remote sensing

MDPI

Article

IoT Enabled Deep Learning Based Framework for Multiple Object Detection in Remote Sensing Images

Imran Ahmed [1], Misbah Ahmad [2], Abdellah Chehri [3], Mohammad Mehedi Hassan [4] and Gwanggil Jeon [5,*]

[1] School of Computing and Information Science, Anglia Ruskin University, Cambridge CB1 1PT, UK
[2] Center of Excellence in IT, Institute of Management Sciences, Peshawar 25000, Pakistan
[3] Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, ON K7K 7B4, Canada
[4] Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[5] Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, Korea
* Correspondence: gjeon@inu.ac.kr

Abstract: Advanced collaborative and communication technologies play a significant role in intelligent services and applications, including artificial intelligence, Internet of Things (IoT), remote sensing, robotics, future generation wireless, and aerial access networks. These technologies improve connectivity, energy efficiency, and quality of services of various smart city applications, particularly in transportation, monitoring, healthcare, public services, and surveillance. A large amount of data can be obtained by IoT systems and then examined by deep learning methods for various applications, e.g., object detection or recognition. However, it is a challenging and complex task in smart remote monitoring applications (aerial and drone). Nevertheless, it has gained special consideration in recent years and has performed a pivotal role in different control and monitoring applications. This article presents an IoT-enabled smart surveillance solution for multiple object detection through segmentation. In particular, we aim to provide the concept of collaborative drones, deep learning, and IoT for improving surveillance applications in smart cities. We present an artificial intelligence-based system using the deep learning based segmentation model PSPNet (Pyramid Scene Parsing Network) for segmenting multiple objects. We used an aerial drone data set, implemented data augmentation techniques, and leveraged deep transfer learning to boost the system's performance. We investigate and analyze the performance of the segmentation paradigm with different CNN (Convolution Neural Network) based architectures. The experimental results illustrate that data augmentation enhances the system's performance by producing good accuracy results of multiple object segmentation. The accuracy of the developed system is 92% with VGG-16 (Visual Geometry Group), 93% with ResNet-50 (Residual Neural Network), and 95% with MobileNet.

Keywords: artificial intelligence; IoT; remote sensing; aerial computing; PSPNet

Citation: Ahmed, I.; Ahmad, M.; Chehri, A.; Hassan, M.M.; Jeon, G. IoT Enabled Deep Learning Based Framework for Multiple Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 4107. https://doi.org/10.3390/rs14164107

Academic Editor: Pedro Melo-Pinto

Received: 19 June 2022
Accepted: 17 August 2022
Published: 22 August 2022

1. Introduction

Today people have been witnessing rapid evolution in almost every field of life because of the emerging trends of collaborative and communication technologies. The rapid advancement, deployment, and integration of IoT, artificial intelligence, drone, robotics, cloud or edge computing, big data analytics, future communication networks, and aerial access networks have also been accepted as key enablers for different smart city applications. Due to aerial access networks, which can be established depending upon either low-altitude or high altitude platforms, when combined with IoT, satellite, and physical infrastructures, enable a complete access network with global coverage and various quality-of-service provisioning. Figure 1 shows the collaboration of IoT, drone cameras, intelligent communication, and aerial access technology that can lead us toward high standards and

quality of life by providing smart public safety disaster management, smart agriculture, environmental aspects, and services. These services may assist in traffic control by providing congestion-free routes to the users, unusual even detection, fire monitoring, energy efficiency, network connectivity, monitoring, and most importantly, security and surveillance applications. Recently, improvements in advanced aerial devices and technologies, such as the growth of small, smart, and economical communication networks, satellites, and the large availability of aerial drones and unmanned vehicles, have transformed numerous security and surveillance applications. Furthermore, when connected with future communication networks and satellites, these devices also enhance real-time assistance for smart cities, as depicted in Figure 1.

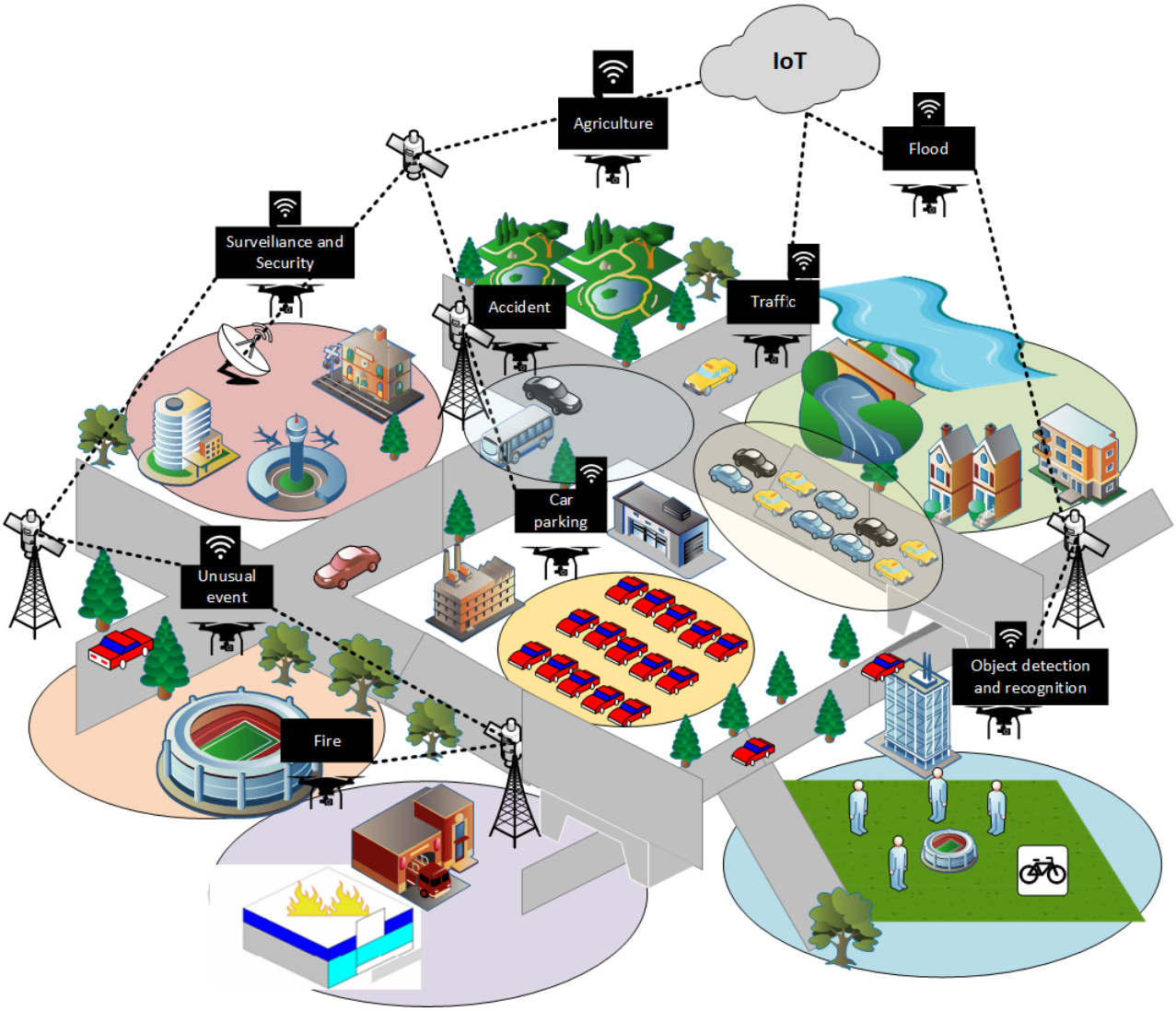

Figure 1. Illustration of collaborative IoT and aerial drone in smart city applications.

