drones

MDPI

Article

Multimodal Few-Shot Target Detection Based on Uncertainty Analysis in Time-Series Images

Mehdi Khoshboresh-Masouleh * and Reza Shah-Hosseini *

School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran 14399-57131, Iran

* Correspondence: m.khoshboresh@ut.ac.ir (M.K.-M.); rshahosseini@ut.ac.ir (R.S.-H.)

Abstract: The ability to interpret multimodal data, and map the targets and anomalies within, is important for an automatic recognition system. Due to the expensive and time-consuming nature of multimodal time-series data annotation in the training stage, multimodal time-series image understanding, from drone and quadruped mobile robot platforms, is a challenging task for remote sensing and photogrammetry. In this regard, robust methods must be computationally low-cost, due to the limited data on aerial and ground-based platforms, yet accurate enough to meet certainty measures. In this study, a few-shot learning architecture, based on a squeeze-and-attention structure, is proposed for multimodal target detection, using time-series images from the drone and quadruped robot platforms with a small training dataset. To build robust algorithms in target detection, a squeeze-and-attention structure has been developed from multimodal time-series images from limited training data as an optimized method. The proposed architecture was validated on three datasets with multiple modalities (e.g., red-green-blue, color-infrared, and thermal), achieving competitive results.

Keywords: multimodal; time-series images; robot; target detection; few-shot learning

Citation: Khoshboresh-Masouleh, M.; Shah-Hosseini, R. Multimodal Few-Shot Target Detection Based on Uncertainty Analysis in Time-Series Images. *Drones* **2023**, *7*, 66. https://doi.org/10.3390/drones7020066

Academic Editors: Giordano Teza, Massimo Fabris, Arianna Pesci and Tina Živec

Received: 29 December 2022
Revised: 13 January 2023
Accepted: 16 January 2023
Published: 17 January 2023

1. Introduction

In large-scale photogrammetry and remote sensing observations (e.g., drone imaging), visual target detection is a difficult issue in multimodal images due to the spectral similarities between the target and the background [1]. In real-world applications, the variety of geospatial information from urban areas can cause difficulties when creating and analyzing datasets, since it is hard to find the right method that matches their learning preferences.

Drone imaging is a good solution for static and dynamic target detection from different objects on an urban scale [2]. Drone and quadruped mobile robots equipped with multimodal sensors (e.g., RGB, color-infrared, and thermal) are an efficient and low-cost method for high-resolution scene understanding in real-world scenarios based on time-series images [3–5] such as crop and weed monitoring, traffic and vehicle management, and search and rescue missions for crisis management. In this regard, target detection is an important task in scene understanding from time-series images [6–8]. Finding an optimized algorithm with a small training dataset is the key challenge for target prediction and its real-world applications.

In drone imaging, a new encoder-decoder deep learning approach based on Fully Convolutional Network (FCN), Dilated Convolutional Neural Network (DCNN), U-Net, and MultiScaleDilation is proposed for multitarget prediction from oblique time-series images [9]. Moreover, a multitask learning method based on an encoder-decoder model is proposed for vehicle and building detection from multimodal drone images [10]. Gao et al. (2021) proposed a few-shot detector based on fully convolutional one-stage object detection (FCOS) for vehicle, storage tank, and plane detection in drone images [11]. Moreover, a new plant location and counting technology, based on a few-shot learning method that uses RGB images acquired from unmanned aerial vehicles (also known as

CenterNet), has been proposed. The quantitative assessments of this study showed that the average precision for a modified CenterNet architecture is about 95% [12]. A multi-stream framework and a deep multimodality learning method has been designed for drone video aesthetic quality assessment from spatial appearance, drone camera motion, and scene structure [13]. Lu and Koniusz (2022) proposed a versatile Few-Shot Keypoint Detection (FSKD) pipeline for unseen species based on uncertainty learning from RGB images [14]. In quadruped mobile robot imaging, an efficient study was proposed based on ERFNet, MAVNet, U-Net, Fast-SCNN, MFNet, and RTFNet from multimodal time-series images for fire-extinguishers, backpacks, hand-drills, and survivors [15]. The encoder block in CNNs for time-series images, represented by an encoding function $y = f(x)$, compresses the time-series input into a latent space, and the decoder block, $z = h(y)$, aims to predict the target from the latent space. Unal (2021) proposed a new visual target detection method based on the Kalman filter and deep learning for human tracking from RGB street-level images. In this study, the Kalman filter was used to estimate the position, velocity, and acceleration of a maneuvering target [16]. In Kiyak and Unal (2021), four deep learning models (i.e., deep convolutional neural networks (DCNN), deep convolutional neural networks with fine-tuning (DCNNFN), transfer learning with deep convolutional neural network (TLDCNN), and fine-tuning deep convolutional neural network with transfer learning (FNDCNNTL)) were proposed to detect small aircraft [17]. In Han et al. (2022), a few-shot learning method (also known as FSOD) was applied to object detection from RGB street-level images based on meta-learning with cross-modal representations [5]. To achieve a stable target detection accuracy regardless of the sensing devices' viewpoint and orientation (drone or quadruped mobile robot), a multimodal few-shot learning method can be utilized [18–22]. According to the related literature, several investigations have demonstrated that there are still some crucial problems, such as: (1) poor model performance due to the lack of pixel-wise training in few-shot learning models in real-world applications, which have not yet been well considered in the relevant studies; and (2) poor generalization ability in different platforms due to the use of single-modal data for training.

In this study, a few-shot machine learning architecture based on squeeze-and-attention structure [23] is proposed for multimodal target detection based on time-series images from drone and quadruped mobile robot platforms with a small training dataset. To build robust algorithms in target detection, a squeeze-and-attention structure has been developed from multimodal time-series data from limited training data as an optimized method. We aim to fill some of the gaps in the supervised method with small training data in the field of target detection of multimodal time-series data from drone and quadruped mobile robot platforms. We summarize our contributions as follows:

1. A few-shot learning model based on an uncertainty analysis of feature extraction is proposed with an encoder-decoder structure and squeeze-and-attention module. The proposed model is composed of two components in the encoder block, including the residual representation extraction and the attention layers;
2. It proposes a novel approach to extracting inherent and latent representation from multimodal images;
3. We conducted several multimodal datasets for different real-world scenarios to investigate the behavior of the proposed few-shot learning method.

2. Materials and Methods

2.1. Definition of the Problem

Despite recent breakthroughs in drone imaging and few-shot learning, confident and robust target detection remains a challenge for remote sensing engineers [20,24]. The purpose of static and dynamic target detection from drone images is to locate the target on large-scale images by using pixel-wise segmentation [25]. Figure 1 shows a comparison of oblique and vertical views from drone and quadruped mobile robot platforms. In Figure 1, targets at different distances need to be processed in different scale spaces and spatial-spectral representations. Due to the expensive and time-consuming nature of

multimodal time-series data annotation in the training stage, multimodal time-series image understanding from drone and quadruped mobile robot platforms is a challenging task for remote sensing and photogrammetry [26]. In this regard, robust methods must be computationally low-cost, due to the limited data on aerial and ground-based platforms, yet accurate enough to meet certainty measures.

