	—	✓	—	—	✓	—	—	—	—
[80]	—	—	✓	✓	—	—	—	—	✓	—	—	Training and prediction time
[81]	✓	✓	—	—	—	—	—	—	✓	—	—	—
[82]	✓	✓	—	—	—	—	—	—	✓	—	—	—
[83]	✓	—	—	—	—	—	—	—	✓	—	—	Frequency weighted IoU (FWIoU)

such as [26, 43, 47, 56, 59, 81] attain 92.63% accuracy rate when compared with other works. The best performances for some of the metrics such as overall accuracy, F1 score, intersection over union, and recall were noted and tabulated as shown in Table 16.

5.3. Research Gaps and Challenges. In recent decades, several semantic segmentation approaches have been designed for different applications such as surveillance systems, traffic monitoring, and analysis on environmental changes. However, manual segmentation methods are time tedious and complex one. Thus, an automated semantic segmentation of aerial images is emerged as the recent hot topic [139]. On the contrary, the semantic segmentation of aerial images is a complex task due to several constraints such as demand for pixel-level accuracy, nonconventional data, and lack of training examples. Each object in the remote-sensing images specifies important information, which requires to be precisely categorized from the neighboring ones. Numerous works have been proposed for solving this problem, which has been focused on improving regularization and FCN such as object boundary details. More numbers of public datasets have been considered for evaluating the performance of the deep learning approaches. Here, infrared and colour satellite images have gained noteworthy performance that is more equivalent to image sets utilized in the portrait and scenic computer vision tasks. From the comprehensive review, the public datasets such as ISPRS datasets get more importance that has guaranteed the implementation of deep learning approaches for facilitating the semantic segmentation [140]. Though, the semantic segmentation on different data or imaging modality and analysis metrics make evaluation complex. Moreover, handling of different modality of remote-sensing images such as UAV, hyperspectral images, and infrared and RGB images are complex to process. It results in lack of accuracy to estimate the nonconventional data.

Sometimes, a large volume of data and a lack of training examples pose complexities in aerial imaging applications. Conversely, it is much more challenging due to the nonconventional data sources such as LiDAR, hyperspectral images, and synthetic aperture radar images [141]. When the deep learning techniques are utilized for processing the nonconventional remote-sensing datasets with labels, it creates complexities. These deep learning methods suffer from the lack of training dataset. Any deep learning model may need a huge set of training images due to the number of classes and complications of the problem [142]. Moreover, the utilization of deep learning is more complicated while considering the expensive and additional remote-sensing data collection [143]. Thus, different augmentation approaches are mostly employed for increasing the variation and number of the dataset. Consequently, the most common datasets called "ISPRS's 2D labeling dataset and IEEE'S GRRS dataset" have been attempted for addressing the data inefficiency through offering the very high-resolution remote-sensing images gathered from UAVs [141].

An additional limitation of deep learning-based semantic segmentation is the necessity of a high number of label dataset, which generally requires manual annotation. This issue has also considerably been solved through public datasets through offering the annotations [142]. However, it is still tedious while taking the own or manual datasets. Existing research works have utilized conventional approaches for producing the annotations. Similarly, the label dataset can be created with the feature of pretrained models. From the meta-analysis results, the deep learning provides enhanced efficiency and shows the superior performance when compared to conventional approaches [143]. Many challenges of deep learning-adopted techniques have been solved and reduced in recent decades, which have to increase the performance. The future research areas in the semantic segmentation of aerial images can integrate the well-known deep learning models with hybrid or new variant metaheuristic approaches. As the deep learning-based semantic segmentation models have emerged their future prospects, it has to create a new future scope on different applications using intelligent algorithms for increasing the accuracy rate [144]. In the future, it has to solve the nonconventional data and labeling problems while preparing a new datasets. Thus, this research helps the researchers to understand the semantic segmentation model with several other possibilities for coming up with new future research perspectives.

6. Conclusion

This study has presented a comprehensive review on conventional semantic segmentation models through deep learning approaches. For this purpose, a set of research works has been taken from recent years. This study has given the information regarding different machine learning or deep learning techniques used, simulation tools, performance metrics, features and challenges of conventional semantic segmentation models, different imaging modalities, and the datasets utilized. Finally, the research gaps and limitations were analyzed for exploring a future research perspective of semantic segmentation systems. On the whole, this study has offered the detailed information on semantic segmentation models, which are helpful for assisting the researchers to present a semantic segmentation model in the upcoming years.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the help of the Vellore Institute of Technology, Vellore, India, for giving excellent assets to this work. Also, the authors would like to thank the individual copyright holders for consent conceded to incorporate referred figures in this work.