Aerial drone cameras can efficiently assist in various control and monitoring applications and also help in the gathering of valuable information [1]. The information can present excellent possibilities for various applications in smart cities, such as urban management, land change monitoring, traffic monitoring, and most importantly, surveillance. These applications include object detection [2], classification [3,4], tracking [5,6], and segmentation [6,7] tasks, which utilized the information obtained remotely (at a distance from height) by using high-resolution visual devices, such as drone cameras or sensors. However, identification, classification, tracking, and segmentation of multiple objects using aerial drone data is also a challenging task [8], which involves multiple factors, such as the

height of the installed and used camera, the appearance of objects, background varieties, and environmental requirements. Researchers have introduced various techniques for identifying multiple objects and regions of interest in aerial data sets in literature. The proposed and developed systems are mainly categorized into two classes, such as conventional feature-based and high-level deep learning-based methods [9–11]. Conventional methods utilized multiple features, including shape, color, edge, textual, background, foreground subtraction, and extraction, to classify different objects. These methods obtain excellent results for various remote surveillance applications; however, they normally need a huge amount of data and are not able to recognize or segment different objects whose properties or characteristics do not exist in the training data set. With the evolution of advanced deep learning, various convolution neural network based models are used by researchers to segment different objects and regions of interest in aerial data sets. These advanced and refined methods significantly enhanced the overall and generalization accuracy of segmentation methods for different applications [12–19].

This article investigates the collaboration performance between the Internet of things (IoT) and drones. This might be essential in providing timely and important wireless communication assistance in various public safety applications, as shown in Figure 1. It could help to explore areas that are difficult to reach and that provide immediate coverage and connectivity for the services' recovery right after a disaster by improving overflow capacity. Thus, to deliver such vital facilities for public safety, the collaboration between IoT and drones can support public safety requirements in the case of disasters, such as real-time monitoring, real-time analytics, and enhanced decision making to support smart city applications. Successful results of the deep learning designs inspired us to introduce an IoT-enabled solution for smart surveillance and monitoring in smart city applications. The system is based on artificial intelligence that uses a deep learning model, PSPNet [20], for object segmentation in the aerial drone data set. The model utilizes a pyramid parsing module that uses global context information for various region-based context collections. The local and global traces collectively produce more stable final predictions. The PSPNet architecture needs the global information of the image in order to predict the local predictions; therefore, it delivers more reliable results on benchmark data sets, such as PASCAL VOC 2012 and cityscapes. Furthermore, the model performs better than FCN because pixel classifiers cannot obtain the context of the entire input image. The overall work performed in this article consists of the following steps: we firstly implement data augmentation techniques to increase the architecture's performance. Then, we apply transfer learning and train the pre-trained segmentation model on the aerial data set. In this work, we use a benchmark publically available data set named Aerial Semantic Segmentation Drone data set (https://www.tugraz.at/index.php?id=22387) (19 June 2022). Finally, we examine and compare the segmentation architecture with three different CNN classification designs, particularly VGG-16, ResNet-50, and MobileNet. The primary goals of the article are presented as follows:

- To introduce an IoT-enabled solution for smart surveillance applications in smart cities using an aerial drone.
- To apply artificial intelligence and develop a system based on the deep learning model for multiple object detection.
- To apply data augmentation techniques and deep transfer learning to increase and enhance the performance of an aerial drone surveillance system.
- To explore training and testing of the deep learning architecture with different CNN classifies using aerial drone images.
- To investigate and analyze the results of the segmentation architecture with different classification models with aerial drone data in terms of efficiency.

The rest of the work performed in the article is categorized into the following parts: In Section 2, a brief summary of related work is presented that is applied to object detection in various aerial drone surveillance applications. In Section 3, we present an IoT-enabled system for intelligent surveillance, which is based on artificial intelligence. We also explain

the segmentation model and CNN classifiers applied for multiple object segmentation using an aerial data set. In Section 4, the summary of the data set utilized for the experiments is briefly discussed. Furthermore, in this part, we also explain the output and performance results. Section 5 provides a discussion about the proposed method. Lastly, in Section 6, we review the given work with viable future trends.

2. Related Work

Recent methods developed and introduced for multiple object classification, detection, and segmentation are classified into conventional hand-crafted features or machine learning and advanced deep learning methods. In conventional methods, researchers used different template based methods for identifying a particular object or region, such as roads with simple visual appearance [21]. Authors in [22] applied gray-scale images and morphological operations for the detection of buildings. Stankov et al. [23] offered unsupervised and supervised machine learning with color-based feature designs for segmentation of roofs in aerial data. In [24], authors employed geometric information to identify objects; it principally encodes prior information by applying parametric generic or particular shape patterns. Moreover, authors in [25] introduced a segmentation method to obtain geographical data, such as the relationship between an object's appearance and spatial contexts. Ref. [26] employed a mean shift algorithm for object segmentation in an aerial data set. These techniques overcome the constraints of traditional pixel based classification systems and have been employed for mapping landslide [27], land cover, and change detection [28]. Some studies focused on background subtraction methods including Gaussian mixture model [29] and machine learning techniques, Markov random fields, random forest, and logistic regression classifier [30,31], for detection of objects in aerial images.

Researchers have recently started utilizing CNN-based techniques for aerial drone applications with deep learning. Similarly, authors in [32] introduced a dynamic neural technique applying a finite state machine for the extraction of roads. Ref. [33] introduced a deep learning method for the identification of targets in aerial data. Some researchers also employed segmentation and deep learning-based detection and classification methods for aerial drone surveillance applications. In [34], scholars executed semantic segmentation paradigm on observation data of Earth. Garg et al. [35] presented an instance segmentation paradigm for aerial views. Segnet and U-Net based techniques are applied by [36], enabling semantic segmentation for high-resolution aerial data sets. Ref. [37] also introduced a segmentation method for detecting multiple objects.

It is concluded from the above discussion that researchers in recent years have done significant work. Various conventional features, machine learning [38], and deep learning-based systems are introduced [39] for many aerial drone applications [40]. Researchers also applied background subtraction and foreground extraction, such as the Gaussian mixture algorithm and the machine learning method. They used neural network architectures for different remote sensing applications, such as ship detection, object detection, road detection, and tree detection. Similarly, they also studied the state-of-the-art models for classification, detection, and segmentation models of multiple objects. Though to the best of our knowledge, they mostly worked on the detection of specifically targeted objects, identification, recognition, classification, and segmentation of particular objects, including cars, ships, other vehicles, buildings, roofs, roads, lands, green farms, and trees, etc. This work presented an IoT-enabled generic surveillance system for multiple object detection, which can assist in various real-time surveillance applications.

3. IoT-Enabled Leep Learning Based Solution for Object Detection Using Aerial Drone

This work presented an IoT-enabled, smart surveillance solution for smart city applications. In Figure 2, we have shown the details of the proposed system. It can be seen that aerial drone vehicles used for monitoring and surveillance are connected through IoT and future generation networks. The collected aerial drone transferred recorded data to the

monitoring and surveillance unit. The mentoring unit gathered the data and generated a true mask or labeled it for all recorded images. A publicly available benchmark data set has been used. The collected data set images are initially given to a pre-processing phase. This phase performs shuffling, image resizing, and normalization. In the pre-processing phase, data augmentation techniques are applied to develop variation in the data set and the system's performance. The collected image samples are split into train and test sets after the pre-processing phase and data augmentation phase.