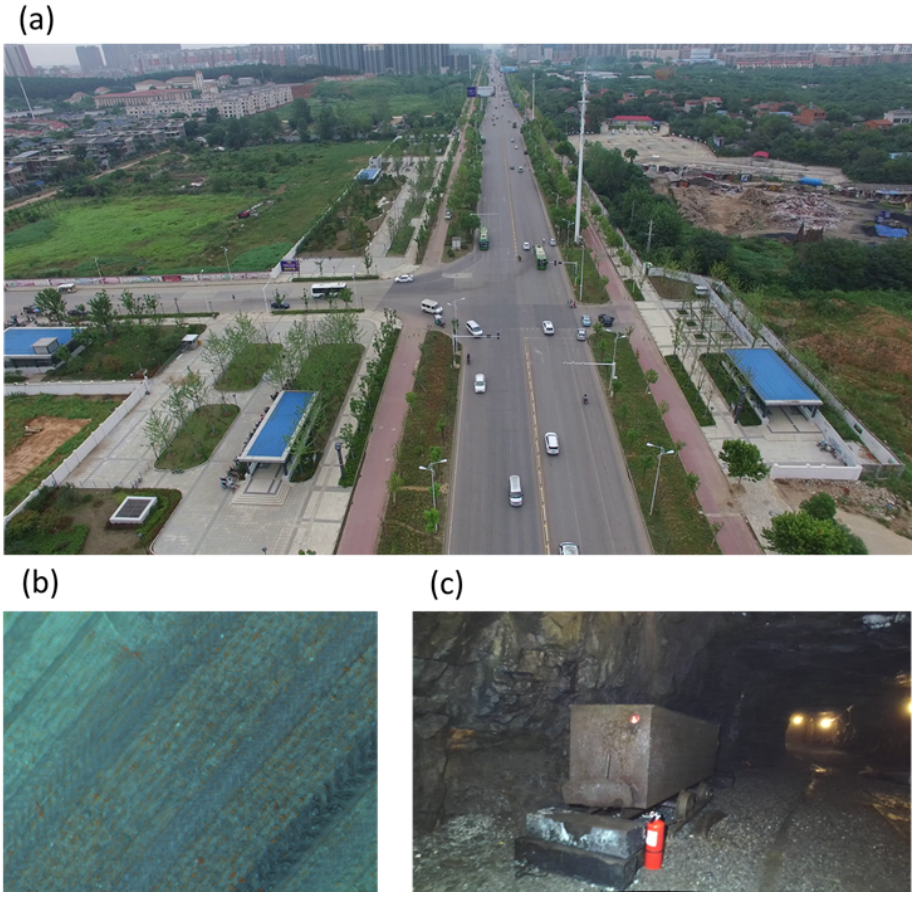

Figure 1. Oblique and vertical views in drone and quadruped mobile robot platforms. (**a**) Drone-based red-green-blue (RGB) image with an oblique view; (**b**) Drone-based col-or-infrared (CIR) image with a vertical view; (**c**) Quadruped mobile robot-based RGB image with an oblique view.

Target localization with the use of uncertainty modeling in few-shot learning for drone images may improve scene understanding from a limited training dataset, while many target detection methods appear to understand single-time localization with large amounts of training data [27]. Few-shot learning can perform unseen tasks after training on a few labeled images and can consider several tasks to produce a predictive function; it is also an inductive transfer system, whose main purpose is to improve generalization ability for multiple tasks [21]. Target localization of the trained network can be assumed as accurate, but not for decision-making in real-world applications [28–30]. For example, timely weed detection in farming tasks is crucial to obtain high-quality crops. In this regard, the uncertainty estimation map should be a crucial stage of the predictive map.

In recent years, the number of studies on the civil applications of few-shot learning in image processing has rapidly increased. A general few-shot learning approach is a pre-trained convolutional neural network [31] which is trained on a large dataset and fine-tuned on another task. A deep network trained on the iSAID dataset [32] can be used in another instance segmentation, since the iSAID contains over 655,451 object instances. This approach is known as transfer learning, and it requires a large training dataset and parameters in order to fine-tune [33]. Transfer learning includes the transfer of knowledge from a preliminary task to a newer task that has many similarities with the old one [34]. In

meta-learning [35], in order to learn from a limited amount of training data with annotated images, a new strategy called few-shot learning was proposed. In this regard, few-shot learning models were developed, and they have since proven to be robust methods which use only limited training data.

A model-agnostic meta-learning approach has been developed for image segmentation [36]. The model-agnostic meta-learning extends gradient descent by optimizing for a model initialization that leads to good performance on a set of related tasks. A gated encoder-decoder convolutional neural network has been proposed for pixel-wise weed detection from multispectral drone images [37]. Although these related models are fairly powerful for some target localization tasks, their performance in confidential static and dynamic target detection is, to the best of our knowledge, still not excellent.

2.2. Proposed Few-Shot Learning Network

A graphical summary of the proposed few-shot learning network for robust target localization is presented in Figure 2. The proposed model contains the following three sets, where each set contains M images: a training set $D_{train} = \{x_i, y_i\}_{i=1}^{M}$, $x_i : D \rightarrow \mathbb{R}^3$, an input image, and $y_i : D \rightarrow \{0, 1\}$, its corresponding ground truth map; a support set $D_{support} = \{x_i, y_i\}_{i=1}^{M_{support}}$; and a test set $D_{test} = \{x_i\}_{i=1}^{M_{test}}$. The proposed model takes the DMF-Net [37], the SA-Net [38], and the MF-Net [26] as the main backbone model for weed detection, vehicle monitoring, and human detection.