References

[1] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 266–270, 2019.

[2] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6054–6068, 2017.

[3] X. Zhang, W. Ma, C. Li, J. Wu, X. Tang, and L. Jiao, "Fully convolutional network-based ensemble method for road extraction from aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 10, pp. 1777–1781, 2020.

[4] R. Manickam, S. Kumar Rajan, C. Subramanian, A. Xavi, G. J. Eanoch, and H. R. Yesudhas, "Person identification with aerial imaginary using SegNet based semantic segmentation," *Earth Science Informatics*, vol. 13, no. 4, pp. 1293–1304, 2020.

[5] B. Du, Z. Zhao, X. Hu et al., "Landslide susceptibility prediction based on image semantic segmentation," *Computers & Geosciences*, vol. 155, 2021.

[6] G. Can, D. Mantegazza, G. Abbate, S. Chappuis, and A. Giusti, "Semantic segmentation on Swiss3DCities: A benchmark study on aerial photogrammetric 3D pointcloud dataset," *Pattern Recognition Letters*, vol. 150, pp. 108–114, 2021.

[7] G. Mandlburger, M. Kölle, H. Nübel, and U. Soergel, "BathyNet: A deep neural network for water depth mapping from multispectral aerial images," *Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, vol. 89, pp. 71–89, 2021.

[8] Bo Li, C. Chen, S. Dong, and J. Qiao, "Transmission line detection in aerial images: An instance segmentation approach based on multitask neural networks," *Signal Processing: Image Communication*, vol. 96, August 2021.

[9] P. K. R. Maddikunta, S. Hakak, M. Alazab et al., "Unmanned aerial vehicles in smart agriculture: Applications, requirements, and challenges," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17608–17619, 2021.

[10] Q. Gao and X. Shen, "ThickSeg: Efficient semantic segmentation of large-scale 3D point clouds using multi-layer projection," *Image and Vision Computing*, vol. 108, Article ID 104161, 2021.

[11] A. S. Edun, K. Perry, J. B. Harley, and C. Deline, "Unsupervised azimuth estimation of solar arrays in low-resolution satellite imagery through semantic segmentation and Hough transform," *Applied Energy*, vol. 298, Article ID 117273, 2021.

[12] P. Wei, D. Chai, T. Lin, C. Tang, M. Du, and J. Huang, "Large-scale rice mapping under different years based on time-series Sentinel-1 images using deep semantic segmentation model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 174, pp. 198–214, 2021.

[13] R. Ch, G. Srivastava, T. R. Gadekallu, P. K. R. Maddikunta, and S. Bhattacharya, "Security and privacy of UAV data using blockchain technology," *Journal of Information Security and Applications*, vol. 55, Article ID 102670, 2020.

[14] G. Bhattacharjee and S. K. Pujari, "Semantic segmentation of aerial images survey," *International Journal of Applied Information Systems (IJAIS)*, vol. 12, no. No. 5, pp. 28–34, 2017.

[15] H. Xiu, P. Vinayaraj, K.-S. Kim, R. Nakamura, and W. Yan, "3D semantic segmentation for high-resolution aerial survey derived point clouds using deep learning (demonstration)," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 588–591, Washington, Seattle, November 2018.

[16] T. P. Singh, R. R. Singh, Himanshu, A. Mishra, and N. Sharma, "Semantic segmentation of satellite images: A survey," *International Research Journal of Engineering and Technology (IRJET)*, vol. 07, no. Issue. 12, 2020.

[17] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, 2021 [18] Juhong Wang, Bin Liu & Kun Xu "Semantic segmentation of high-resolution images," Science China Information Sciences, vol. 60, 2017, Article ID 114417.

[18] M. Chouai, M. Merah, and M. Mimi, "Correction to: CH-Net: deep adversarial autoencoders for semantic segmentation in X-ray images of cabin baggage screening at airports," *Journal of Transportation Security*, vol. 13, no. 1-2, p. 91, 2020.

[19] J. Wang, B. Liu, and K. Xu, "Semantic segmentation of high-resolution images," *Science China Information Sciences*, vol. 60, Article ID 123101, 2017.

[20] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, 2015.

[21] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, T. R. Gadekallu, and G. Srivastava, "SP2F: A secured privacy-preserving framework for smart agricultural Unmanned Aerial Vehicles," *Computer Networks*, vol. 187, p. 1, Article ID 07819, 2021.

