remote sensing

MDPI

Article

Deep TEC: Deep Transfer Learning with Ensemble Classifier for Road Extraction from UAV Imagery

J. Senthilnath [1,*], Neelanshi Varia [2], Akanksha Dokania [3], Gaotham Anand [4] and Jón Atli Benediktsson [5]

[1] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore
[2] Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar 382007, India; neelanshiV2@gmail.com
[3] Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati 781039, India; akankshadokania@gmail.com
[4] Department of Aerospace Engineering, Indian Institute of Science, Bangalore 560012, India; gauthama.anand@gmail.com
[5] Electrical and Computer Engineering, University of Iceland, 101 Reykjavik, Iceland; benedikt@hi.is
* Correspondence: jsenthilnath@alum.iisc.ac.in

Received: 16 November 2019; Accepted: 7 January 2020; Published: 10 January 2020

Abstract: Unmanned aerial vehicle (UAV) remote sensing has a wide area of applications and in this paper, we attempt to address one such problem—road extraction from UAV-captured RGB images. The key challenge here is to solve the road extraction problem using the UAV multiple remote sensing scene datasets that are acquired with different sensors over different locations. We aim to extract the knowledge from a dataset that is available in the literature and apply this extracted knowledge on our dataset. The paper focuses on a novel method which consists of deep TEC (deep transfer learning with ensemble classifier) for road extraction using UAV imagery. The proposed deep TEC performs road extraction on UAV imagery in two stages, namely, deep transfer learning and ensemble classifier. In the first stage, with the help of deep learning methods, namely, the conditional generative adversarial network, the cycle generative adversarial network and the fully convolutional network, the model is pre-trained on the benchmark UAV road extraction dataset that is available in the literature. With this extracted knowledge (based on the pre-trained model) the road regions are then extracted on our UAV acquired images. Finally, for the road classified images, ensemble classification is carried out. In particular, the deep TEC method has an average quality of 71%, which is 10% higher than the next best standard deep learning methods. Deep TEC also shows a higher level of performance measures such as completeness, correctness and F1 score measures. Therefore, the obtained results show that the deep TEC is efficient in extracting road networks in an urban region.

Keywords: UAV; remote sensing; road extraction; deep learning; transfer learning; ensemble classifier

1. Introduction

Recent advances in remote sensing technologies have paved the way for a simpler and better way of monitoring geographical regions [1]. Traditionally, satellite remote sensing (SRS) has provided limited spatial and temporal resolution for applications like land cover mapping, weather, meteorology, mineralogy, etc. Specifically, with the increased popularity of unmanned aerial vehicles (UAVs) for varietal remote sensing applications, we have overcome the shortcomings of SRS with regards to spatial and temporal resolution [2]. Also, compared to the manned aerial systems, the UAV can be used in inaccessible areas, low altitudes and some places without endangering human life for applications [3]. Thus, UAV can facilitate better spatial and temporal resolutions at a lesser cost for various remote sensing applications [4,5]. The applications range from photogrammetry, feature extraction, target detection, urban monitoring, vegetation analysis, etc. [6].

This study focuses on addressing one such application of UAV remote sensing in urban monitoring, in particular, road extraction. Road extraction from street view as well as satellite and aerial view has become an integral problem for traffic management, self-driving vehicles, global positioning system (GPS)-based utilities, urban mapping and various other applications. In the past, geometrical and statistical methods have been suggested for road extraction [7]. Road extraction is difficult due to occlusions in the form of vehicles, trees, buildings and other non-road objects. With the advancement in machine learning techniques recently, a lot of work is available in the literature. In particular, deep learning techniques have reached a pinnacle in problems like object detection, semantic segmentation, classification, etc. [8–11]. Also, the advantage of sequential data in large amounts has helped to better perform tasks like change detection, pattern modeling, etc., and deep recurrent neural networks have been widely used for the same [12].

In this paper, a new deep transfer learning with ensemble classifier (deep TEC) is proposed for road extraction using UAV imagery. The proposed method deep TEC consists of two stages. In the first stage, we have implemented three deep transfer learning methods, namely, conditional generative adversarial networks (cGAN), CycleGAN and fully convolutional metwork (FCN) for road extraction in diverse types of backgrounds on a benchmark UAV road extraction dataset that is available in the literature [13] (Dataset-A). This extracted knowledge (pre-trained model) of road regions is then tested for road extraction in our real-time UAV-acquired images (Dataset-B) and prove their generalization ability (domain adaptation) based on transfer learning. In the next stage, we perform the ensemble-based classification model. The ensemble classifier aggregates the outcome of the previous stage used for testing the domain adaptability (three deep transfer learning methods) with a majority voting. The algorithms extracted the curves of roads as well as intersections quite efficiently. The results of the methods were analyzed using different measures along with the time taken for the segmentation.

In Section 2, we review the literature on deep learning methods applied to road extraction. In Section 3, we discuss the benchmark training UAV dataset (Dataset-A) with the specifications of the UAV used to capture the test dataset (Dataset-B). In Section 4, we discuss the methods and architecture applied for the task at hand and in Section 5, we discuss the evaluation metrics used for performance analysis of the results. In Section 6, we discuss the results obtained on 13 UAV images for road extraction. The paper is concluded in Section 7.

2. Related Work

With the advent of recent success with deep neural networks, especially with the help of multiple fast processing GPUs and the availability of remote sensing data, researchers have solved the road extraction problem using remote sensing data. The recently proposed JointNet utilizes the focal loss function to improve road extraction while maintaining a larger receptive field at the same time [14]. The JointNet is a combination of dense connectivity and atrous convolution that effectively extracts both road and building regions. Another recent work proposes the use of a convolutional neural network (CNN) to extract structural features and then apply multi-scale Gabor filters and edge-preserving filters to improve feature extraction [15]. Y-Net, a quite recent deep learning method, combines feature extraction and a fusion module that can better segment multi-scale roads in high-resolution images. In comparison with other methods, Y-Net performs better in extracting narrow roads. [16].

Few researchers have applied shallow neural networks for the road extraction [17]; some of which include larger trainable weights and take advantage of local spatial coherence of the output [18]. The recent advancement in computational speed and increased data resources have greatly fueled the usage of deep neural networks (DNN). Gao et al. [19], by taking advantage of high-resolution remote sensing data, proposed multi-level semantic features. In their study, a novel loss function is proposed to overcome misclassification error and helps to focus on the spare set of real labeled pixels in the training stage. A convolutional neural network (CNN) with different variants like fusion with a line integral convolution-based algorithm [20], a combination of deep convolutional neural network and finite-state machine [21], derivatives such as the road structure refined convolutional neural network (RSRCNN) [22], and DenseNet methods [23] were successfully applied on road segmentation. Fully convolutional networks (FCN) have recently gained a lot of popularity, and depending on the data availability and computational power a decision can be made on whether to use pre-trained nets like the VGG-Net. Generally, dense networks like VGG-Net and ResNet have a large number of layers which require a very long training time and hence their pre-trained weights are used to perform different tasks [12]. These nets are trained on datasets like ImageNet which have around 1000 classes. Training a network on such a large number of classes, with an extremely large dataset and very deep neural networks is difficult. Hence, pre-trained networks (weights) are used to perform specific tasks owing to sparse available dataset and computation power. Further nets have been proposed, including a U-shaped FCN comprising of a series of convolutions with the corresponding series of deconvolutions including skip connections in between [24], a network consisting of ResNet-34 which is pre-trained on ImageNet and the decoder is based on vanilla U-Net [25], and FCN with improved tuning [25] were successfully applied in road segmentation. Additionally, the refined deep residual convolutional neural network (RDRCNN) employs mathematical morphology and a tensor-based voting methodology to get improved results on complex road networks [11]. In this study, we have a sufficient training dataset (Dataset-A with 189 images) as well as a network (FCN) which isn't as deep. Therefore, taking advantage of the available dataset and lighter network, FCN can be trained for our task and does not require a pre-trained network.

In most of the aforementioned CNN-based deep learning techniques, the precision and accuracy of the segmentation increases greatly with the help of deep Networks architecture. However, they also prone to a lot of computational power and require large datasets. In the literature, generative adversarial networks (GANs) [26] have given better results in various problem areas like text-to-image, and image segmentation with much less computational requirements [27]. A lot of work has been done for road extraction from street view but not much work has been performed on road extraction from UAV remote sensing images. Along with deep neural networks for hyperspectral images [28], GANs have lately been used for hyperspectral image classification [29]. In some of the applications, the performance of GAN is better than CNN for road detection [30]. Furthermore, StreetGAN for road network synthesis [31] and other GANs [32,33] have been proposed.