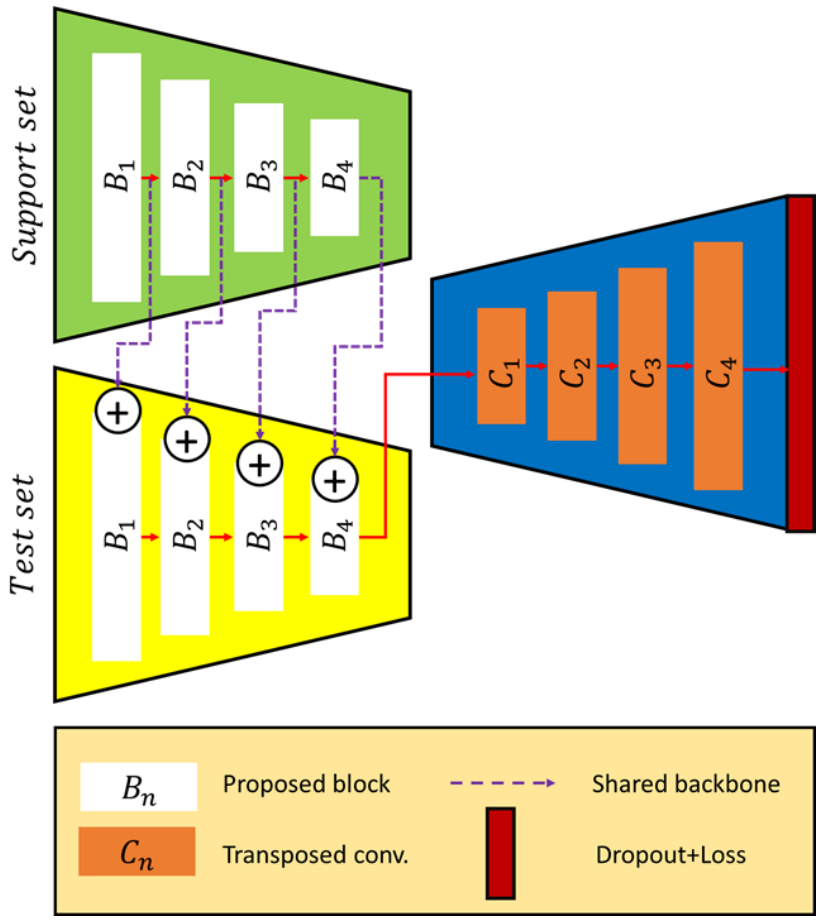

Figure 2. Overview of the proposed network for multimodal target prediction from time-series images. The network takes time-series multimodal images as a support set and test set, and then outputs a target detection map.

In this study, we designed and trained DMF-Net, SA-Net, and MF-Net for learning augmentation and for improving the generalization ability of the trained model. DMF-Net is a lightweight gated convolutional neural network model with a model size of 70 MB. This model is trained using a lightweight dataset that includes 150 samples of multispectral drone images with a size of 480 × 360 pixels. The gated CNN was efficient at minimizing the unnecessary transmission of data by using a convolutional layer for extracting optimized single-date or time-series images. SA-Net and MF-Net are few-shot learning models based on the region-based temporal aggregation Monte Carlo dropout [39], which can further improve the uncertainty modeling to guide vehicle and human monitoring.

Our model is different from similar networks in multiscale feature extraction likewise based on convolutional neural networks, since it overcomes the constraints of convolutional blocks and binary mask generation for static and dynamic target localization. The proposed method utilizes a pair of encoders and decoders. The encoder is composed of four squeeze-and-attention modules and the decoder is composed of one transpose convolution module and interpolation, which learns non-local spatial-spectral features. To aggregate multistep non-local representations, we adopted four squeeze-and-attention modules on the multistep outputs of the backbone model, resulting in better target boundaries. In the proposed method, we focus on region-based temporal aggregation Monte Carlo dropout as the uncertainty estimator for target localization. We propose an encoder-decoder network based on a new squeeze-and-attention block B_n, and a transposed coevolution T_D for target detection from time-series images with different platforms for few-shot learning (Figure 2). The convolutional layers include 5 × 5 kernels, which are applied to the input representation maps using stride one. The proposed few-shot learning network is composed of 10 squeeze-and-attention layers followed by batch normalization (*Bt*) and rectified linear unit (*ReL*) functions to generate feature maps, as well as max-pooling layers to reduce the size of feature maps (Table 1).

In action, we define a training dataset T_D employed for the training step, a test dataset Q_D, and a support set S_D employed for the testing step.

We summarize the proposed network's components as follows:

Gated Module. This module is based on a guided feature extractor, feature fusion layer, dilation CNN, inception layer, and encoder-decoder blocks. The gated module extracts the object-level representation from time-series images from multimodal data. We used the DMF $G_i = [Bt(L \otimes k_{3\times 3})] \odot [UP(U \otimes k_{3\times 3})]$ as a gated module [37];

Squeeze-and-Attention Layer. To aggregate multi-scale non-local representations, we adopt squeeze-and-attention layers $B = U(RL(f_{att}(Pool(I))) \times x_{res} + U(ReLU(f_{att}(Pool(I)))$ on the multi-scale outputs of the gated module model, resulting in better target boundaries;

Weighted Binary Cross-Entropy. This loss function $E(cp, \hat{cp}) = -(\gamma \log(\hat{cp}) + (1 - cp) \log(1 - \hat{cp}))$ is used for the issue of the target imbalance in which all positive pixels get weighted by an amount close to one;

System Implementation Details. The proposed method was trained, using PyTorch, on a single NVIDIA TESLA K80 with a batch size of 12 for 150 epochs for target prediction. The learning rate and momentum were about 10^{-2} and 0.9, respectively, for the stochastic gradient descent approach;

Backbones. To train the proposed network from three multimodal datasets for target detection, two backbones were selected. The experiments were carried out with DMF-Net, SA-Net, and MF-Net backbones.

2.3. Datasets

The experimental areas include oblique and vertical images from drone and quadruped mobile robot platforms. Table 2 shows the details of multimodal datasets for target detection.

Table 1. The proposed architecture for multimodal target detection. n@C × R: number of channels, columns, and rows.

Encoder	Components	Size	Stride	Result
Input	-	-	-	$n@C_{Input} \times R_{Input}$
B_1	CV1 + Bt + ReL	64@5 × 5	1	64@C × R
	CV2 + Bt + ReL + B	64@5 × 5	1	64@C × R
	MP	2 × 2	2	-
B_2	CV3 + Bt + ReL	128@5 × 5	1	128@C × R
	CV4 + Bt + ReL + B	128@5 × 5	1	128@C × R
	MP	2 × 2	2	-
B_3	CV5 + Bt + ReL	256@5 × 5	1	256@C × R
	CV6 + Bt + ReL	256@5 × 5	1	256@C × R
	CV7 + Bt + ReL + B	128@5 × 5	1	256@C × R
	MP	2 × 2	2	-
B_4	CV8 + Bt + ReL	512@5 × 5	1	512@C × R
	CV9 + Bt + ReL	512@5 × 5	1	512@C × R
	CV10 + Bt + ReL + B	512@5 × 5	1	512@C × R
	MP	2 × 2	2	-
Decoder				
C_1	UP	2 × 2	2	-
	DCV1 + Bt + ReL	1024@5 × 5	1	1024@C × R
	DCV2 + Bt + ReL	1024@5 × 5	1	1024@C × R
	DCV3 + Bt + ReL	512@5 × 5	1	512@C × R
C_2	UP	2 × 2	2	-
	DCV4 + Bt + ReL	512@5 × 5	1	512@C × R
	DCV5 + Bt + ReL	512@5 × 5	1	512@C × R
	DCV6 + Bt + ReL	256@5 × 5	1	256@C × R
C_3	UP	2 × 2	2	-
	DCV7 + Bt + ReeL	256@5 × 5	1	256@C × R
	DCV8 + Bt + RL	256@5 × 5	1	256@C × R
	DCV9 + Bt + ReL	128@5 × 5	1	128@C × R
C_4	UP	2 × 2	2	-
	DCV10 + Bt + ReL	128@5 × 5	1	128@C × R
	DCV11 + Bt + ReL	64@5 × 5	1	64@C × R
Output	UP	2 × 2	2	-
	DCV12 + Bt + ReL	64@5 × 5	1	64@C × R
	DCV13 + Bt + ReL	2@5 × 5	1	2@C × R
	Loss	-	-	-

Table 2. Details of datasets for multimodal target prediction.