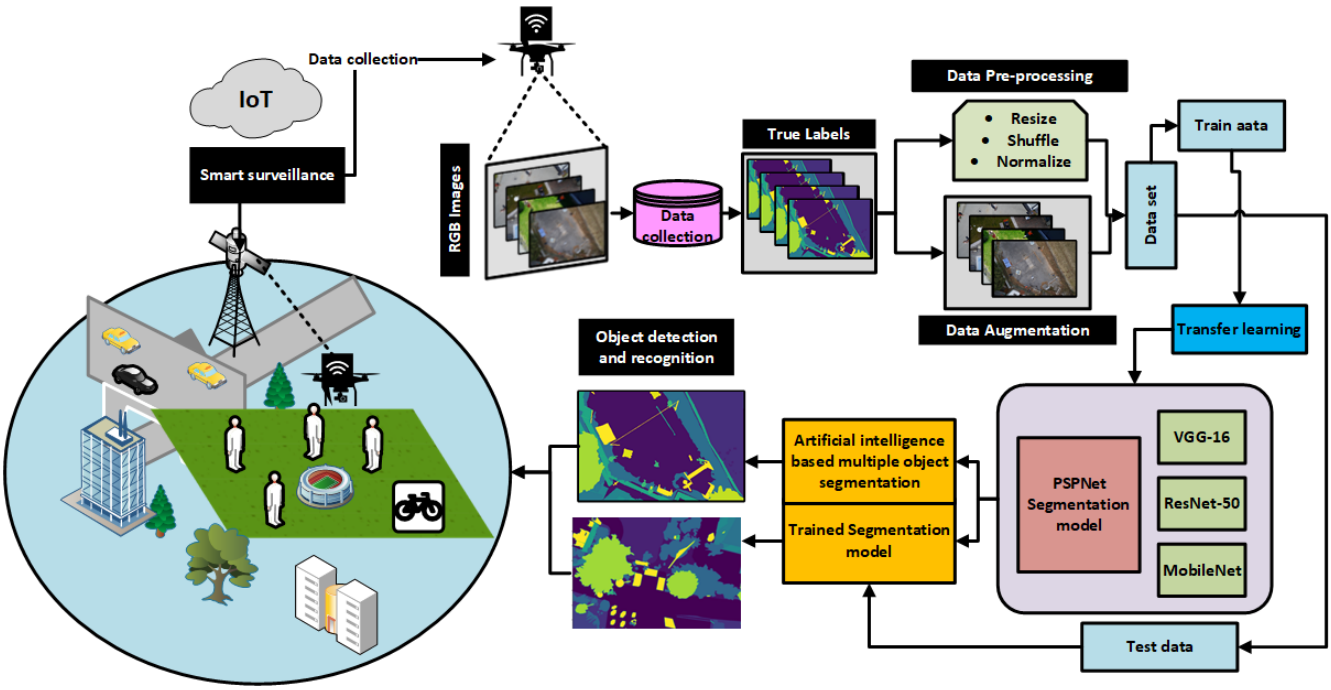

Figure 2. An IoT-enabled multiple object segmentation system based on deep learning using aerial drone data set. The overall system is based on a segmentation model with and CNN classifiers.

As detailed in Figure 2, the deep learning architecture is used for the segmentation model of multiple objects, which is based on deep learning. We applied the PSPNet [20], a deep learning model for object segmentation basically developed for the segmentation of multiple objects. As we know, mostly semantic segmentation paradigms mainly consist of two sections, i.e., an encoder and a decoder. We used an encoder to extract features from the image, whereas for the prediction class pixels, we used a decoder at the end. This work used three CNN classification architectures as base networks for object classification, particularly VGG-16, ResNet 50, and MobileNet. The convolution layers are transferred as the encoder part in the first step. Taking advantage of the deep transfer learning, as explained in Figure 2, we combine convolution layers with deconvolution layers as the decoder. The architecture raining is performed, and the segmentation model is trained with the CNN classification architectures. We utilize a test set to estimate the segmentation system's performance for three separate architectures. This training output is the segmentation map of a specific object in various test sample images. The segmentation results are assessed by applying a ground truth or true label mask. Finally, the usual evaluation metrics are applied to define the results of the developed system.

3.1. Pre-Processing and Data Augmentation

We performed the necessary pre-processing in order to keep the consistency of the segmentation model during training. Moreover, to maintain variation in image appearances, we performed image normalization, such as managing the brightness and contrast of an image. Lastly, the data set is shuffled into training and testing sample images. The overall data set utilized in this article mainly contains 400 training samples and 200 testing samples. It is

a comparatively small data set for the deep learning model training. Hence, we implement data augmentation techniques to preserve and obtain a moderate amount of sample images and to evade the over-fitting difficulty by ensuring the adequacy, fairness, and sufficiency in the variance and robustness of the architectures. For classification architectures, we resize the input images and the ground truth mask to $224 \times 224 \times 3$. We applied real-time data augmentation techniques to the data set, similar to [41]. As a result, images in the aerial data set are flipped, shifted, and rotated as described in Figure 3. This provides a considerably greater number of images for the training of models and architectures.

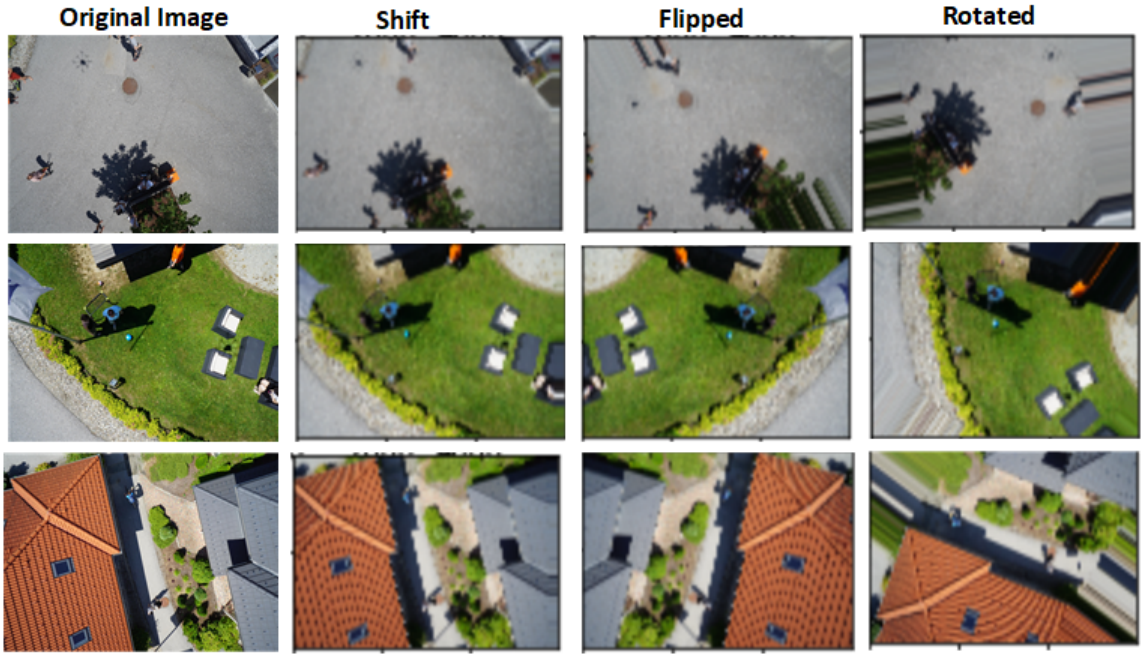

Figure 3. Results of data augmentation performed to increase variation in the data set.

3.2. CNN Based Classifiers as Base Architectures

As discussed earlier, different CNN architectures are applied for classification and feature extraction of the image later used by the encoder. The deep learning classification architectures mainly contain a collection of layers, especially convolutional, pooling, activation, ReLu, and fully connected layers. We used VGG-16 [42], a CNN model that is fine-tuned by adjusting some layers to overcome overfitting. It consists of 16 convolutional layers, having an input image shape of $224 \times 224 \times 3$, a fixed filter size of 3×3, and five max-pooling layers of size 2×2 in the entire network. There is a softmax layer at the head of the two fully connected layers. VGG-16 is a large network with nearly 138 million parameters. It is accumulating several convolutional layers to develop deep neural networks that enhance the capacity to learn hidden and deep features.