In real-time applications, the data sources differ at times and so it is not feasible to retrain models repetitively. Hence, learning from a source data distribution for a model based on a different target data distribution is handy. A lot of labelled data is available to train models in some domains, but to generate our data with labels and annotations is expensive. The operation of collecting and labelling data is quite time consuming and hence, the necessity of leveraging available data for related domains arises which results in "transfer learning". Transfer learning has been used in many important tasks including road extraction [34,35], but not much literature is available in remote sensing-based road extraction [36]. Additionally, various neural network algorithms give different results and each of them has its speciality in extracting roads, i.e., their results differ in terms of correctness, completeness, linearity, etc. Consequently, ensemble learning becomes helpful to extract the best qualities of all the classifier outcomes. While using deep learning networks, with different structures, the results of semantic segmentation differ for each of them [34,35]. It is important to note that different networks learn features in different ways and hence bringing a combination/fusion of several methods can result in a better outcome. Over the past few years, quite a few ensemble-based techniques have been proposed for various tasks including speech recognition [37], disease detection [38], semantic segmentation of roads [39] etc., and deep learning-based models for remote sensing [40,41] have been proposed. However, for road extraction using remote sensing data, only one method based on an ensemble classifier [42] has been proposed.

3. UAV for the Detection of Road Networks

The deep learning methods are implemented on labelled images on two different data sources. "Dataset-A" that was used in the paper created by Zhou [13] is a large UAV remote sensing RGB image dataset with varying image sizes and locations, captured from different heights. Dataset-A was acquired using a UAV that flew in different locations in Australia. "Dataset-B" consists of the images captured by a UAV and a camera as described in Tables 1 and 2. It had 13 different RGB image frames from a video that covers partly/fully road networks, including a variety of suborbital roads. The frames were carefully chosen to represent different scenes from the continuous capture of a camera mounted on a UAV. The dataset had been acquired in different conditions including the angle of capture, device capturing the scene, day-light, height, etc., and such differences affect the data distribution in the feature space. Dataset-B was solely used for testing purposes. Field visits to the site of the road (Dataset-B) revealed that there were rarely any other built-up regions near these major roads/highways or at least within 250 m buffer from the sides/edges of these roads. A major reason is the Government regulations which do not allow any human activity in the vicinity of these roads (250 m). However, our observations were that detecting the major roads was not a problem with high spatial resolution images. Besides, roads are also differentiable from other surrounding land use categories because of their linear shapes and contiguity between similar reflectance pixels (tar and cement).

Table 1. Unmanned aerial vehicle (UAV) specifications.

Characteristic	Rating
Motor	BLDC, 630 w, 650 kv
Battery	60C 3S 5000 mAh LiPo
Remote Control	2.4 GHz, 8 channel
ESC	60 amp
Servo Motor	9 g, 2.2 Kg torque, 5 numbers
Max Gross Weight	2.2 kg
Flight stabilizer	Pixhawk
Thrust-weight ratio	1.2
Specific thrust	7.07 g/W
Prop-rotor Diameter	12 × 4.5 inch
Length	1210 mm
Height	300 mm
Stall Speed	5 m/s
Cruise Speed	15 m/s
Wingspan	2 m
Endurance	30 min

In the experiments, the number of training and test images used was 189 from Dataset-A and 13 from Dataset-B respectively. The images had a variety in terms of orientation, angle of capture and shapes. It contained occlusions and noise as well and thus a proper mix of data is used to train the network. In this study, we implement FCN-32, cGAN and CycleGAN on RGB images obtained from UAV remote sensing dataset and semantically segment the images into classes of "road" and "non-road". The methods are first trained on Dataset-A and then tested on Dataset-B to check the efficiency of its domain adaptability. Further, ensemble classifier is used to improve the road extraction.

Table 2. Camera specifications.

Parameter	Value
Camera make	Go Pro Hero 3
Video resolution	1080 p
Frame rate	60 Hz
Sensor resolution	12 mega pixel
Sensor Size	6.17 mm × 4.55 mm
Field of view	170 degrees
Weight	75 g
Video format	H.264 MP4

The UAV platform used in this study was the custom-built fixed-wing aircraft. The UAV was designed to be naturally very stable and modular for easy deployment in the field. The weight of the entire platform was around 2.2 kg and the time to deploy from transport to flight was only 10 min. Detailed specifications of the UAV are mentioned in Table 1.

Figure 1 shows the system description of the UAV. The onboard flight control system was an arm-based open-source Pixhawk autopilot. The flight controller was equipped with a GPS and an inertial measurement unit (IMU) to measure the UAV location and flight attitude, respectively. A ground-based laptop installed with an open-source flight control system (QGroundControl) remotely connected to the aircraft in real-time using wireless radio communication devices. A GoPro 3 (GoPro HD Hero 3, San Mateo, California) camera, which has its specifications mentioned in Table 2, was affixed to the plane aft location to reduce vibration of the motor affecting the video quality. The desired flight path covering an area was delineated with waypoint coordinates in the QGroundControl software. The UAV during the flight mission followed these waypoint coordinates automatically while simultaneously recording its flight path, location and attitude.

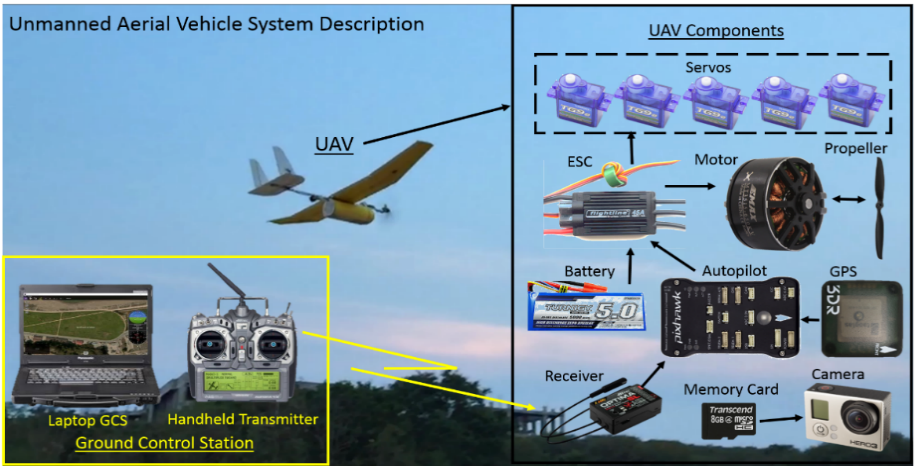

Figure 1. System description of the UAV.

4. Methodology

We propose deep TEC, a deep transfer learning with an ensemble classifier-based framework. The proposed deep TEC method adopts spectral-spatial-based approach that combines both the reflectance property and the spatial arrangements of the roads pixels that are linear features, and they are trained from Dataset-A. The method is divided mainly into two stages, namely, deep transfer learning and ensemble classifiers as shown in Figure 2.

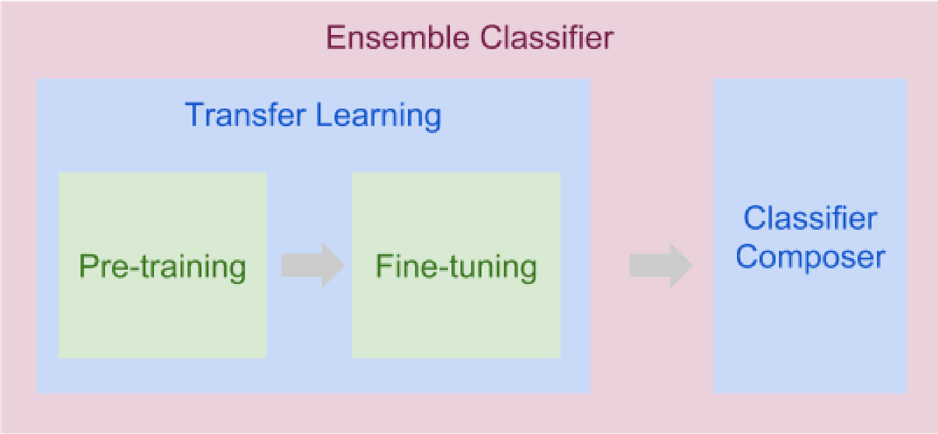

Figure 2. Block diagram of the deep transfer learning with ensemble classifier (TEC) approach.