Dataset	Modality	Patch Size (Pixels)	Samples		
			Train	Support	Test
UAVid	RGB	3840 × 2160	500	55	45
WeedMap	CIR	480 × 360	320	100	60
PST900	Thermal	1280 × 720	190	100	60

In this study, three datasets from different scenarios (Table 3), including vehicle detection, weed mapping, and human detection, were utilized to evaluate the proposed method. The UAVid dataset [9] consists of RGB images with an oblique view from a drone platform for vehicle detection. The UAVid dataset consists of a time-series dataset targeting semantic labeling for urban scene analysis from an oblique drone perspective. The characteristics that make the UAVid dataset a standard dataset are (1) the time-series

high-resolution images and (2) the different landscape types, including different types of vehicles. The WeedMap dataset [40] consists of color-inferred images with a vertical view from a Sequoia sensor located on a Mavic Pro platform in Eschikon, Switzerland, and it is used as an example to study crop and weed detection assessment. Crops on the Eschikon fields were sowed on 5 April 2017 and were arranged in 50 cm rows. Their growth stage was about one month at the moment of data collection, and the sizes of the crops and weeds ranged from 5 to 10 cm. The PST900 dataset [15] consists of multimodal images (thermal) with an oblique view from a quadruped mobile robot platform. This dataset comprises synchronized and calibrated thermal time-series images with a size of 1280 × 720 pixels for real-time target detection.

Table 3. List of datasets for target detection.

Dataset	Type	View	Data Source	Target Class
UAVid	Video	Oblique	Drone	Vehicle
WeedMap	Orthophoto	Vertical	Drone	Weed
PST900	Video	Oblique	Quadruped robot	Human

3. Results

3.1. Metrics

To evaluate the predicted targets and anomalies, the standard quality measures of the Jaccard index (*J*), correctness (C), and entropy (*E*) of the probabilities p_n associated with the histogram of input I were used, and they are computed as follows:

$$J = \frac{True\ Positive}{True\ Positive + False\ Negative + False\ Positive} \quad (1)$$

$$C = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$E = -\sum_{n=0}^{255} p_n(I).\log_2(p_n(I)) \quad (3)$$

3.2. Uncertainty Analysis of Feature Extraction

In this subsection, we evaluate the qualitative uncertainty of the feature extraction step for multimodal images. In action, the CNN models by learning the geometric shapes in the first layer and then evolving to learn representations of the input data in the deeper blocks, resulting in a more accurate target prediction. The proposed method consists of the input layer (multimodal data), hidden block (the proposed layer), and output (target), whereas the hidden block can include a multilayer of low-level properties to high-level properties. Figure 3 shows the difference between low-level and high-level representation extraction by the proposed model.

3.3. Exprimental Results

To evaluate the performance of the trained network for each scenario, a test set and a support set were selected outside the train set. The results of the multimodal targets and anomalies are shown in Figure 4. Vegetation regions and shadow regions in the test set are an important challenge in target localization. To improve the training set, we added time-series data that included various grass and tree covers to make the model robust for target prediction in the presence of vegetative regions. Moreover, the proposed method was trained using shadow-included multimodal time-series images from the various drone and quadruped robot images to learn how the model tackles that problem.

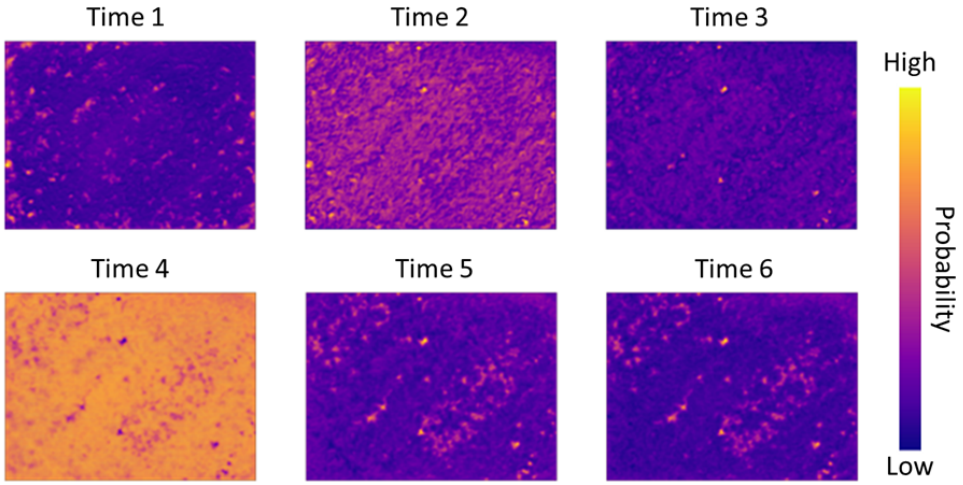

Figure 3. Examples of uncertainty maps learned by the proposed method for weed and crop mapping from the color-infrared images. This example shows the performance of the proposed network on time-series images. Higher uncertainty at crop/weed boundaries was observed.