The second architecture used for the classification and extraction of image features is ResNet50 [43]. It has a 50-layer Residual network with approximately 26 million parameters. The residual network is a deep CNN model that Microsoft proposed in 2015. Rather than learning features, the residual network learns from residuals. To deliver data over layers, it uses the skip connections and combines nth input layer direct to some $(n + x)$th layer, allowing extra layers to be accumulated and building a deep architecture. It has forty-eight convolutional layers, a single max pooling, and one intermediate pooling layer.

The last classification architecture is MobileNet [44]. It is also an effective and interchangeable CNN architecture used in several applications. MobileNet introduces two distinct global hyper-parameters that are called width and resolution multiplier. These hyperparameters allow researchers to trade off efficiency or latency rates depending on their requirements. It applies depth-wise convolutions rather than conventional convolutions allowed in previous architectures to produce lighter designs. All convolutional layers are

composed of depth and point-wise convolutional layers. Evaluating these convolutions as separate layers, it contains 28 layers and 4.2 million parameters which can be reduced by tuning hyperparameters.

3.3. Semantic Segmentation Model

We applied the PSPNet model for the segmentation of objects in aerial images. It is developed by [20], one of the common well-accepted object and image segmentation models. The architecture of PSPNet predicts the local level predictions using the global context of the input image, providing better results on benchmark data sets. The model is better than FCN-based pixel classifiers as it could not obtain the meaning of the entire image. The overall architecture is presented in Figure 4 mainly divided into two main stages: encoder and decoder. As discussed earlier, that encoder extracts output features from the image, whereas the decoder is used to predict the pixel class at the end. The encoder part of PSPNet has comprised the CNN-based backbone architecture with dilated convolutions and the pyramid pooling module as described in Figure 4.

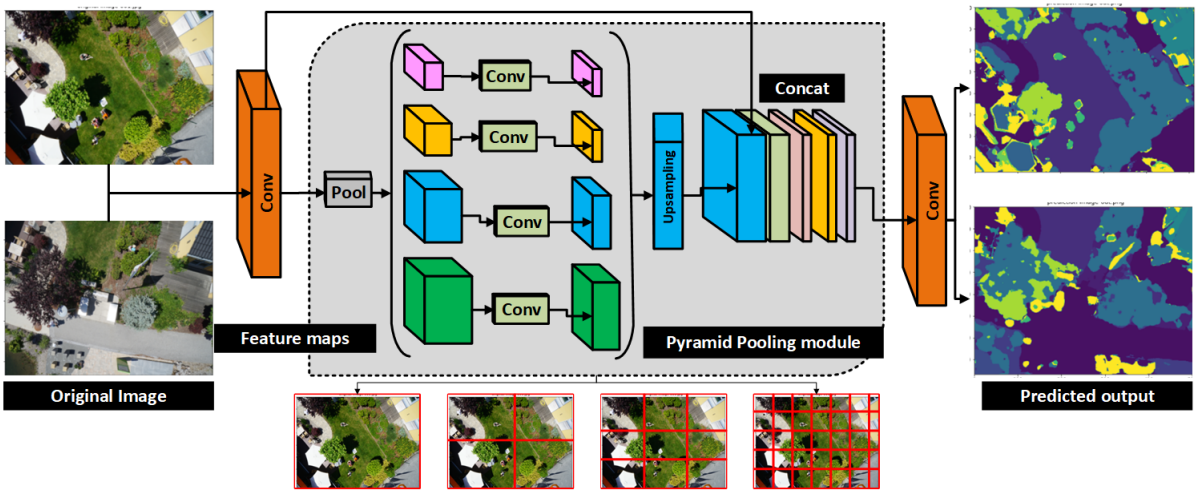

Figure 4. The general architecture of the deep learning-based segmentation model used for the segmentation of multiple objects using aerial drone data set.

The pyramid pooling module is the central element as it supports the segmentation model to obtain the global information of the input image. It also aids in classifying the image and the pixels using global information. As shown in Figure 5 extracted from the backbone CNN, the feature map is pooled at various sizes and later transferred within convolution layers. These feature maps are obtained using CNN architectures. At the final layer, we applied a binary cross entropy loss to train the classifier, given as,

$$L(\hat{y}, y) = -ylog\hat{y} - (1-y)log(1-\hat{y}) \quad (1)$$

$$\hat{y} = \frac{1}{1 + exp(-(\sum_j w_j x_j + b))} \quad (2)$$

where $y \epsilon \{0,1\}$ is the class, $x = (x1, \ldots, xj, \ldots, xn)$ indicates representative feature n, and w is used to correspond to x. The same size image feature as the original upsampling takes place on the pooled features. Lastly, the original feature maps are concatenated with upsampled features and transferred to the decoder part. This method combines the features of various scales, therefore adding the entire and complete context.

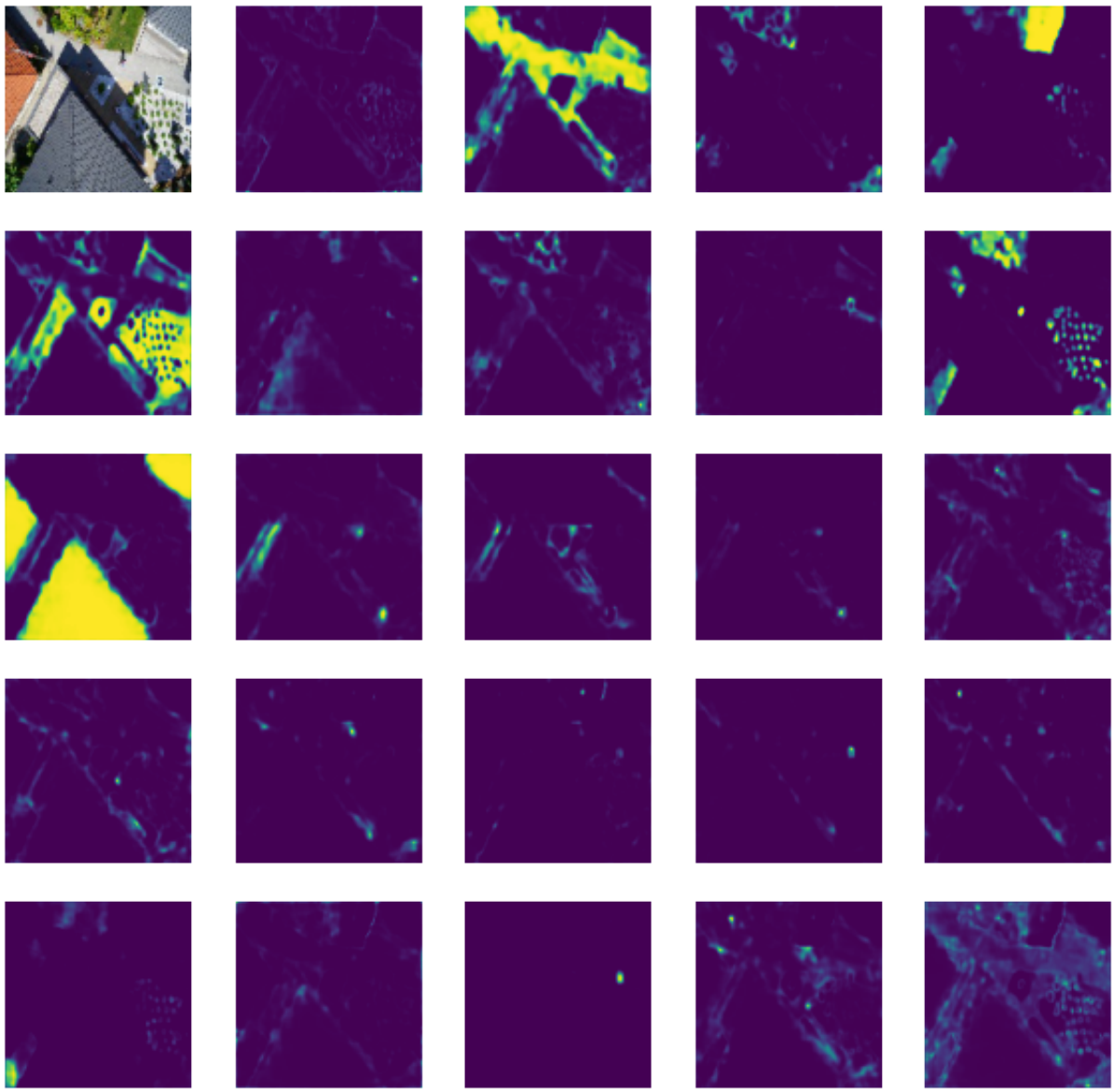

Figure 5. Predicted feature maps from the backbone CNN architecture; these feature maps are pooled at various sizes and later transferred within convolution layers.