4.1. Deep Transfer Learning

As discussed in the literature [43], a domain D consists of two parts: Feature space X and a marginal probability distribution P(X), where X = {x_1, ..., x_n}, and n is the number of data instances. For a particular domain, D = {X, P(X)}, a task T comprises of two parts: Objective predictive function $f(.)$, denoted by $T = \{Y, f(.)\}$, and a label space Y. Here, the task T is learned from the training data, which consist of pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$, where $i = 1, 2, ..., n$ and the function $f(.)$ maps the training data X to the label Y. During the training phase, knowledge is extracted and the testing phase gives the output for a new X. For the current task, we have defined one source domain D_S and, one target domain D_T. Here $D_S = \{(x_s^1, y_s^1), ..., (x_s^n, y_s^n)\}$, where $x_s^i \in X_S$ is the data instance, i.e., the input image, $y_s^i \in Y_S$ is the corresponding output image and n is the number of instances. Similarly, D_T is defined as $\{(x_T^1, y_T^1), ..., (x_T^m, y_T^m)\}$, where the input image x_T^i is in X_T, $y_T^i \in Y_T$ is the corresponding output image and m is the number of instances in the test set.

In Figure 3, transfer learning phase is defined as a source domain (D_S) and a task (T_S), a target domain (D_T) and learning task (T_T), and knowledge transfer has a goal to improve the learning of the output prediction function $f_T(.)$ in D_T using the extracted knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$. In our case, the source domain and target domain are Dataset-A and Dataset-B respectively, where $D_S \neq D_T$. The source domain and target domain are not the same due to multiple reasons. The geographical locations where datasets were captured are widely different. Dataset-A was captured in Australia whereas Dataset-B was captured in India. The elevation and angle of the capturing of datasets are also different which makes the resolution, size and orientation of images varietal. Dataset-A was captured from an altitude of 130–170 m while Dataset-B was captured from an altitude of 50–80 m. There is also a stark difference in style of road networks and the landscape in terms of complexity and degree of convolution of road networks, colors, occlusion by vehicles, trees, etc. For the current experiment, we train the classifier to learn extraction of road networks from Dataset-A. Since we want the trained classifier to extract road networks (not other classes like vegetation, vehicles, buildings, etc.), the tasks are the same and hence $T_S = T_T$.

The transfer learning is divided mainly into two stages: (i) Source task selection (defining T_S) and pre-training (developing a model on D_S), and (ii) fine-tuning weights (defining T_T) and reusing the model on the target domain (applying the model on D_T). For the current task, since $T_S = T_T$, we do not fine-tune the trained model.

For the current setup, the source task (T_S) was to classify the pixels of an image into a "road" pixel and a "non-road" pixel for all the images of a UAV remote sensing test dataset, i.e., to identify road networks in any given bird-view image. For training, we can consider any of the suitable varieties of DNN. For this case, we implemented three methods, namely, fully convolutional network's 32 derivative [44], CycleGAN [45] and conditional generative adversarial network (pix2pix) [46]. The architecture and parameters had been taken the same as proposed earlier and proved to be efficient [47] on the source domain dataset, i.e., Dataset-A as mentioned in Section 3.

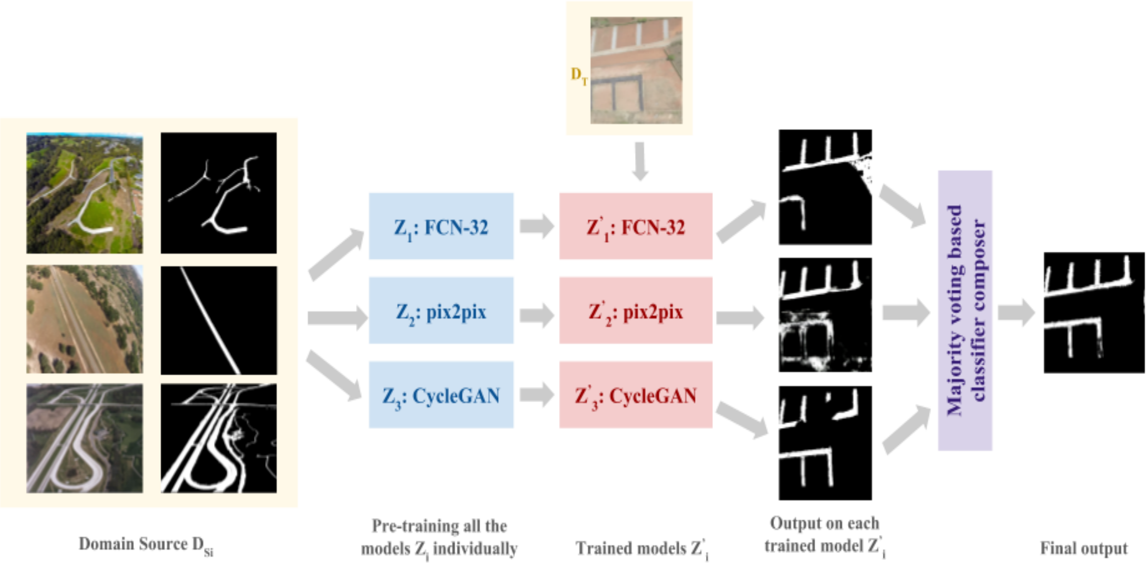

Figure 3. The deep TEC architecture.

The FCN mainly consists of three types of components, such as convolution, pooling and activation function. The output dimension of the FCNs is the dimension of data (images in this case) and d is the number of channels. The input data point, denoted by t the location (i, j) has output as shown by Equation (1).

$$y_{i,j} = f_{ks}(\{X_{si+\delta i,sj+\delta j}\}0 \leq \delta i, \delta j \leq k), \tag{1}$$

where f_{ks} determines the type of the layer. Here k is the kernel size and s is the stride. For individual pixel classification, training is done with back-propagation to minimize the softmax loss function. In this setup, we used convolution filters of size 3 × 3 with stride 1. rectified linear unit (ReLU) and leaky ReLU functions were used alternatively for all the layers.

Compared with most deep learning techniques, FCN has the advantage to adopt the fully connected layers to the convolutional layers to extract the object of any spatial distribution. This provides the freedom to train the dataset of any resolution (varying spatial size) of the object of interest. In our study, Dataset-A contains images of different sizes and that is why FCN gives the freedom to train a network with images of any size to extract road segments.

The pix2pix method uses U-Net as the generator part of the GAN. U-Net possesses skip layers to help pass the low-level information directly to the corresponding decoding network without having it to pass through all the other layers. The discriminator network uses PatchGAN for discriminating between real and generated image at a patch level, meaning that it penalizes area of the image that are not real. For this setup, the down-sampling stack consists of two convolutional layers, the filters of size 3 × 3, a ReLU layer followed by a max-pooling layer with stride 2 and this stack of layers is repeated.

CycleGAN was built upon the power of unpaired images, and learned to translate on a combination of cycle loss and adversarial loss. It learned a mapping $G : X \implies Y$ in such a way that the distribution of images from $G(X)$ was indistinguishable from the distribution $G(Y)$. It did the task via adversarial loss coupled with an inverse mapping $F : Y \implies X$ such that it had a cycle consistency loss to enforce $F(G(X)) \approx X$ (and vice versa). In simpler words, two adversarial discriminators D_X and D_Y were introduced where D_X discriminated between a set of images $\{x\}$ and translated a set of images $\{F(y)\}$ whereas D_Y distinguished between a set of images $\{y\}$ and the set $\{G(x)\}$. For this setup, we used two strides, two convolutions and residual blocks. We also used two fractionally strided convolutions with stride $\frac{1}{2}$ as proposed in the original paper. We used nine blocks for the images and applied instance normalization. The discriminator network used 70 × 70 PatchGANs to classify the patches as in cGAN.

Let the denotations of the techniques be, (Z_i) where $i = 1, 2, 3$ such that FCN-32 (Z_1), cGAN (Z_2) and CycleGAN (Z_3). For each and every network Z_i, the parameters were fine-tuned so that they gave better performances for the given dataset. Let this trained network be denoted as Z'_i shown in Figure 3.

This type of transfer learning where the source and target tasks are the same is known as transductive transfer learning. Here, the parameters of the trained networks are completely "frozen" and the trainable parameters need not be retrained. Also, the last layer of the network, i.e., the softmax layer in our case, remains the same as it has to predict the same classes, i.e., "road" and "non-road".