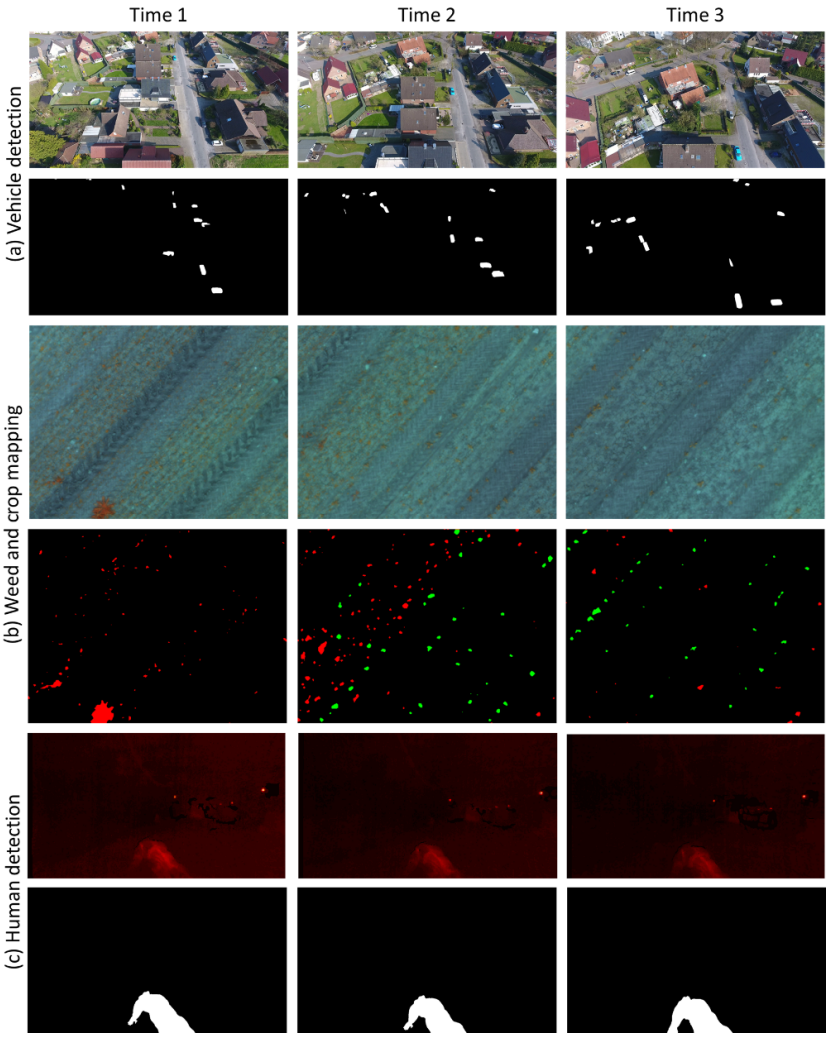

Figure 4. Examples of target prediction from drone and quadruped robot time-series images by the proposed model. (**a**) Vehicle detection from RGB images; (**b**) Weed (red) and crop (green) mapping with color-infrared images; (**c**) Human detection from thermal images.

Table 4 shows the numerical predictions of the multimodal test performed using the proposed method compared to the different Jaccard index and entropy metrics. As shown in Table 4, the Jaccard index and entropy metrics have values of 89.1% and 31.30% for vehicle detection from RGB images, 92.83% and 24.33% for crop and weed mapping from color-infrared images, and 85.40% and 34.00% for human detection from thermal images, respectively. The low computational cost required by the proposed model to classify a given multimodal image makes feasible its integration into real-time applications such as drones.

Table 4. The quality metrics for the predicted targets of multimodal datasets.

Data	Target	Samples	Backbone	J (%)	E	T (ms) [1]
UAVid	Vehicle detection	Set-1, n = 15	SA-Net	83.4	0.41	80
		Set-2, n = 15		89.3	0.32	76
		Set-3, n = 15		94.6	0.21	78
WeedMap	Weed and crop mapping	Set-1, n = 20	DMF-Net	89.6	0.28	59
		Set-2, n = 20		96.9	0.19	62
		Set-3, n = 20		92.0	0.26	49
PS500	Human detection	Set-1, n = 20	MF-Net	83.2	0.37	43
		Set-2, n = 20		87.7	0.36	48
		Set-3, n = 20		85.3	0.29	39

[1] On an NVIDIA Tesla K80.

3.4. Comparative Study of Few-Shot Learning and Classical Methods

In this section, two classical methods, including Support Vector Machine (SVM) [41] and Random Forest (RF) [42], were used for comparisons. For a fair comparison, all methods were trained in this study from the beginning using the same training dataset (Section 2.3) and feature extraction method (Section 2.2) that was applied for the training of the proposed few-shot learning model. The quantitative assessments of this study show that the average IoU scores for the proposed method, SVM, and RF are about 78%, 47%, and 39%, respectively. Moreover, the qualitative comparison of the proposed few-shot learning model prediction with SVM and RF is shown in Figure 5. The qualitative results show the ability of the proposed few-shot learning model to detect smaller target regions in a scene while producing a confident result.

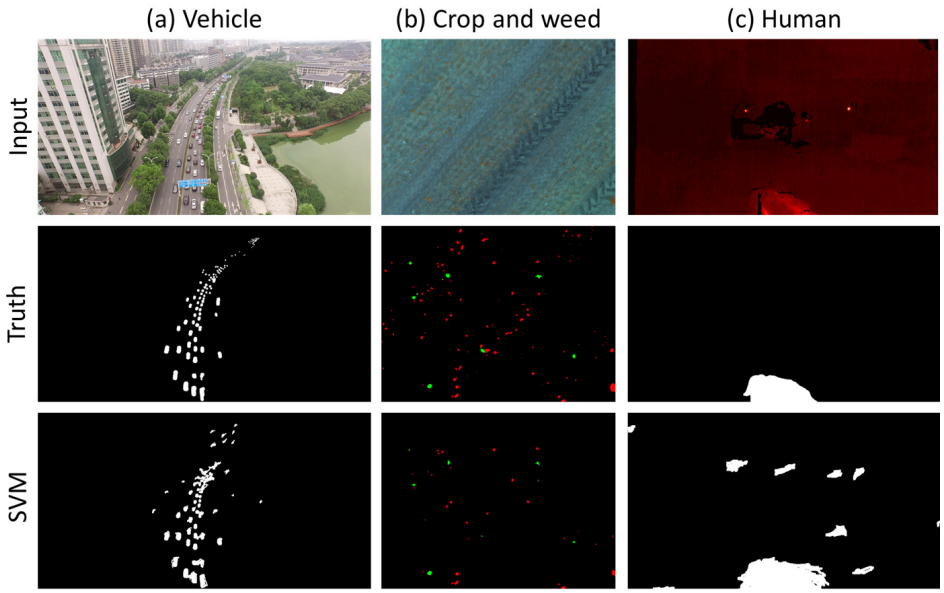

Figure 5. *Cont.*

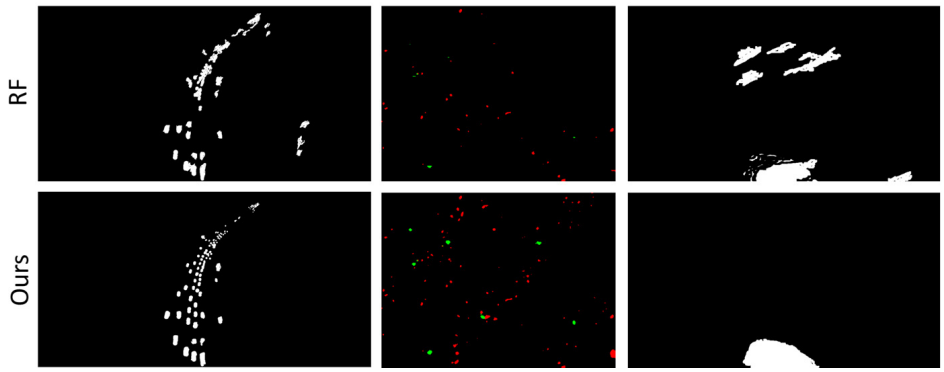

Figure 5. Qualitative assessment of the target detection with SVM, RF, and the proposed few-shot learning method. (**a**) Vehicle detection from obliques RGB images; (**b**) Weed (red) and crop (green) detection with color-infrared images; (**c**) Human detection from thermal images.