[22] N. M. Balamurugan, S. Mohan, M. Adimoolam, A. John, and W. Wang, "DOA tracking for seamless connectivity in beamformed IoT-based drones," *Computer Standards & Interfaces*, vol. 79, Article ID 103564, 2022.

[23] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2021.

[24] S. Saito, R. Arai, and Y. Aoki, "Seamline determination based on semantic segmentation for aerial image mosaicking," *IEEE Access*, vol. 3, pp. 2847–2856, 2015.

[25] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation OF aerial images with an ensemble OF CNNS," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. III-3, pp. 473–480, 2016.

[26] A. Holliday, M. Barekatain, J. Laurmaa, C. Kandaswamy, and H. Prendinger, "Speedup of deep learning ensembles for semantic segmentation using a model compression technique," *Computer Vision and Image Understanding*, vol. 164, pp. 16–26, 2017.

[27] K. Chen, K. Fu, M. Yan, X. Gao, X. Sun, and X. Wei, "Semantic segmentation of aerial images with shuffling convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 173–177, Feb2018.

[28] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3252–3261, 2018.

[29] K. Chen, M. Weinmann, X. Sun et al., "Semantic segmentation OF aerial imagery via multi-scale shuffling convolutional neural networks with deep supervision," *ISPRS Annals*

of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. IV-1, pp. 29–36, 2018.

[30] M. Volpia and D. Tuia, "Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 144, pp. 48–60, 2018.

[31] Y. Sun, X. Zhang, Q. Xin, and J. Huang, "Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 143, pp. 3–14, 2018.

[32] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, no. Part A, pp. 60–77, 2018.

[33] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[34] D. M. Vo and S.-W. Lee, "Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions," *Multimedia Tools and Applications*, vol. 77, no. 14, pp. 18689–18707, 2018.

[35] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 2612–2626, 2019.

[36] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "High-resolution aerial images semantic segmentation using deep fully convolutional network with Channel attention mechanism," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3492–3507, 2019.

[37] Y. Li, B. Peng, L. He, K. Fan, and L. Tong, "Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2279–2287, 2019.

[38] S. M. Azimi, P. Fischer, M. Korner, and P. Reinartz, "Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 2920–2938, 2019.

[39] L. Wang, Z. Chen, D. Hua, and Z. Zheng, "Semantic segmentation of transmission lines and their accessories based on UAV-taken images," *IEEE Access*, vol. 7, pp. 80829–80839, 2019.

[40] Z. Cao, K. Fu, X. Lu et al., "End-to-End DSM fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 11, pp. 1766–1770, 2019.

[41] T. Nguyen, J. Wozencraft, C. J. Taylor et al., "MAVNet: An effective semantic segmentation micro-network for MAV-based tasks," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3908–3915, 2019.

[42] Z. Guo, G. Wu, X. Song et al., "Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery," *IEEE Access*, vol. 7, pp. 99381–99397, 2019.

[43] D. M. Igonin and Yu. V. Tiumentseva, "Comparative efficiency analysis for various neuroarchitectures for semantic segmentation of images in remote sensing applications," *Optical Memory & Neural Networks*, vol. 28, no. 4, pp. 306–320, 2019.

[44] M. Wu, C. Zhang, J. Liu, L. Zhou, and X. Li, "Towards accurate high resolution satellite image semantic segmentation," *IEEE Access*, vol. 7, pp. 55609–55619, 2019.

[45] M. K. Masouleh and R. Shah-Hosseini, "Development and evaluation of a deep learning model for real-time ground vehicle semantic segmentation from UAV-based thermal infrared imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 155, pp. 172–186, 2019.

[46] N. Audebert, A. Boulch, B. Le Saux, and S. Lefèvre, "Distance transform regression for spatially-aware deep semantic segmentation," *Computer Vision and Image Understanding*, vol. 189, Article ID 102809, 2019.

[47] F. Mohammadimanesha, B. Salehic, M. Mahdianparia, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 151, pp. 223–236, 2019.

[48] L. Ding, T. Tang, and L. Bruzzone, "Improving semantic segmentation of aerial images using patch-based attention," 2019, https://arxiv.org/abs/1911.08877.

[49] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with Channel attention and domain specific transfer learning," *Remote Sensing*, vol. 11, no. Issue. 1, p. 83, 2019.

[50] W. Liu, Y. Zhang, H. Fan, Y. Zou, and Z. Cui, "A new multi-channel deep convolutional neural network for semantic segmentation of remote sensing image," *IEEE Access*, vol. 8, pp. 131814–131825, 2020.