4.2. Classifier Composer

An ensemble classifier usually contains a training set, base inducer and combiner (also called composer) [48]. In our case, the following components are identified: (i) Training set represents a labelled dataset for ensemble training, which in the current case is source domain (D_S). Let it be denoted as $X = \{x_1, ..., x_i, ..., x_n\}$ where n is the number of inputs and y is the target attribute i.e., "road" and "non-road"; (ii) base inducer is the methods Z_i, where i = 1, 2, 3 for FCN-32, pix2pix and CycleGAN respectively; (iii) the composer is responsible for combining the classifications of the various classifiers, and it gets the input images from the inducers as tuples and gives the final output. We define a weighting-based ensemble method known as plurality vote (PV) or basic ensemble method (BEM). Classification of pixels of the input images is performed according to the class ("road" and "non-road") that gets the highest number of votes.

Mathematically this can be written as defined in Equation (2).

$$class(x) =_{c_j \in dom(y)} (\Sigma_k g(y_{z'_i}(x), c_j)), \tag{2}$$

where $y_{z'_i}(x)$ is the classification of the Z'^{th}_i trained network (classifier), and $j = \{0, 1\}$ and $g(y, c)$ is an indicator function as defined in Equation (3).

$$g(y, c) = \begin{cases} 1, & \text{if } y = c \\ 0, & \text{otherwise} \end{cases}. \tag{3}$$

The main steps of the deep TEC are given in Algorithm 1.

Algorithm 1 Deep transfer learning and ensemble classifier (deep TEC)

```
Input: ${(x_s^1, y_s^1), ..., (x_s^n, y_s^n)}$, where $x_s^i \in X_S$ is the data instance, i.e., the input image and $y_s^i \in Y_S$ is the corresponding output image for $i$ = 1, 2, ..., $N$ instances
Output: Domain target $D_{Tj}$ target labels for $j$ = 1, 2, ..., $M$ test images

begin
Define P suitable networks for the given task $Z_k$, $k$ = 1, 2, ..., $P$
for k = 1 to $P$ do
    Pre-train the network $Z_k$ on $D_S$ for $N$ instances by finding the optimum parameters by Equation (1)
    Define regularization parameters for network $Z_k$
Save the model as $Z'_k$

for models $Z_k$: $k$ = 1 to $P$ do
    for the test images $D_T$: $j$ = 1 to $M$ do
        Give image $D_{Tj}$ as input to saved model $Z'_k$
        Save the classified target label pixels $Y_{kj}$

for output image $D_{Tj}$: $j$ = 1 to $M$ do
    for all the pixels q of image $D_{Tj}$ do
        for all the methods 1 to P by Equation (2) do
            if pixel q is road then
                vote $q_{road} + +$
            else
                vote $q_{non-road} + +$
            if $q_{road} + + > q_{non-road} + +$ then
                assign q = 1 (i.e., road) by Equation (3)
            else
                assign q = 0 (i.e., non-road) by Equation (3)
end
```

5. Evaluation Metrics

In this study, the performance measures, namely, correctness, completeness, quality and F1 score were used to analyze [49]. The performance of road extraction was compared for different deep learning methods such as FCN-32, pix2pix and CycleGAN outputs along with the proposed deep transfer learning with the ensemble classifier (deep TEC). Correctness essentially describes the purity of the positive length (road segments) prediction relative to the ground truth, completeness describes the totality of the positive detection (road segments) relative to the ground truth and quality is the accuracy relative to the ground truth. The F1 score is the harmonic mean of completeness and correctness. All these parameters are defined in terms of true positives (TP), false positives (FP) and false negatives (FN). TP is the length of the road extracted as the road and the corresponding output label also indicates it is a road. If the classified length is not a road but the ground truth indicates it to be road, then it is counted as FN. FP is the length of the road that is indicated as a road but the ground-truth indicates otherwise. Based on these parameters, the performance measures completeness, correctness, quality and F1 score are defined as shown in Equations (4)–(7) respectively.

$$Completeness = \frac{TP}{TP + FN}, \tag{4}$$

$$Correctness = \frac{TP}{TP + FP}, \tag{5}$$

$$Quality = \frac{TP}{TP + FP + FN}, \tag{6}$$

$$F1\,score = \frac{2 * Completeness * Correctness}{Completeness + Correctness}. \tag{7}$$

Apart from these, we also calculated gap density (Equation (8)) which was essentially the average number of gap pixels per gap in the image. A gap was identified as the region where all the connected road pixels were detected as non-road by a classifier.

$$Gap\ Density = \frac{\Sigma_{i=0}^{n} a_i}{n}, \tag{8}$$

where a is the total number of pixels covered by gaps and n is the total number of gaps. This parameter helps in analyzing the fragmentation of the extracted output. The lesser the gap density is, the more consistent the output.

6. Results and Discussion

In this section, we compare the performance of the proposed method, deep TEC with the well-known deep learning methods such as FCN-32, pix2pix and CycleGAN on 13 test images. The performance of these methods is evaluated using completeness, correctness, quality, F1 score and the gap density. As discussed in Section 3, the training is performed on the standard annotated Dataset-A that are available in [50,51]. Further, the test images (Dataset-B) consist of the UAV remote sensing dataset that has been acquired by the UAV and camera discussed in Section 3. Test images were chosen to be representative of different nature (concrete and non-concrete), shape and structure (linear and non-linear) road characteristics.

6.1. Performance Evaluation

Figure 4 compares the performance of the algorithms on 13 UAV images (Dataset-B) in terms of completeness. The completeness metric gives us the fraction of the true road pixels that were correctly detected in the output. The median completeness value for deep TEC is around 0.92, greater than other algorithms, implying that it has correctly identified 90 percent of the true road pixels in at least half the number of images. Deep TEC also performs better in terms of average completeness and the variation in the completeness value observed across 13 images than pix2pix, FCN and CycleGAN (Table 3).

Similarly, Figure 5 illustrates the performance of algorithms on 13 UAV images (Dataset-B) in terms of correctness. The correctness parameter tells us the probability that a road pixel detected by a classifier is actually a road pixel. Deep TEC is able to perform better than pix2pix, CycleGAN and FCN in most of the images. Its mean correctness 0.82 is higher than that of pix2pix (0.73), FCN (0.72) and CycleGAN (0.76) (Table 3). The median correctness value for deep TEC is approximately 0.82, which indicates that at least 80 percent of predicted road pixels by deep TEC are true roads in half of the images.

Correctness primarily focuses on accuracy and overlooks the missing data while completeness overlooks accuracy. It becomes imperative for us to use another metric F1 score, defined as the harmonic mean of completeness and correctness, which balances both the metrics. Figure 6 brings out a comparison of algorithms on the basis of their F1 scores across 13 UAV images using a box plot. Deep TEC, again, has the highest median and smallest inter-quartile range among all the algorithms. Its each quartile values, i.e., Q1, Q2 and Q3 are greater than other algorithms.

In terms of the quality measure, the deep TEC performs better than the other classifiers for each and every image and has a better margin. This can be observed in Figure 7 where the quality trend for the deep TEC is better than the mean quality of the FCN-32, pix2pix and CycleGAN methods. Hence, the proposed deep TEC gives the better performance in terms of the completeness, correctness as well as quality than all the three methods (pix2pix, FCN and CycleGAN) and can be seen in the

Table 3 and Figure 7 and that is why the proposed majority voting-based classifier proves to be more efficient than the sole classifier.

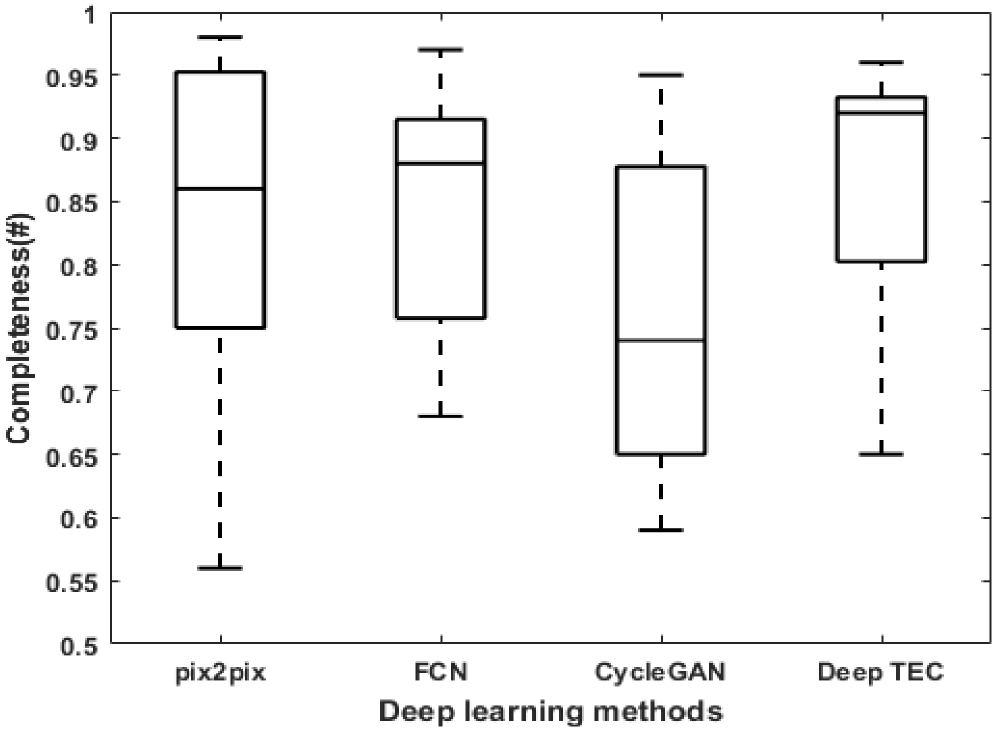

Figure 4. Evaluation of completeness for road extraction using deep learning methods.