4. Discussion

Table 5 shows the quantitative results for the weed and vehicle detection from the drone images. In this study, we compared the proposed method against the new method PFE-Net [32]. A visualization of static and dynamic target detection results of the test stage is shown in Figures 6–8. The proposed model achieves mean correctness for the vehicle, weed, and human detection of 88.4%, 82.2%, and 86.25% for the 45 RGB, 60 CIR, and 60 thermal tested images, respectively, while the mean entropy for the vehicle, weed, and human detection is 21.9%, 23.7%, and 20.75%, respectively. As a result, there is an inverse relationship between IoU and entropy value in target detection from large-scale remote sensing images. Experimental results verified this point for the proposed few-shot learning model. Three unique aspects of this work are:

1. Our findings indicate that the proposed few-shot learning model offers better generalization performance and outperforms other methods. The proposed few-shot learning model achieved the highest mIoU in comparison to other evaluated methods, such as PFE-Net (as a few-shot learning model), SVM, and RF, for each dataset in multimodal images. Therefore, compared with the state-of-the-art and classical models, the target detection accuracy was improved;
2. To our knowledge, we have presented the first multimodal few-shot learning method with a low computational cost for RGB, CIR, and thermal modalities based on uncertainty estimation;
3. The proposed method uses unique features of multimodal images during the feature extraction from squeeze-and-attention layers.

Table 5. Static and dynamic target detection comparisons on different types of test scenes from the proposed method and PFE-Net.

				IoU		**C**		**E**	
Methods	Scenario	Backbone	Samples	1 shot	10 shot	1 shot	10 shot	1 shot	10 shot
Ours	Vehicle	SA-Net	45	46.7	77.2	83.7	93.1	26.4	17.3
PFE-Net				47.2	73.1	80.1	86.4	34.1	24.8
Ours	Weed	DMF-Net	60	51.5	89.4	73.2	91.2	29.1	18.3
PFE-Net				51.9	84.2	72.1	83.5	36.3	27.1
Ours	Human	MF-Net	60	63.4	91.5	79.4	93.1	24.1	17.4
PFE-Net				57.3	87.3	74.3	83.4	34.8	29.6

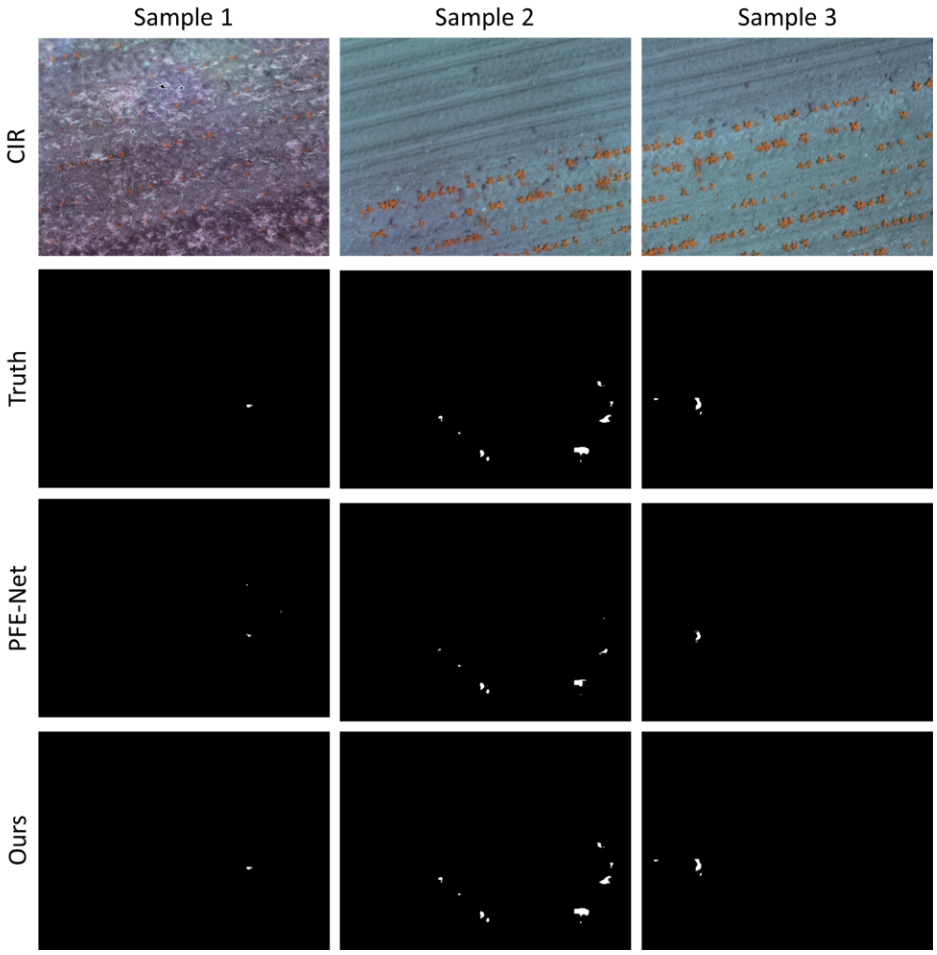

Figure 6. The proposed method's predictions for weed detection scene test samples.

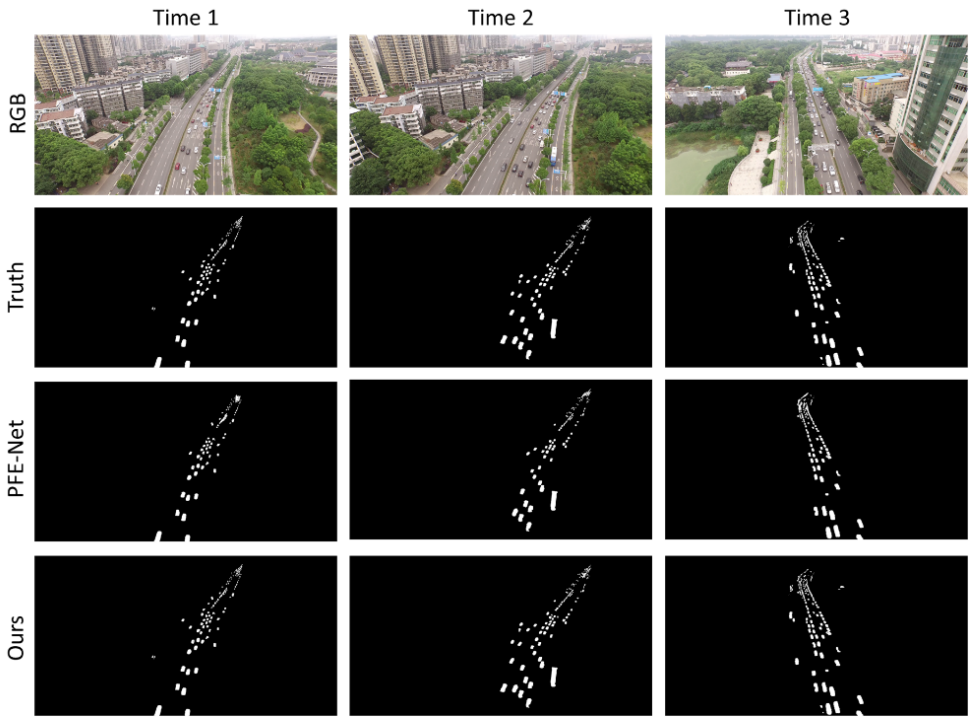

Figure 7. The proposed method's predictions for vehicle detection test samples.