[51] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7557–7569, 2020.

[52] F. Wang, S. Piao, and J. Xie, "CSE-HRNet: A context and semantic enhanced high-resolution network for semantic segmentation of aerial imagery," *IEEE Access*, vol. 8, pp. 182475–182489, 2020.

[53] G. Martinez-Soltero, A. Y. Alanis, N. Arana-Daniel, and C. Lopez-Franco, "Semantic segmentation for aerial mapping," *Mathematics*, vol. 8, no. 9, p. 1456, 2020.

[54] W. Jiawe, H. Xiong, H. Wang, and X. Nian, "ADSCNet: Asymmetric depthwise separable convolution for semantic segmentation in real-time," *Applied Intelligence*, vol. 50, no. issue. 12, pp. 1045–1056, 2020.

[55] J. Deng, Z. Zhong, H. Huang, Y. Lan, Y. Han, and Y. Zhang, "Lightweight semantic segmentation network for real-time weed mapping using unmanned aerial vehicles," *Applied Sciences*, vol. 10, no. 20, p. 7132, 2020.

[56] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[57] D. Chai, S. Newsam, and J. Huang, "Aerial image semantic segmentation using DCNN predicted distance maps," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 161, pp. 309–322, 2020.

[58] Z. Song, Z. Zhang, S. Yang, D. Ding, and J. Ning, "Identifying sunflower lodging based on image fusion and deep semantic segmentation with UAV remote sensing imaging," *Computers and Electronics in Agriculture*, vol. 179, Article ID 105812, 2020.

[59] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[60] L. Ye, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 108–119, 2020.

[61] L. C. L. Bianco, J. Beltrán, G. F. López, F. García, and A. Al-Kaff, "Joint semantic segmentation of road objects and lanes using Convolutional Neural Networks," *Robotics and Autonomous Systems*, vol. 133, Article ID 103623, 2020.

[62] Li Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 140–152, 2020.

[63] K. Zhang, Yu Han, J. Chen, Z. Zhang, and S. Wang, "Semantic segmentation for remote sensing based on RGB images and lidar data using model-agnostic meta-learning and partical Swarm optimization," *IFAC-papersOnLine*, vol. 53, no. Issue 5, pp. 397–402, 2020.

[64] W. Boonpook, Y. Tan, and Bo Xu, "Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry," *International Journal of Remote Sensing*, vol. 42, no. Issue 1, pp. 1–19, 2021.

[65] H. Yang, Bo Yu, J. Luo, and F. Chen, "Semantic segmentation of high spatial resolution images with deep neural networks," *GIScience and Remote Sensing*, vol. 56, no. Issue. 5, 2019.

[66] A. Mehra, N. Jain, and H. S. Srivastava, "A novel approach to use semantic segmentation based deep learning networks to classify multitemporal SAR data," *Geocarto International*, vol. 37, no. 1, pp. 163–178, 2020.

[67] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7178–7193, 2020.

[68] N. Venugopal, "Automatic semantic segmentation with DeepLab dilated learning network for change detection in remote sensing images," *Neural Processing Letters*, vol. 51, no. 3, pp. 2355–2377, 2020.

[69] S. Girisha, U. Verma, M. M. Manohara Pai, and R. M. Pai, "UVid-net: Enhanced semantic segmentation of UAV aerial videos by embedding temporal information," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4115–4127, 2021.

[70] J. Huang, X. Zhang, Y. Sun, and Q. Xin, "Attention-guided label refinement network for semantic segmentation of very high resolution aerial orthoimages," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4490–4503, 2021.

[71] A. Abdollahi, B. Pradhan, G. Sharma, K. N. A. Maulud, and A. Alamri, "Improving road semantic segmentation using generative adversarial network," *IEEE Access*, vol. 9, pp. 64381–64392, 2021.

[72] M. Alam, J.-F. Wang, C. Guangpei, L. Yunrong, and Y. Chen, "Convolutional neural network for the semantic segmentation of remote sensing images," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 200–215, 2021.

[73] A. Anagnostis, A. C. Tagarakis, D. Kateris et al., "Orchard mapping with deep learning semantic segmentation," *Sensors*, vol. 21, no. 11, 2021.

[74] P. Li, H. Zhang, Z. Guo et al., "Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning," *Advances in applied energy*, vol. 4, Article ID 100057, 2021.