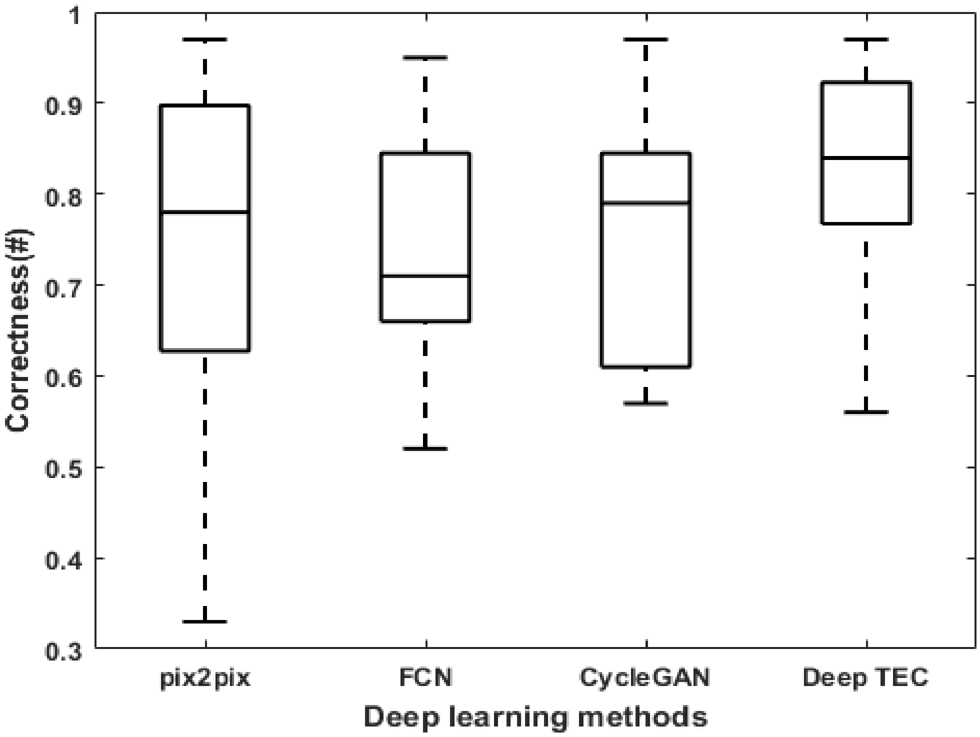

Figure 5. Evaluation of correctness for road extraction using deep learning methods.

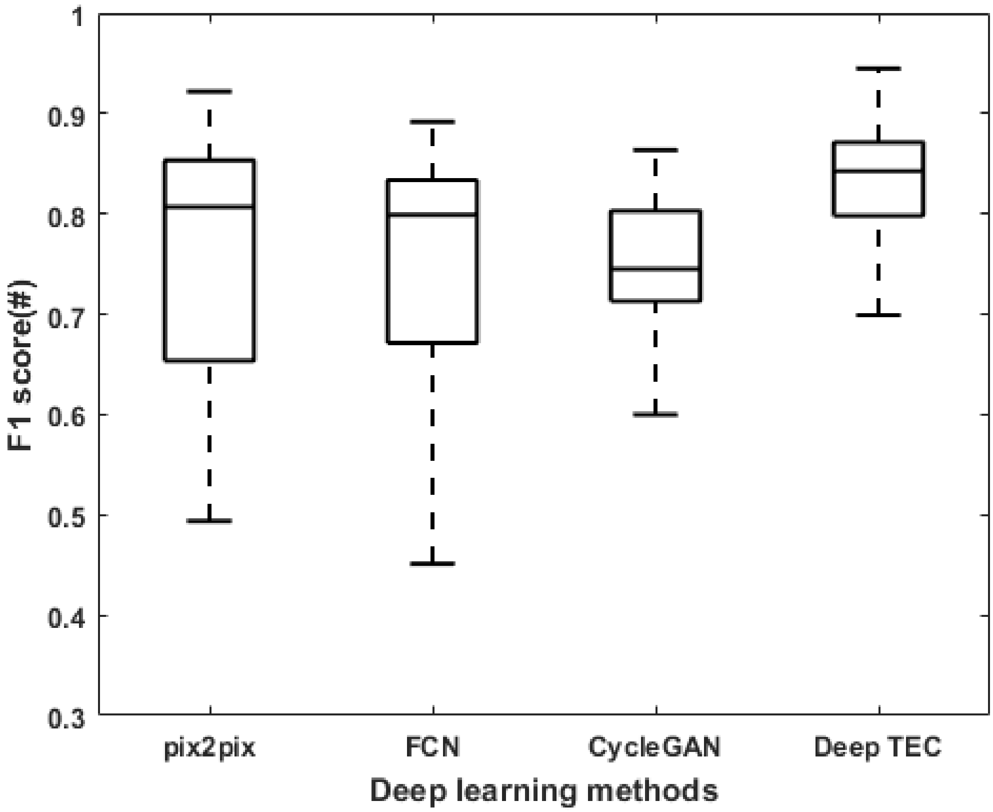

Figure 6. Evaluation of F1 score for road extraction using deep learning methods.

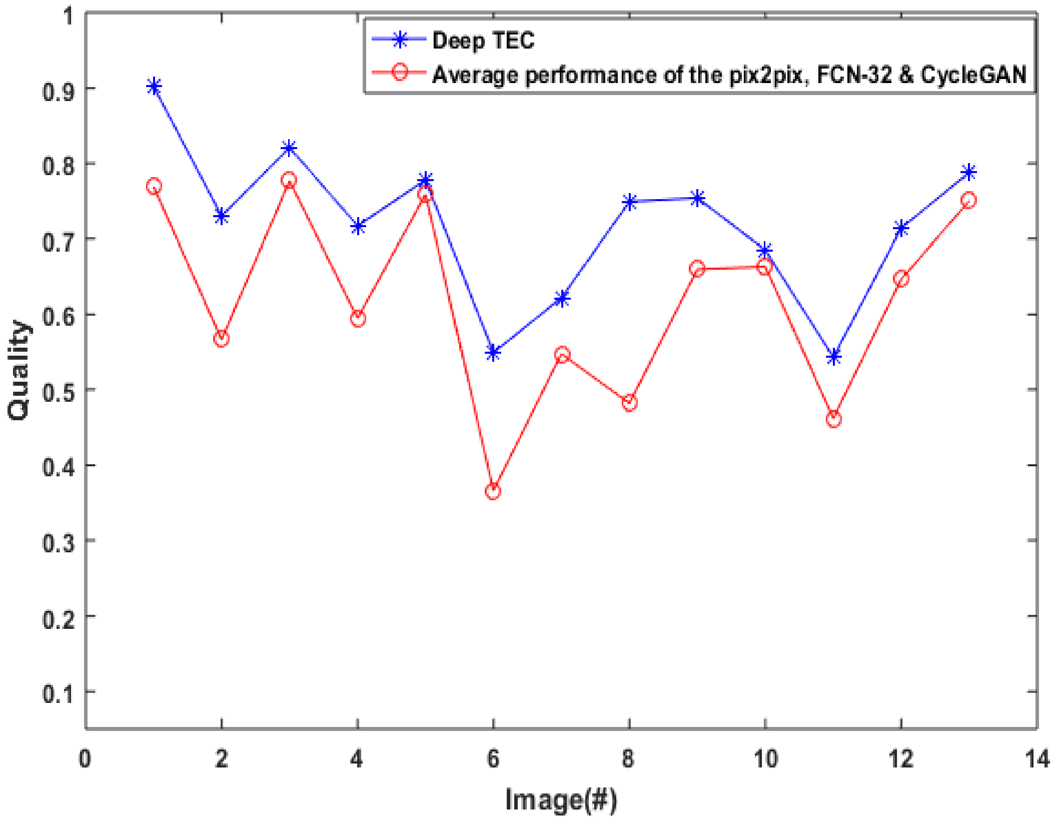

Figure 7. Comparison of the quality trend for the deep TEC with the mean quality of the FCN-32, pix2pix and CycleGAN methods.