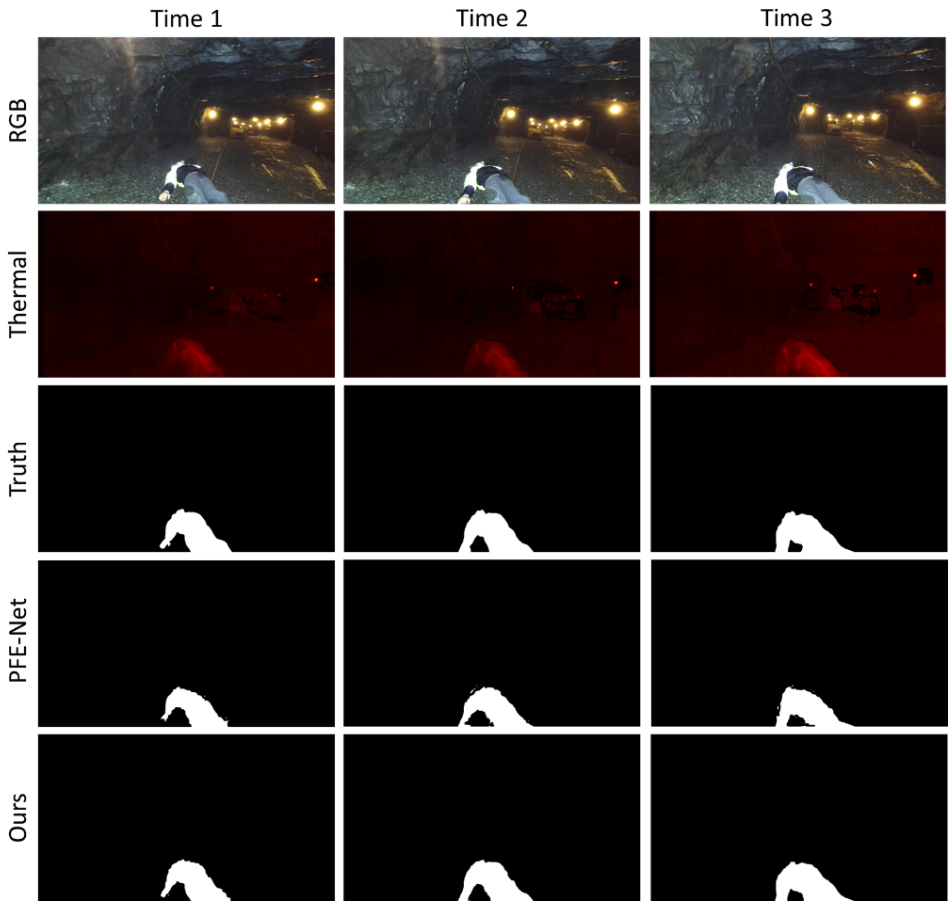

Figure 8. The proposed method's predictions for human detection test samples.

In this subsection, we present an ablation study to compare some different model variants, such as different loss functions for training, multiple modalities, and CNN backbones, and justify our design choices. Table 6 shows some ablation study results to investigate the behavior of the proposed method.

4.1. Loss Functions for Training

In this subsection, all models were evaluated with different loss functions. Although the proposed method with a Weighted Cross entropy (WeC) loss function delivers good results, the proposed method was tested with another two loss functions, consisting of Dice Loss (DiL) [43] and Weighted Bootstrapped Cross-entropy (WBC) [44]. We trained our network using the absolute error between the ground truth map and the model's prediction.

4.2. Multiple Modalities

The proposed target detection method was tested with different data modalities, such as red, green, blue, near-infrared, and red edge channels.

4.3. CNN Backbones

The proposed model can be set up with different backbones for target detection. We selected two backbones for the ablation study. The experiments were carried out with the ResNet-101 and HRNet.v2 [45] backbones.

Table 6. Ablation study on using different loss functions, modalities, and backbones for the proposed method.

Feature Extractor	Backbone	Loss	Modalities	IoU	C	E
Proposed layer ℓ_{out}	ResNet-101	Dice	RGB (Vehicle)	64.2	84.2	23.4
	HRNet.v2			63.1	81.3	24.6
	ResNet-101	WBC		70.2	84.3	21.6
	HRNet.v2			69.1	82.1	22.7
	ResNet-101	WeC		73.5	85.4	19.8
	HRNet.v2			76.1	87.4	18.2
	ResNet-101	Dice	CIR (Weed)	79.3	84.1	21.4
	HRNet.v2			82.5	84.9	20.7
	ResNet-101	WBC		83.4	85.2	19.8
	HRNet.v2			84.1	86.6	18.9
	ResNet-101	WeC		86.2	89.3	18.7
	HRNet.v2			87.1	90.2	18.5
	ResNet-101	Dice	Red + Green + Red edge + Near-infrared (Weed)	82.4	87.2	19.6
	HRNet.v2			85.5	88.3	18.5
	ResNet-101	WBC		84.2	88.7	18.3
	HRNet.v2			86.8	88.5	18.2
	ResNet-101	WeC		87.8	90.7	18.6
	HRNet.v2			88.0	91.0	17.9
	ResNet-101	Dice	Thermal	71.3	84.2	26.1
	HRNet.v2			74.2	85.5	20.3
	ResNet-101	WBC		70.3	83.1	27.3
	HRNet.v2			72.9	85.3	24.7
	ResNet-101	WeC		83.4	89.4	18.3
	HRNet.v2			86.7	90.6	17.6

5. Conclusions

Target localization from drone and quadruped robot platforms has become the dominant tool for real-world applications such as vehicle monitoring for traffic management, weed mapping for smart farming, and human detection for search and rescue missions. Target segmentation with the use of uncertainty estimation in few-shot learning for drone images can potentially improve scene understanding with a small training dataset, while many target detection methods appear to understand single-time localization with a big training dataset. This paper presents a new few-shot learning architecture for some core remote sensing problems. Our method outperforms the state-of-the-art models on many challenging remote sensing datasets.

In this study, we have performed three multimodal experiments for target detection from the drone and quadruped robot time-series images based on a new squeeze-and-attention method for few-shot learning from a small training dataset. The method has achieved impressive performances in various real-world tasks, which can be grouped into three categories: weed and crop mapping; vehicle and traffic monitoring; and human tracking. To improve the generalization of the trained models for 3D target reconstruction, model architecture needs to be developed in future studies.

Author Contributions: Conceptualization, M.K.-M.; completed the implementation of the scheme, M.K.-M. and R.S.-H.; writing—original draft preparation, M.K.-M.; writing—review and editing, M.K.-M. and R.S.-H.; theoretical guidance, R.S.-H.; project administration, M.K.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The UAVid dataset (images and labels) is available at [9]: https://uavid.nl/ (accessed on 19 March 2022); The WeedMap dataset (images and labels) is available at [40]: https://projects.asl.ethz.ch/datasets/doku.php?id=weedmap:remotesensing2018weedmap (accessed on 19 March 2022); The PST900 dataset (images and labels) is available at [15]: https://drive.google.com/file/d/1hZeM-MvdUC_Btyok7mdF00RV-InbAadm/view (accessed on 19 March 2022).