[75] F. Wang, X. Luo, Q. Wang, and Lu Li, "Aerial-BiSeNet: A real-time semantic segmentation network for high resolution aerial imagery Author links open overlay," *Chinese Journal of Aeronautics*, vol. 34, no. 9, pp. 47–59, 2021, Available online.

[76] L. Vasquez-Espinoza, M. Castillo-Cara, and L. Orozco-Barbosa, "On the relevance of the metadata used in the semantic segmentation of indoor image spaces," *Expert Systems with Applications*, vol. 184, Article ID 115486, 2021.

[77] Z. Chen, D. Ting, R. Newbury, and C. Chen, "Semantic segmentation for partially occluded apple trees based on deep learning," *Computers and Electronics in Agriculture*, vol. 181, Article ID 105952, 2021.

[78] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "DAugNet: Unsupervised, multisource, multitarget, and lifelong domain adaptation for semantic segmentation of satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1067–1081, 2021.

[79] F. X. Li, X. Lyu, H. Gao et al., "Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images," *International Journal of Remote Sensing*, vol. 42, no. 9, pp. 3583–3610, 2021.

[80] Yi-Z. Jiang, "Semantic segmentation of remote sensing image based on convolutional neural network and mask generation," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[81] W. Liu, Y. Zhang, J. Yan, Y. Zou, and Z. Cui, "Semantic segmentation network of remote sensing images with dynamic loss fusion strategy," *IEEE Access*, vol. 9, pp. 70406–70418, 2021.

[82] Z. Zheng, X. Zhang, P. Xiao, and Z. Li, "Integrating gate and attention modules for high-resolution image semantic segmentation," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4530–4546, 2021.

[83] S. Ouyang and Y. Li, "Combining deep semantic segmentation network and graph convolutional neural network for semantic segmentation of remote sensing imagery," *Remote Sensing*, vol. 13, no. Issue. 1, p. 119, 2021.

[84] S. Mouakket and A. M. Bettayeb, "Investigating the factors influencing continuance usage intention of Learning management systems by university instructors: The Blackboard system case," *International Journal of Web Information Systems*, vol. 11, no. 4, pp. 491–509, 2015.

[85] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," arXiv preprint arXiv:1705.07204, 2017.

[86] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[87] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: A survey and review," arXiv preprint arXiv:1412.7584, 2014.

[88] D. Zhang, X. Chen, D. Wang, and J. Shi, "A survey on collaborative deep learning and privacy-preserving," in *Proceedings of the 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pp. 652–658, IEEE, Guangzhou, China, June 2018.

[89] L. Alzubaidi, J. Zhang, A. J. Humaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, no. 1, pp. 53–74, 2021.

[90] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[91] M. Shervin, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[92] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[93] R. Mottaghi, X. Chen, X. Liu et al., "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, Columbus, OH, USA, June 2014.

[94] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: Common objects in context," in *Proceedings of the European conference on computer vision*, Springer, Cham, pp. 740–755, 2014.

[95] M. Cordts, O. Mohamed, S. Ramos et al., "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, Las Vegas, NV, USA, June 2016.

[96] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, Honolulu, HI, USA, Julu 2017.

[97] Ce Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1972–1979, IEEE, Miami, FL, USA, August 2009.

[98] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proceedings of the 2009 IEEE 12th international conference on computer vision*, pp. 1–8, IEEE, Kyoto, Japan, 2009, September.

[99] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,"vol. 2, pp. 416–423, in *Proceedings of the Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, pp. 416–423, IEEE, Vancouver, BC, Canada, 2001, July.

[100] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3282–3289, IEEE, Providence, RI, USA, 2012, June.

[101] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[102] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proceedings of the 2011 International Conference on Computer Vision*, pp. 991–998, IEEE, Barcelona, Spain, November 2011.

[103] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1978, Columbus, OH, USA, June 2014.

[104] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, Las Vegas, NV, USA, June 2016.

[105] X. Shen, A. Hertzmann, J. Jia et al., "Automatic portrait segmentation for image stylization," *Computer Graphics Forum*, vol. 35, no. 2, pp. 93–102, 2016.

[106] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the European conference on computer vision*, Springer, Berlin, Heidelberg, pp. 746–760, 2012.

[107] J. Xiao, A. Owens, and A. Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in *Proceedings of the IEEE international conference on computer vision*, pp. 1625–1632, Sydney, NSW, Australia, December 2013.

[108] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, Boston, MA, USA, June 2015.

[109] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *Proceedings of the 2011 IEEE international conference on robotics and automation*, pp. 1817–1824, IEEE, Shanghai, China, May 2011.