To analyze the performance further, the gap analysis of the extracted images was also performed. For each image, we observed the road regions which were not extracted by the classifier, computed the number of pixels representing these regions and the number of such regions. A new performance measure (gap density) where the number of the gap pixels per gap was formulated and used. Figure 8 shows the boxplot for pix2pix, FCN, CycleGAN and deep TEC which is useful to statistically analyze the variation of gap density across 13 images. We can observe that deep TEC and pix2pix have lesser minimum and maximum ranges in comparison with the FCN and CycleGAN. Further, the median, first and second quartile range for deep TEC are less than that of pix2pix, FCN and CycleGAN. CycleGAN had better correctness among all the methods, but when we consider gap density, CycleGAN has higher variation than pix2pix and FCN. Overall, pix2pix has lower gap density which is second best but deep EC performs best among the four methods.

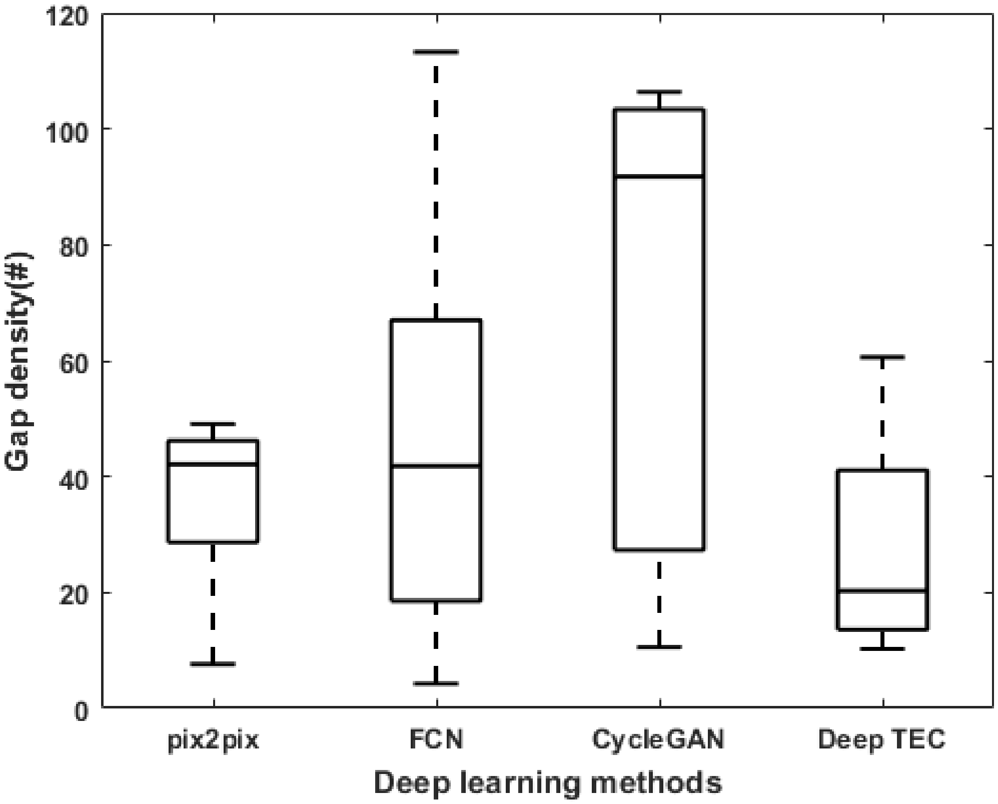

Figure 8. Gap density analysis for road extraction using deep learning methods.

Table 3. Analysis of performance measures for road extraction using deep learning methods.

Deep Learning Methods	**Completeness** ($\mu \pm \sigma$)	**Correctness** ($\mu \pm \sigma$)	**Quality** ($\mu \pm \sigma$)	**Gap Density** ($\mu \pm \sigma$)
pix2pix	0.84 ± 0.13	0.73 ± 0.22	0.62 ± 0.17	43.5 ± 28.3
FCN	0.83 ± 0.12	0.72 ± 0.17	0.62 ± 0.15	47.3 ± 32.4
CycleGAN	0.76 ± 0.11	0.76 ± 0.12	0.61 ± 0.09	134.6 ± 190.7
Deep TEC	**0.87** ±0.09	**0.82** ±0.13	**0.71** ±0.10	**30.9** ±25.0

6.2. Visual Analysis

From Figure 9, we can see that for simpler images like one to four, pix2pix was able to extract better correctness than CycleGAN. And for the fifth image as well, it is very similar. As mentioned, pix2pix has better completeness on average. This is because the algorithm learns to generate images based on individual image and ground-truth pair, and hence it is easy to extract characteristics like simple roads, perfect grey color, etc. accurately from any given image but when it comes to complexities, the variety in sizes and networks of roads make it difficult to transfer the learning. That is why it is very good at extracting simple roads and it has comparatively less false positives in complex ones. This also explains the standard deviation of pix2pix's completeness being much higher than others shown in Table 3, as the consistency in the results is very much dependent on the complexity and nature of roads.