To obtain accurate results and better segmentation performance, a convolutional conditional random field [45] is used for optimization, calculated as:

$$P(X = \hat{x}|\hat{I} = I) = \frac{1}{Z(I)} exp(-E(\hat{x}|I)) \tag{3}$$

In the above equation, if image I has P pixels and k classes, the predicted label graph for an image I is defined by a random field. Thus, a predicted label graph X is produced to increase the conditional probability, which can be modeled as a conditional random field under the Gibbs distribution illustrated in Equation (3). $Z(I)$ is the normalization function. The energy function [46] $E(\hat{x}|I)$ is given as:

$$E(\hat{x}|I) = \sum_{i \geqq N} \sigma_p(\hat{x}_i|I) + \sum_{i \geqq N} \sigma_p(\hat{x}_i, \hat{x}_j|I) \tag{4}$$

In the above equation, $\sigma_p(\hat{x}_i|I)$ is the unitary potential, whereas the $\sigma_p(\hat{x}_i, \hat{x}_j|I)$ is the pairwise potential. The first one only considers each pixel's category label, without any

other pixels' information. In contrast, the second considers the joint distribution between the pixel i and pixel j. Unlike the first one, it describes the interaction between the two pixels. When the output features of the image are extracted from an encoder, it is now transmitted to the decoder part. The decoder takes these features and transforms them into predictions by moving them within its layers. It is simply an extra network that uses features and makes predictions. It is worth noting here that the PSPNet model is not a perfect segmentation architecture; it is simply an encoder. Thus, we need decoders to implement PSPNet, commonly using a convolution layer followed by a bilinear upsampling. To achieve this, the FPN (Feature Pyramid Network) decoder is applied the same as as in U-Net [41]. Therefore, we combined the FPN decoder with the PSPNet encoder, which can obtain the small characteristics from the input image. Later, various upsampled stages of feature maps are concatenated with the primary feature maps. Finally, these feature maps are combined as global information at the end of the pyramid pooling module. Ultimately, it is accompanied by convolution layers to produce the final prediction maps. The PSPNet model used auxiliary loss during training. A weight of α 0.4 is joined to auxiliary loss to evaluate the losses.

4. Experimental and Performance Results

This section provided the training and testing observations of the developed system. We also discussed the segmentation or prediction results of the model for multiple object segmentation. Furthermore, in the end, we provided the evaluation results for aerial drone data.

4.1. Data Set

This data set named Aerial Semantic Segmentation Drone data set (https://www.tugraz.at/index.php?id=22387) (19 June 2022). principally focuses on understanding semantic information of urban areas/scenes in smart cities for developing the security of the independent aerial drone camera and landing operations. It includes images of more than twenty houses collected by using a high-resolution camera from a bird's eye view at the altitude of 5 to 30 m from the ground. The image size is 6000 $\times$ 4000 pixels (24Mega Pixels). The data set contains a total of 400 publicly available images for training and 200 private images for the testing set. Moreover, the data set provides pixel accurate annotation for similar training and testing samples. The data set complexity is limited to objects, usually twenty, including people, trees, land, grass, water, gravel, rocks, pools, paved areas, bicycles, cars, cats, dogs, doors, windows, roofs, walls, fence poles, fences, and obstructions.

4.2. Training and Testing

The proposed system has been executed using a python programming language (Pytourch library) with OpenCV 3.6. The implementation is the same as that presented in [20]. We utilize a conventional per-pixel Softmax Cross-Entropy Loss to train PSPNet. In addition, the learning rate is given by:

$$lr = baselr \times (1 - iter)^{power} \quad (5)$$

We defined the learning rate to be 0.01 and the power to be 0.9 for our experiments. The performance of the model can be increased by improving the training iteration. The weight decay and momentum of the model are set to 0.0001 and 0.9, respectively. Due to insufficient physical space on GPU, we initiated the "batch size" to 16 throughout training. The loss curve for training and testing of the PSPNet with CNN classification architectures is shown in Figures 6 and 7.

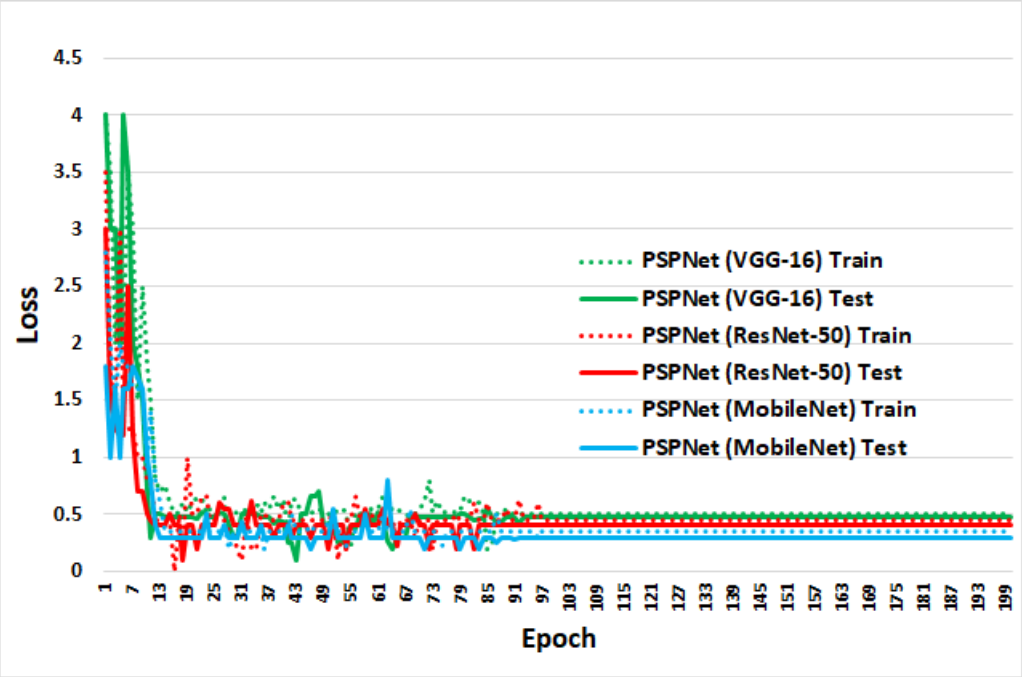

Figure 6. Train and Test Loss with CNN classification architectures.