[110] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, Honolulu, HI, USA, July 2017.

[111] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," arXiv preprint arXiv:1702.01105, 2017.

[112] A. X. Chang, T. Funkhouser, L. Guibas et al., "Shapenet: An information-rich 3d model repository," arXiv preprint arXiv:1512.03012, 2015.

[113] M. De Deuge, A. Quadros, C. Hung, and D. Bertrand, "Unsupervised feature learning for classification of outdoor 3d scans," *Australasian Conference on Robitics and Automation*, vol. 2, p. 1, 2013.

[114] W. Song, N. Zheng, R. Zheng, X. Zhao, and A. Wang, "Digital image semantic segmentation algorithms: A survey," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 10, no. 1, pp. 196–211, 2019.

[115] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, "Weakly-supervised dual clustering for image semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2075–2082, Portland, OR, USA, June 2013.

[116] R. Guo, X. P. Yang, and J. Wang, "The optimization analysis of the image segmentation and denoising based on the improved FCM clustering algorithm [J]," *CAAI Transactions on Intelligent Systems*, vol. 11, no. 6, pp. 227–233, 2016.

[117] X. X. Zhang, "Research on image semantic segmentation based on probability Graph Model [D]," Dissertation, Xiamen University, Xiamen, Fujian, China, 2014.

[118] C. F. Zhang, "Image semantic segmentation based on conditional random _led [J]," *Computer CD software and applications*, no. 9, pp. 21–23, 2012.

[119] X. M Zuo, Z. Zhao, and T. T Gou, "RGB-D image segmentation method based on interactive conditional random _elds [J]," *Computer applications and software*, vol. 34, no. 3, pp. 174–180, 2017.

[120] L. Mao and M. Xie, "Image semantic segmentation based on higher-order CRF model [J]," *Application research of compute*, vol. 30, no. 11, pp. 3514–3517, 2013.

[121] L. J. Wang, Y. Q. Zhong, and H. Guo, "Improved image segmentation algorithm based on order conditional random _eld model [J]," *Computer Engineering*, vol. 42, no. 6, pp. 241–246, 2016.

[122] H. X. Chen, "Semantic segmentation based on convolutional neural networks [D]," Dissertation, Zhejiang University, Hangzhou, China, 2016.

[123] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, Santiago, Chile, December 2015.

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

[125] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1925–1934, Honolulu, HI, USA, July 2017.

[126] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," arXiv preprint arXiv:1506.04924, 2015.

[127] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[128] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 11–19, Honolulu, HI, USA, July 2017.

[129] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," arXiv preprint arXiv:1606.02147, 2016.

[130] Md A. Islam, M. Rochan, N. D. B. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3751–3759, Honolulu, HI, USA, July 2017.

[131] N. Souly, C. Spampinato, and M. Shah, "Semi and weakly supervised semantic segmentation using generative adversarial network," arXiv preprint arXiv:1703.09695, 2017.

[132] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, Venice, Italy, October 2017.

[133] Yi Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2359–2367, Honolulu, HI, USA, July 2017.

[134] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3150–3158, Las Vegas, NV, USA, June 2016.

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

[136] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018.

[137] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*, Springer Science & Business Media, Boston, MA, 2011.

[138] https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c [Multi-Class Metrics Made Simple, Part III: the Kappa Score (aka Cohen's Kappa Coefficient)].

[139] M. Zhang, Y. Zhou, J. Zhao, Y. Man, B. Liu, and R. Yao, "A survey of semi- and weakly supervised semantic segmentation of images," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4259–4288, 2020.

[140] Q. Geng, Z. Zhou, and X. Cao, "Survey of recent progress in semantic image segmentation with CNNs," *Science China Information Sciences*, vol. 61, Article ID 051101, 2018.

[141] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017.

[142] U. Srinivasan, S. Pfeiffer, S. Nepal, M. Lee, L. Gu, and S. Barrass, "A survey of MPEG-1 audio, video and semantic analysis techniques," *Multimedia Tools and Applications*, vol. 27, no. 1, pp. 105–141, 2005.

[143] S. I. Anishchenko and M. V. Petrushan, "Optimal feature space for semantic image segmentation," *Pattern Recognition and Image Analysis*, vol. 24, no. 4, pp. 502–505, 2014.

[144] Q. Ning, J. Zhu, and C. Chen, "Very fast semantic image segmentation using hierarchical dilation and feature refining," *Cognitive Computation*, vol. 10, no. 1, pp. 62–72, 2018.