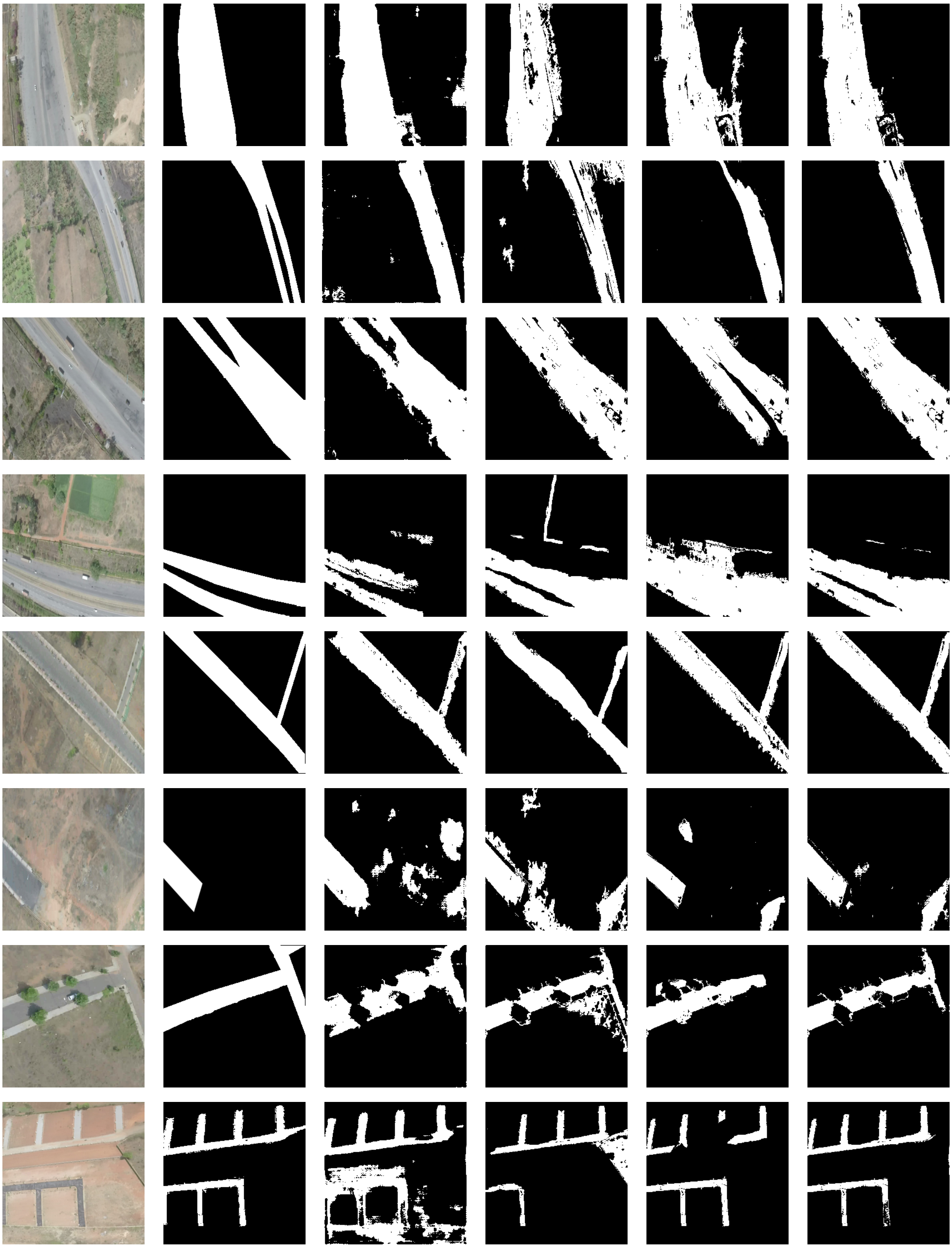

Figure 9. *Cont.*

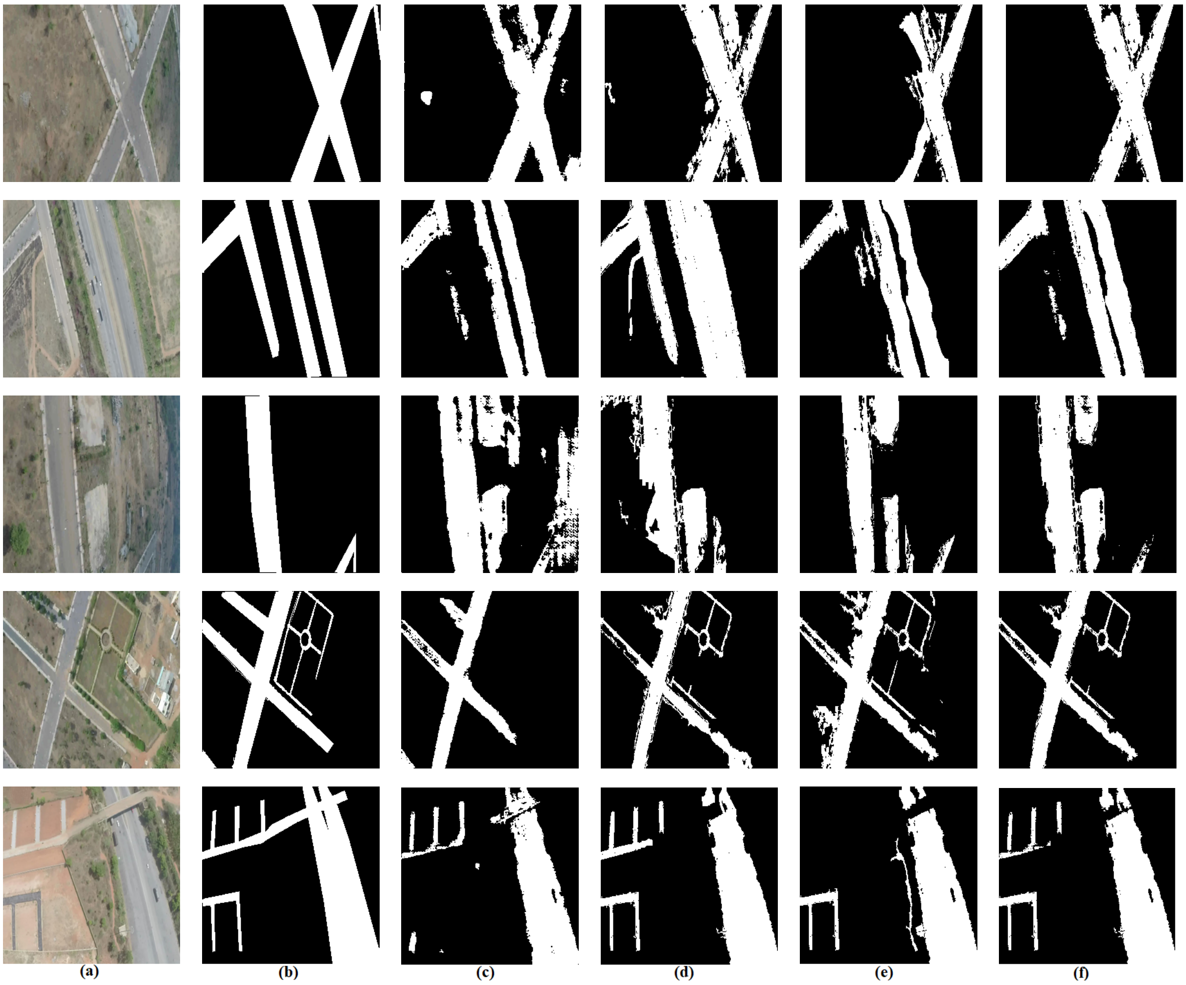

Figure 9. Visual comparison of four deep transfer learning methods applied on thirteen images: (**a**) Input image; (**b**) ground-truth; (**c**) pix2pix-based road network extraction; (**d**) FCN-based road network extraction; (**e**) CycleGAN-based road network extraction; (**f**) deep TEC-based road network extraction.

After the fifth image in Figure 9, overall CycleGAN performs very well in terms of correctness. CycleGAN learns generating images based on the overall semantics of the given data and not individual pairs. This helps in extracting not just the color-based features but also curvatures and edges and that is why it results in better correctness. This also proves better consistency (least standard deviation shown in Table 3) of the results as the way in which it extracts is style dependent and not particularly image dependent.

From Table 3, we can observe that the proposed deep TEC of combining the transfer learning and ensemble classifier gives the best performance for all performance measures. The mean and median of its completeness, correctness and F1 score values are much higher than other three methods (pix2pix, FCN and CycleGAN) as can be seen in the Figure 9, and that is why the proposed ensemble classifier with the majority voting-based classifier proves to be more efficient than the individual deep transfer learning approach; this effect can also be seen in Figure 9.

6.3. Computational Complexity

All the algorithms were run on the same system with python environment. The system configuration was 16 GB RAM with an i-7 processor. The time taken to train each of the models with 189 images is shown in Table 4.

Table 4 shows that cGAN converges to the solutions the quickest amongst the deep learning methods followed by FCN. The average training times for cGAN and FCN are around 300 s and 370 s respectively. The average time for cGAN is significantly lower than CycleGAN.

Table 4. Time taken for execution.ss.

Method Name	Mean Training Time	Number of Images	Total Training Time
FCN	∼370 s	189	∼20 h
cGAN	∼300 s	189	∼16 h
CycleGAN	∼420 s	189	∼23 h

In this study, we have applied deep learning methods on Dataset-A, which are more efficient in extracting features but take longer computational times. In the literature, clustering methods are applied on Dataset-A which takes an average of 0.283 s whereas our proposed method deep TEC takes around 370–420 s on average for training. Deep TEC combines the advantages of three deep learning methods (with extracted knowledge and ensemble) and is deployed to transfer knowledge to a different domain on 13 testing images with a mean detection time of ∼2 s.

7. Conclusions

In this paper, a deep TEC (deep Transfer learning with Ensemble Classifier) for road extraction is presented using UAV remote sensing RGB images which perform better. Initially, transfer learning is applied on Dataset-A which is an annotated standard road extraction dataset available in the literature. The trained model of pix-2-pix, FCN-32 and CycleGAN are then applied on UAV test images (Dataset-B). Then the ensemble classifier is implemented on the principle of the majority voting method, obtained by implementing transfer learning on three deep networks. It can be seen that the transfer learning performed on Dataset-B gives satisfactory outputs. Based on the evaluation matrices, we can observe that pix2pix gives an overall better completeness outcome and CycleGAN gives a better correctness outcome. At the same time, for some images, FCN performs better than the pix2pix and CycleGAN. Hence, the proposed deep TEC gives better output for all the results than all the three standalone methods.

In the future, the dataset for training can be further improved with the addition of images from different UAVs and the extraction of roads in complex sites like city roads and avenues. The ensemble classifier can also be further enhanced by exploiting more deep learning networks and by implementing a weighted vote-based method. This could lead to designing the ensemble classifier based on the weighted vote-based method, which might improve the ensemble results.

Author Contributions: J.S. conceived and designed the research; J.S. and G.A. Data curation; N.V., J.S., and A.D. developed and analyzed the data; N.V., J.S., A.D., G.A., and J.A.B. wrote and edited the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank S.N. Omkar and Chandrashekar from IISc, Bangalore, India, for providing us with the UAV remote sensing data (Dataset-B).