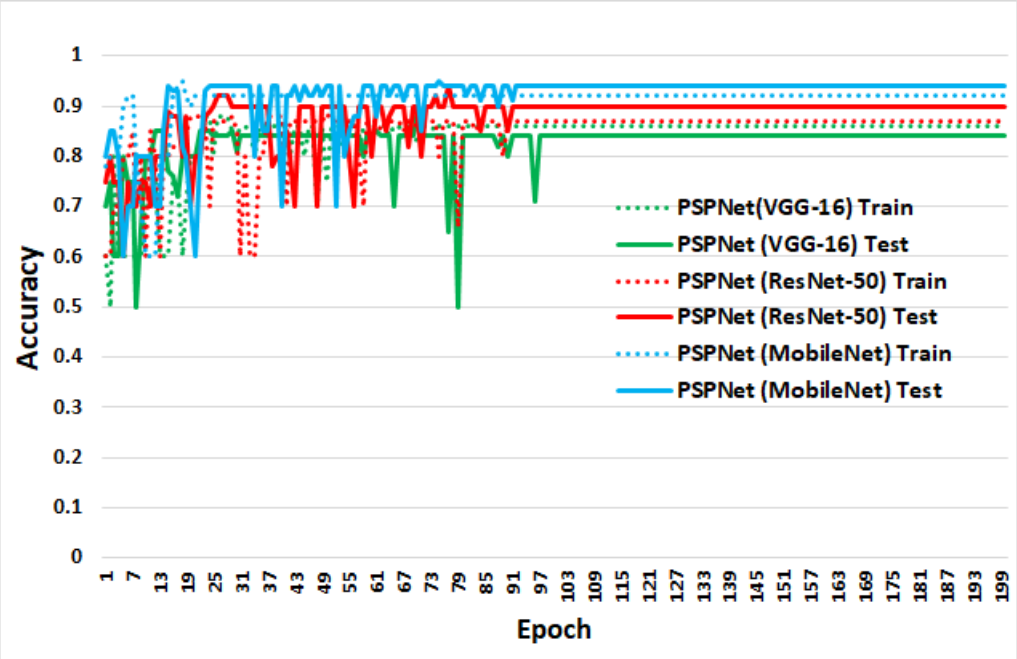

Figure 7. Train and Test Accuracy with CNN classification architectures.

4.3. Segmentation Results

The segmentation results of the above discussed system are presented in Figure 8. It can be observed that the system achieves good results for various objects in aerial drone data. We tested the developed system for different test images. In Figure 8, the first column visualizes the primary input images; the second and third column displays the true label masks and segmentation images, whereas the last two columns represent predicted mask and segmentation results. We have presented segmentation results for five testing images (row-wise). From Figure 8a, the segmentation architecture effectively segmented the road and ground in the image. In Figure 8b, the road area along with trees and multiple people

at different locations is accurately segmented. Similarly, the grass and road are accurately segmented in Figure 8c,d. The design also segmented the covered building efficiently, as observed in the last row of Figure 8e.

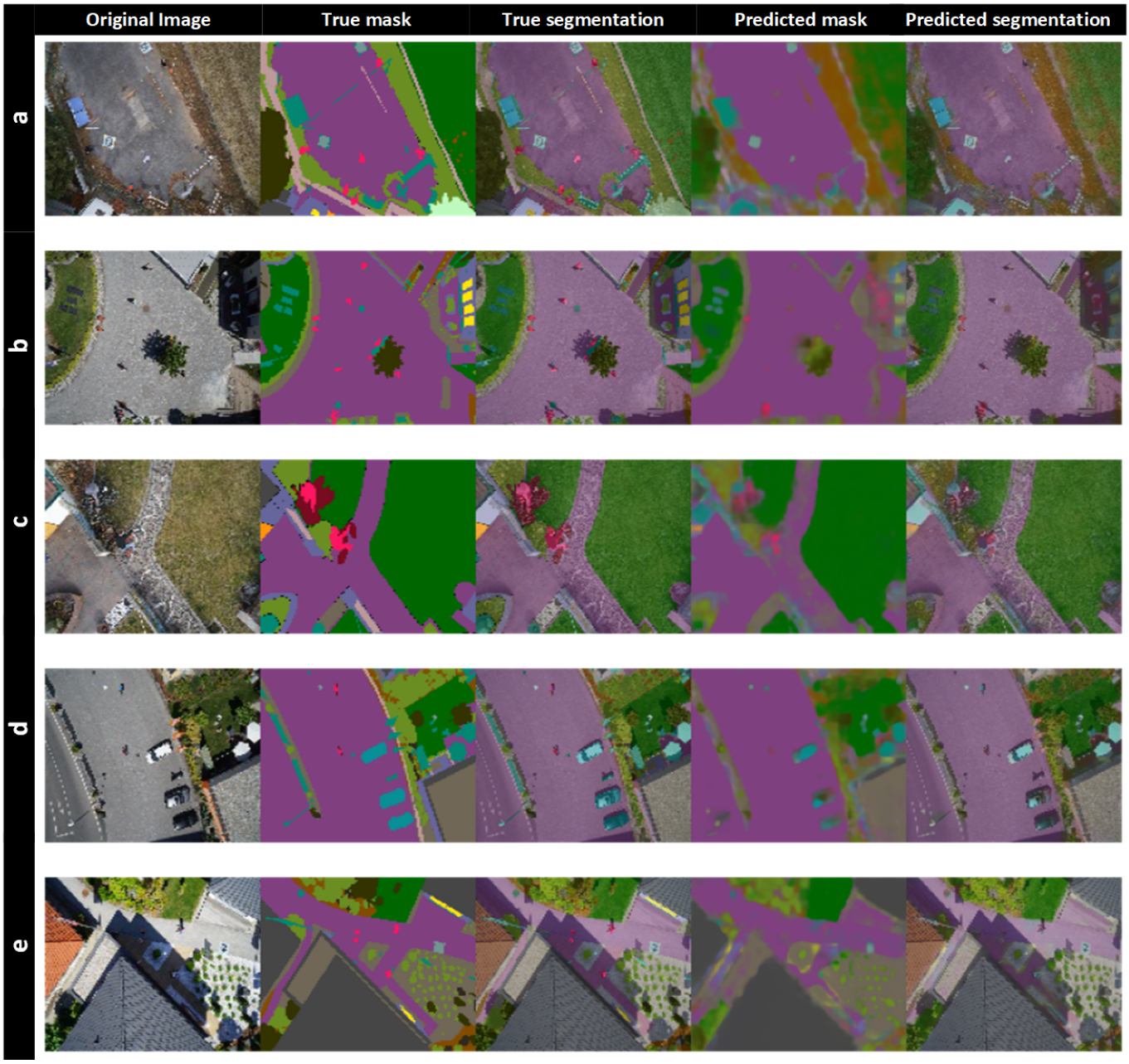

Figure 8. Multiple object segmentation results are shown row-wise from (**a**–**e**) for the model using aerial drone data. The first column represents original images, the second presents true mask, the third shows the true segmented mask, whereas the last two columns show the predicted mask and segmentation results.

4.4. Performance Evaluation

The developed system is evaluated in terms of accuracy, and different evaluation matrices are applied [12]. Every pixel of the test image is categorized into four classes according to prediction and ground truth results, which are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Using different parameters given in Figure 9, we estimated the precision, recall, and accuracy of the system as described in Figure 10. It can be recognized that the developed system produces excellent results

with various classification architectures. The precision rate of the segmentation paradigm with VGG-16 is 72%, with ResNet-50 is 75%, and with MobileNet is 80%. For VGG-16, ResNet-50, and MobileNet, the Recall rate is 86%, 90%, and 93%, respectively. The F1-Score ranges between 80% and 90% for all classification architectures. The system's accuracy is high with MobileNet, i.e., 96%, whereas for the other two, it is 93% and 95%, respectively.

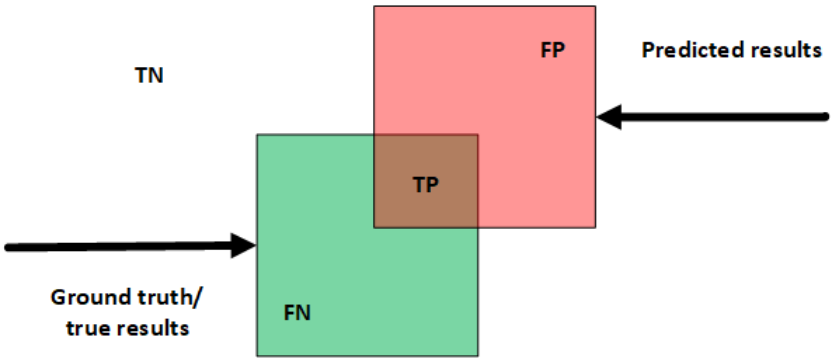

Figure 9. Relationship between TP, TN, FN, and FP.

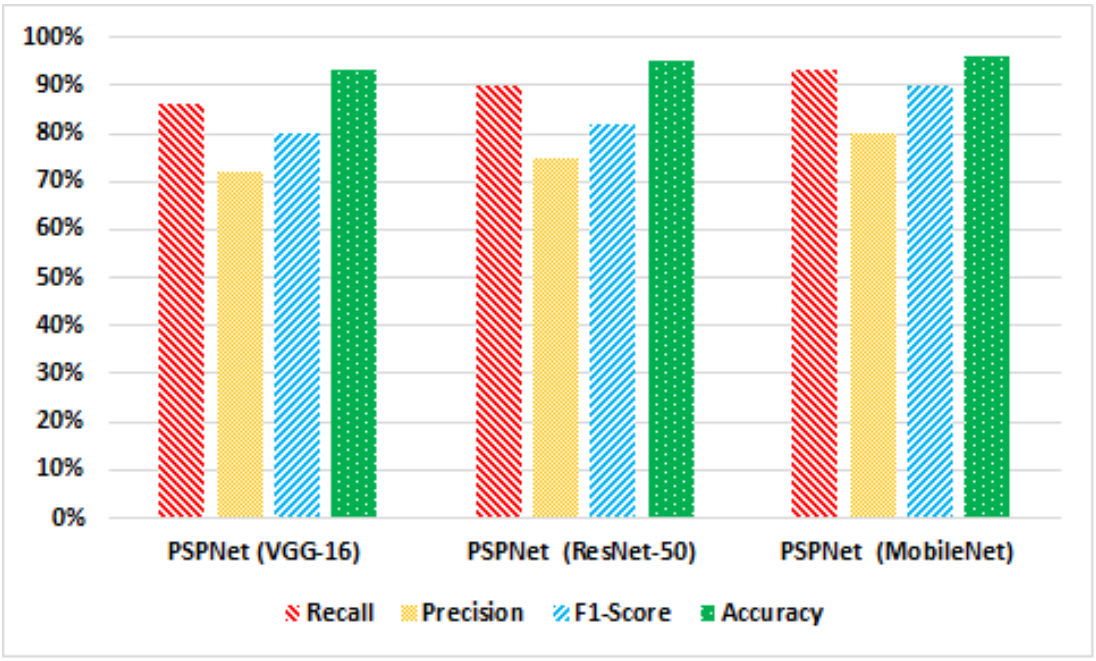

Figure 10. Evaluation results with different classification architectures.

To estimate the segmentation results, we used Pixel Accuracy (P_{acc}), defined as the accuracy of pixel-wise prediction and provided as:

$$P_{acc} = \frac{\sum_{i=0}^{K}(p_{ii})}{\sum_{i=0}^{K}\sum_{j=0}^{K}(p_{ij})} \tag{6}$$

The number of pixels in the testing image is defined as K, p_{ii} is the predicted pixels for class i, and p_{ij} the true label of object class. The segmentation results of the system have also been evaluated using Intersection over Union (IoU), and Mean-($mIoU$) IoU, mathematically provided as:

$$IoU = J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

It is generally defined as the intersection and union area ratio of the predicted segmentation map B to the ground truth label masks A for k object classes. Furthermore, ($mIoU$) is estimated as;

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{\sum_{j=0}^{k} FN + \sum_{j=0}^{k} FP - FN} \quad (8)$$

The segmentation performance of the model with different classification architectures is described in Table 1. The P_{acc} and $mIoU$ with VGG-16 is 80%, and 82%, with ResNet-50 80%, and 83%, whereas with MobileNet it gives excellent results with P_{acc} and $mIoU$, 83% and 85%, respectively.

Table 1. Comparison results P_{acc} and $mIoU$ of PSPNet and Un-Net with different base architectures.

S.No	Model Name	P_{acc}	$mIoU$
1	U-Net (VGG-16)	81%	82%
2	U-Net (ResNet-50)	80%	78%
3	U-Net (MobileNet)	83%	82%
4	PSPNet (VGG-16)	80%	82%
5	PSPNet (ResNet-50)	81%	83%
6	PSPNet (MobileNet)	83%	85%

5. Discussion

In the past few years, a large number of Earth-observing satellites have been deployed by space agencies. Researchers have proposed various machine learning and image processing-based methods that have played a useful role in environmental aerial drone surveillance applications. Deep learning also holds great promise in fulfilling the challenging needs of remote sensing applications [6]. It leverages the huge computing power of modern GPUs to perform human-like reasoning and extract compact features that embody input images' semantics. The interest of the remote sensing community toward deep learning techniques is developing fast, and many architectures have been presented in the last few years to address remote sensing problems, often with an outstanding performance.

This article provided an IoT-enabled deep learning-based system for object detection applications using aerial drone images. In Figure 8, we provided the visual results of the presented model with high accuracy results. The main focus of our study is various object classification and detection. This method is not only helpful for scene understanding by detecting and classifying various kinds of objects but also for several other applications worth mentioning, including fusion, segmentation, and change detection. The method also provided real-time surveillance applications as embedded with IoT. We also investigated the collaboration performance between the Internet of things (IoT) and drones, which might be important in delivering timely and important wireless communication services in different public safety applications. The accurate detection results help to explore those areas which are difficult to reach. The system can help to provide immediate coverage and connectivity for the service recovery right after a disaster by improving overflow capacity. Therefore, to provide such important facilities for public safety, the collaboration between IoT and drones can improve safety requirements in the case of disasters. It can help to monitor in real time and enhance decision making to sustain smart city applications.

6. Conclusions

This article presented an IoT-enabled smart surveillance solution for smart city applications. The surveillance system identified multiple object identification through segmentation. In particular, we provided a new concept of collaborative aerial drones, artificial intelligence, and IoT for developing surveillance applications in smart cities. We presented an artificial intelligence-based system using the deep learning based segmentation model PSPNet. We implement data augmentation techniques and leverage deep transfer learning to boost the accuracy of the system. We investigated and compared the system's accuracy

and performance with different CNN based architectures. The experimental outcomes illustrate that data augmentation increases the system's overall performance by producing good accuracy results of multiple object segmentation. The system's accuracy is 93%, 95%, and 96% with CNN architectures VGG-16, ResNet-50, and MobileNet, respectively. We will continue this work in the future with other deep learning architectures and models. We will compare the results of different deep learning models with other benchmark data sets. Moreover, we expect to employ fine tuning to improve the system's performance with several publicly available data sets.

Author Contributions: Conceptualization, I.A., M.A., A.C., M.M.H. and G.J.; methodology, I.A., M.A., A.C., M.M.H. and G.J.; formal analysis, I.A. and M.A.; investigation, I.A. and M.A.; resources, I.A. and M.A.; data curation, I.A., M.A., A.C., M.M.H. and G.J.; writing, I.A., M.A., A.C., M.M.H. and G.J.; visualization, I.A., M.A., A.C., M.M.H. and G.J.; supervision, G.J.; project administration, G.J.; funding acquisition, A.C., M.M.H. and G.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the King Saud University Riyadh, Saudi Arabia, through the Researchers Supporting Project under Grant RSP-2021/18.