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

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
RS	Remote Sensing
FCN-32	Fully Convolutional Neural Network-32 derivative
cGAN	Conditional Generative Adversarial Network
CycleGAN	Cycle Generative Adversarial Network

ReLU	Rectified Linear Unit
pix2pix	Image to image translation cGAN
Deep TEC	Deep Transfer Ensemble Classifier
TP	True Positives
FP	False Positives
FN	False Negatives

References

1. Senthilnath, J.; Kumar, D.; Benediktsson, J.A.; Zhang, X. A novel hierarchical clustering technique based on splitting and merging. *Int. J. Image Data Fusion* **2016**, *7*, 19–41. [CrossRef]
2. Guo, Y.; Senthilnath, J.; Wu, W.; Zhang, X.; Zeng, Z.; Huang, H. Radiometric calibration for multispectral camera of different imaging conditions mounted on a UAV platform. *Sustainability* **2019**, *11*, 978. [CrossRef]
3. Bhola, R.; Krishna, N.H.; Ramesh, K.N.; Senthilnath, J.; Anand, G. Detection of the power lines in UAV remote sensed images using spectral-spatial methods. *J. Environ. Manag.* **2018**, *206*, 1233–1242. [CrossRef] [PubMed]
4. Senthilnath, J.; Dokania, A.; Kandukuri, M.; Ramesh, K.N.; Anand, G.; Omkar, S.N. Detection of tomatoes using spectral-spatial methods in remotely sensed RGB images captured by UAV. *Biosyst. Eng.* **2016**, *146*, 16–32. [CrossRef]
5. Senthilnath, J.; Kandukuri, M.; Dokania, A.; Ramesh, K.N. Application of UAV imaging platform for vegetation analysis based on spectral-spatial methods. *Comput. Electron. Agric.* **2017**, *140*, 8–24. [CrossRef]
6. Ma, L.; Li, M.; Tong, L.; Wang, Y.; Cheng, L. Using unmanned aerial vehicle for remote sensing application. In Proceedings of the 21st International Conference on Geoinformatics, Kaifeng, China, 20–22 June 2013.
7. Movaghati, S.; Moghaddamjoo, A.; Tavakoli, A. Road Extraction From Satellite Images Using Particle Filtering and Extended Kalman Filtering. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 2807–2817. [CrossRef]
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv* **2015**, arXiv:1506.01497.
9. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:1608.06993.
10. Kendall, A.; Badrinarayanan, V.; Cipolla, R. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *arXiv* **2015**, arXiv:1511.02680.
11. Gao, L.; Song, W.; Dai, J.; Chen, Y. Road Extraction from High-Resolution Remote Sensing Imagery Using Refined Deep Residual Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 552. [CrossRef]
12. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
13. Zhou, H.; Kong, H.; Wei, L.; Creighton, D.; Nahavandi, S. Efficient Road Detection and Tracking for Unmanned Aerial Vehicle. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 297–309. [CrossRef]
14. Zhang, Z.; Wang, Y. JointNet: A Common Neural Network for Road and Building Extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
15. Liu, R.; Qiguang, M.; Jianfeng, S.; Yining, Q.; Yunan, L.; Pengfei, X.; Jing, D. Multiscale road centerlines extraction from high-resolution aerial imagery. *Neurocomputing* **2019**, *329*, 384–396. [CrossRef]
16. Li, Y.; Xu, L.; Rao, J.; Guo, L.; Yan, Z.; Jin, S. A Y-Net deep learning method for road segmentation using high-resolution visible remote sensing images. *Remote Sens. Lett.* **2019**, *10*, 381–390. [CrossRef]
17. Mokhtarzade, M.; Zoej, M.V. Road detection from high-resolution satellite images using artificial neural networks. *Int. J. Appl. Earth Obs. Geoinform.* **2007**, *9*, 32–40. [CrossRef]
18. Mnih, V.; Hinton, G.E. Learning to Detect Roads in High-Resolution Aerial Images. In Proceedings of the Computer Vision—ECCV 2010 Lecture Notes in Computer Science, Crete, Greece, 5–11 September 2010; pp. 210–223.
19. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An End-to-End Neural Network for Road Extraction From Remote Sensing Imagery by Multiple Feature Pyramid Network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]

20. Li, P.; Zang, Y.; Wang, C.; Li, J.; Cheng, M.; Luo, L.; Yu, Y. Road network extraction via deep learning and line integral convolution. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016.
21. 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]
22. Wei, Y.; Wang, Z.; Xu, M. Road Structure Refined CNN for Road Extraction in Aerial Image. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 709–713. [CrossRef]
23. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
24. Kestur, R.; Farooq, S.; Abdal, R.; Mehraj, E.; Narasipura, O.; Mudigere, M. UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle. *J. Appl. Remote Sens.* **2018**, *12*, 1. [CrossRef]
25. Henry, C.; Azimi, S.M.; Merkle, N. Road Segmentation in SAR Satellite Images with Deep Fully Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1867–1871. [CrossRef]
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; pp. 2672–2680.
27. Majurski, M.; Manescu, P.; Padi, S.; Schaub, N.; Hotaling, N.; Simon, C., Jr.; Bajcsy, P. Cell Image Segmentation Using Generative Adversarial Networks, Transfer Learning, and Augmentations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019.
28. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
29. Zhu, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Generative Adversarial Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *26*, 5046–5063. [CrossRef]
30. Shi, Q.; Liu, X.; Li, X. Road Detection from Remote Sensing Images by Generative Adversarial Networks. *IEEE Access* **2018**, *6*, 25486–25494. [CrossRef]
31. Hartmann, S.; Weinmann, M.; Wessel, R.; Klein, R. StreetGAN: Towards Road Network Synthesis with Generative Adversarial Networks. Available online: https://otik.uk.zcu.cz/bitstream/11025/29554/1/Hartmann.pdf (accessed on 13 December 2019).
32. Tao, Y.; Xu, M.; Zhong, Y.; Cheng, Y. GAN-Assisted Two-Stream Neural Network for High-Resolution Remote Sensing Image Classification. *Remote Sens.* **2017**, *9*, 1328. [CrossRef]
33. Costea, D.; Marcu, A.; Leordeanu, M.; Slusanschi, E. Creating Roadmaps in Aerial Images with Generative Adversarial Networks and Smoothing-Based Optimization. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017.
34. Wang, Q.; Fang, J.; Yuan, Y. Adaptive road detection via context-aware label transfer. *Neurocomputing* **2015**, *158*, 174–183. [CrossRef]
35. Ros, G.; Stent, S.; Alcantarilla, P.F.; Watanabe, T. Training constrained deconvolutional networks for road scene semantic segmentation. *arXiv* **2016**, arXiv:1604.01545.
36. Mattyus, G.; Wang, S.; Fidler, S.; Urtasun, R. Enhancing road maps by parsing aerial images around the world. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1689–1697.
37. Deng, L.; Platt, J.C. Ensemble deep learning for speech recognition. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
38. Bashir, S.; Qamar, U.; Javed, M.Y. An ensemble based decision support framework for intelligent heart disease diagnosis. In Proceedings of the International Conference on Information Society (i-Society 2014), London, UK, 10–12 November 2014; pp. 259–264.
39. Alvarez, J.M.; LeCun, Y.; Gevers, T.; Lopez, A.M. Semantic road segmentation via multi-scale ensembles of learned features. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 586–595.
40. Han, M.; Zhu, X.; Yao, W. Remote sensing image classification based on neural network ensemble algorithm. *Neurocomputing* **2012**, *78*, 133–138. [CrossRef]

41. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]
42. Samadzadegan, F.; Hahn, M.; Bigdeli, B. Automatic road extraction from LIDAR data based on classifier fusion. In Proceedings of the 2009 Joint Urban Remote Sensing Event, Shanghai, China, 20–22 May 2009; pp. 1–6.
43. Pan, S.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
44. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
45. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
46. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
47. Varia, N.; Dokania, A.; Senthilnath, J. DeepExt: A Convolution Neural Network for Road Extraction using RGB images captured by UAV. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1890–1895.
48. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **2009**, *33*, 1–39. [CrossRef]
49. Heipke, C.; Mayer, H.; Wiedemann, C.; Jamet, O. Evaluation of automatic road extraction. *Int. Arch. Photogramm. Remote Sens.* **1997**, *32*, 151–160.
50. Available online: https://www.dropbox.com/sh/99cbjs6v73211fk/AABlmOeaPY6NAKUykqAc_E2ra (accessed on 8 July 2018)
51. Available online: https://www.dropbox.com/sh/w8e3a8j5eksfi7o/AADIqsM8Uy7XrceceR6x8NFoa?dl=0 (accessed on 8 July 2018)