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

References

1. Yang, C.; Wong, D.; Miao, Q.; Yang, R. *Advanced GeoInformation Science*; CRC Press: Boca Raton, FL, USA, 2010.
2. Ahmed, I.; Din, S.; Jeon, G.; Piccialli, F. Exploring deep learning models for overhead view multiple object detection. *IEEE Internet Things J.* **2019**, *7*, 5737–5744. [CrossRef]
3. Ahmed, I.; Ahmad, M.; Nawaz, M.; Haseeb, K.; Khan, S.; Jeon, G. Efficient topview person detector using point based transformation and lookup table. *Comput. Commun.* **2019**, *147*, 188–197. [CrossRef]
4. Ahmad, M.; Ahmed, I.; Khan, F.A.; Qayum, F.; Aljuaid, H. Convolutional neural network–based person tracking using overhead views. *Int. J. Distrib. Sens. Netw.* **2020**, *16*, 1550147720934738. [CrossRef]
5. Ullah, K.; Ahmed, I.; Ahmad, M.; Rahman, A.U.; Nawaz, M.; Adnan, A. Rotation invariant person tracker using top view. *J. Ambient. Intell. Humaniz. Comput.* **2019** , 1–17. [CrossRef]
6. Ahmed, I.; Ahmad, M.; Jeon, G. A real-time efficient object segmentation system based on U-Net using aerial drone images. *J. Real-Time Image Process.* **2021**, *18*, 1745–1758. [CrossRef]
7. Ahmed, I.; Ahmad, M.; Khan, F.A.; Asif, M. Comparison of deep-learning-based segmentation models: Using top view person images. *IEEE Access* **2020**, *8*, 136361–136373. [CrossRef]
8. Pires de Lima, R.; Marfurt, K. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sens.* **2020**, *12*, 86. [CrossRef]
9. Ahmed, I.; Ahmad, M.; Rodrigues, J.J.; Jeon, G.; Din, S. A deep learning-based social distance monitoring framework for COVID-19. *Sustain. Cities Soc.* **2021**, *65*, 102571. [CrossRef]
10. Ahmad, M.; Ahmed, I.; Ullah, K.; khan, I.; Khattak, A.; Adnan, A. Person Detection from Overhead View: A Survey. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*. [CrossRef]
11. Zaitoun, N.M.; Aqel, M.J. Survey on image segmentation techniques. *Procedia Comput. Sci.* **2015**, *65*, 797–806. [CrossRef]
12. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEee Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [CrossRef] [PubMed]
13. Ahmed, I.; Ahmad, M.; Rodrigues, J.J.; Jeon, G. Edge computing-based person detection system for top view surveillance: Using CenterNet with transfer learning. *Appl. Soft Comput.* **2021**, *107*, 107489. [CrossRef]
14. Guzzo, A.; Sacca, D.; Serra, E. An effective approach to inverse frequent set mining. In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, Miami Beach, FL, USA, 6–9 December 2009, pp. 806–811.
15. Ahmed, I.; Ahmad, M.; Jeon, G.; Piccialli, F. A framework for pandemic prediction using big data analytics. *Big Data Res.* **2021**, *25*, 100190. [CrossRef]
16. Guzzo, A.; Moccia, L.; Saccà, D.; Serra, E. Solving inverse frequent itemset mining with infrequency constraints via large-scale linear programs. *ACM Trans. Knowl. Discov. Data (TKDD)* **2013**, *7*, 1–39. [CrossRef]
17. Ahmed, I.; Ahmad, M.; Ahmad, A.; Jeon, G. Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: Within 5G infrastructure. *Int. J. Mach. Learn. Cybern.* **2020**, *12*, 3053–3067. [CrossRef]
18. Ahmed, I.; Jeon, G.; Chehri, A.; Hassan, M.M. Adapting Gaussian YOLOv3 with transfer learning for overhead view human detection in smart cities and societies. *Sustain. Cities Soc.* **2021**, *70*, 102908. [CrossRef]
19. Ahmed, I.; Ahmad, M.; Ahmad, A.; Jeon, G. IoT-based crowd monitoring system: Using SSD with transfer learning. *Comput. Electr. Eng.* **2021**, *93*, 107226. [CrossRef]

20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
21. Chaudhuri, D.; Kushwaha, N.; Samal, A. Semi-automated road detection from high resolution satellite images by directional morphological enhancement and segmentation techniques. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1538–1544. [CrossRef]
22. Stankov, K.; He, D.C. Detection of buildings in multispectral very high spatial resolution images using the percentage occupancy hit-or-miss transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4069–4080. [CrossRef]
23. Stankov, K.; He, D.C. Building detection in very high spatial resolution multispectral images using the hit-or-miss transform. *IEEE Geosci. Remote Sens. Lett.* **2012**, *10*, 86–90. [CrossRef]
24. Leninisha, S.; Vani, K. Water flow based geometric active deformable model for road network. *ISPRS J. Photogramm. Remote Sens.* **2015**, *102*, 140–147. [CrossRef]
25. Drăguţ, L.; Csillik, O.; Eisank, C.; Tiede, D. Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 119–127. [CrossRef] [PubMed]
26. Ming, D.; Li, J.; Wang, J.; Zhang, M. Scale parameter selection by spatial statistics for GeOBIA: Using mean-shift based multi-scale segmentation as an example. *ISPRS J. Photogramm. Remote Sens.* **2015**, *106*, 28–41. [CrossRef]
27. Li, X.; Cheng, X.; Chen, W.; Chen, G.; Liu, S. Identification of forested landslides using LiDar data, object-based image analysis, and machine learning algorithms. *Remote Sens.* **2015**, *7*, 9705–9726. [CrossRef]
28. Contreras, D.; Blaschke, T.; Tiede, D.; Jilge, M. Monitoring recovery after earthquakes through the integration of remote sensing, GIS, and ground observations: The case of L'Aquila (Italy). *Cartogr. Geogr. Inf. Sci.* **2016**, *43*, 115–133. [CrossRef]
29. Arı, Ç.; Aksoy, S. Detection of compound structures using a Gaussian mixture model with spectral and spatial constraints. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6627–6638. [CrossRef]
30. Benedek, C.; Shadaydeh, M.; Kato, Z.; Szirányi, T.; Zerubia, J. Multilayer Markov random field models for change detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2015**, *107*, 22–37. [CrossRef]
31. Dong, Y.; Du, B.; Zhang, L. Target detection based on random forest metric learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1830–1838. [CrossRef]
32. Wang, J.; Song, J.; Chen, M.; Yang, Z. Road network extraction: A neural-dynamic framework based on deep learning and a finite state machine. *Int. J. Remote Sens.* **2015**, *36*, 3144–3169. [CrossRef]
33. Jain, A.; Ramaprasad, R.; Narang, P.; Mandal, M.; Chamola, V.; Yu, F.; Guizani, M. AI-enabled Object Detection in UAVs: Challenges, Design Choices, and Research Directions. *IEEE Netw.* **2021**, *35*, 129–135. [CrossRef]
34. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In Proceedings of the Asian Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2016; pp. 180–196.
35. Garg, P.; Chakravarthy, A.S.; Mandal, M.; Narang, P.; Chamola, V.; Guizani, M. Isdnet: Ai-enabled instance segmentation of aerial scenes for smart cities. *ACM Trans. Internet Technol.* **2020**, *21*, 1–18. [CrossRef]
36. Abdollahi, A.; Pradhan, B.; Alamri, A.M. An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images. *Geocarto Int.* **2022**, 37, 3355–3370. [CrossRef]
37. Marcu, A.; Costea, D.; Licaret, V.; Leordeanu, M. Towards automatic annotation for semantic segmentation in drone videos. *arXiv* **2019**, arXiv:1910.10026.
38. Maulik, U.; Chakraborty, D. Remote Sensing Image Classification: A survey of support-vector-machine-based advanced techniques. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 33–52. [CrossRef]
39. Song, J.; Gao, S.; Zhu, Y.; Ma, C. A survey of remote sensing image classification based on CNNs. *Big Earth Data* **2019**, *3*, 232–254. [CrossRef]
40. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
41. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
45. Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P.H. Conditional random fields as recurrent neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1529–1537.
46. He, C.; Fang, P.; Zhang, Z.; Xiong, D.; Liao, M. An end-to-end conditional random fields and skip-connected generative adversarial segmentation network for remote sensing images. *Remote Sens.* **2019**, *11*, 1604. [CrossRef]