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

References

1. Bi, F.; Sun, X.; Liu, W.; Sun, P.; Wang, H.; Sun, J. Multiscale Anti-Deformation Network for Target Tracking in UAV Aerial Videos. *JARS* **2022**, *16*, 022207. [CrossRef]
2. Lv, J.; Cao, Y.; Wang, X.; Yun, Z. Vehicle Detection Method for Satellite Videos Based on Enhanced Vehicle Features. *JARS* **2022**, *16*, 026503. [CrossRef]
3. Touil, A.; Ghadi, F.; El Makkaoui, K. Intelligent Vehicle Communications Technology for the Development of Smart Cities. In *Machine Intelligence and Data Analytics for Sustainable Future Smart Cities*; Ghosh, U., Maleh, Y., Alazab, M., Pathan, A.-S.K., Eds.; Studies in Computational Intelligence; Springer International Publishing: Cham, Switzerland, 2021; pp. 65–84, ISBN 978-3-030-72065-0.
4. Faraj, F.; Braun, A.; Stewart, A.; You, H.; Webster, A.; Macdonald, A.J.; Martin-Boyd, L. Performance of a Modified YOLOv3 Object Detector on Remotely Piloted Aircraft System Acquired Full Motion Video. *JARS* **2022**, *16*, 022203. [CrossRef]
5. Han, G.; Ma, J.; Huang, S.; Chen, L.; Chellappa, R.; Chang, S.-F. Multimodal Few-Shot Object Detection with Meta-Learning Based Cross-Modal Prompting. *arXiv* **2022**, arXiv:2204.07841.
6. Khoshboresh-Masouleh, M.; Shah-Hosseini, R. 2D Target/Anomaly Detection in Time Series Drone Images Using Deep Few-Shot Learning in Small Training Dataset. In *Integrating Meta-Heuristics and Machine Learning for Real-World Optimization Problems*; Houssein, E.H., Abd Elaziz, M., Oliva, D., Abualigah, L., Eds.; Studies in Computational Intelligence; Springer International Publishing: Cham, Switzerland, 2022; pp. 257–271, ISBN 978-3-030-99079-4.
7. Ma, R.; Angryk, R. Distance and Density Clustering for Time Series Data. In Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017; pp. 25–32.
8. Ma, R.; Ahmadzadeh, A.; Boubrahimi, S.F.; Angryk, R.A. Segmentation of Time Series in Improving Dynamic Time Warping. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 3756–3761.
9. Lyu, Y.; Vosselman, G.; Xia, G.-S.; Yilmaz, A.; Yang, M.Y. UAVid: A Semantic Segmentation Dataset for UAV Imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
10. Bayanlou, M.R.; Khoshboresh-Masouleh, M. Multi-Task Learning from Fixed-Wing UAV Images for 2D/3D City Modelling. *arXiv* **2021**, arXiv:2109.00918. [CrossRef]
11. Gao, Y.; Hou, R.; Gao, Q.; Hou, Y. A Fast and Accurate Few-Shot Detector for Objects with Fewer Pixels in Drone Image. *Electronics* **2021**, *10*, 783. [CrossRef]
12. Karami, A.; Crawford, M.; Delp, E.J. Automatic Plant Counting and Location Based on a Few-Shot Learning Technique. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5872–5886. [CrossRef]
13. Kuang, Q.; Jin, X.; Zhao, Q.; Zhou, B. Deep Multimodality Learning for UAV Video Aesthetic Quality Assessment. *IEEE Trans. Multimed.* **2020**, *22*, 2623–2634. [CrossRef]
14. Lu, C.; Koniusz, P. Few-Shot Keypoint Detection with Uncertainty Learning for Unseen Species. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19394–19404.
15. Shivakumar, S.S.; Rodrigues, N.; Zhou, A.; Miller, I.D.; Kumar, V.; Taylor, C.J. PST900: RGB-Thermal Calibration, Dataset and Segmentation Network. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 9441–9447.
16. Unal, G. Visual Target Detection and Tracking Based on Kalman Filter. *J. Aeronaut. Space Technol.* **2021**, *14*, 251–259.
17. Kiyak, E.; Unal, G. Small Aircraft Detection Using Deep Learning. *AEAT* **2021**, *93*, 671–681. [CrossRef]
18. Moon, J.; Le, N.A.; Minaya, N.H.; Choi, S.-I. Multimodal Few-Shot Learning for Gait Recognition. *Appl. Sci.* **2020**, *10*, 7619. [CrossRef]
19. Bodor, R.; Drenner, A.; Fehr, D.; Masoud, O.; Papanikolopoulos, N. View-Independent Human Motion Classification Using Image-Based Reconstruction. *Image Vis. Comput.* **2009**, *27*, 1194–1206. [CrossRef]

20. Hu, Y.; Chen, M.; Saad, W.; Poor, H.V.; Cui, S. Distributed Multi-Agent Meta Learning for Trajectory Design in Wireless Drone Networks. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 3177–3192. [CrossRef]
21. Nishino, Y.; Maekawa, T.; Hara, T. Few-Shot and Weakly Supervised Repetition Counting With Body-Worn Accelerometers. *Front. Comput. Sci.* **2022**, *4*, 925108. [CrossRef]
22. Sugimoto, M.; Shimada, S.; Hashizume, H. RefRec+: Six Degree-of-Freedom Estimation for Smartphone Using Floor Reflecting Light. *Front. Comput. Sci.* **2022**, *4*, 856942. [CrossRef]
23. Zhong, Z.; Lin, Z.Q.; Bidart, R.; Hu, X.; Daya, I.B.; Li, Z.; Zheng, W.-S.; Li, J.; Wong, A. Squeeze-and-Attention Networks for Semantic Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13062–13071.
24. Li, H.; Wu, L.; Niu, Y.; Wang, C.; Liu, T. Small Sample Meta-Leaming Towards Object Recognition Through UAV Observations. In Proceedings of the 2019 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 17–19 October 2019; pp. 860–865.
25. Tan, S.; Yan, J.; Jiang, Z.; Huang, L. Approach for Improving YOLOv5 Network with Application to Remote Sensing Target Detection. *JARS* **2021**, *15*, 036512. [CrossRef]
26. Khoshboresh-Masouleh, M.; Shah-Hosseini, R. Real-Time Multiple Target Segmentation with Multimodal Few-Shot Learning. *Front. Comput. Sci.* **2022**, *4*, 1062792. [CrossRef]
27. Khoshboresh-Masouleh, M.; Shah-Hosseini, R. Uncertainty Estimation in Deep Meta-Learning for Crop and Weed Detection from Multispectral UAV Images. In Proceedings of the 2022 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Istanbul, Turkey, 7–9 March 2022; pp. 165–168.
28. Kendall, A.; Cipolla, R. Modelling Uncertainty in Deep Learning for Camera Relocalization. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 4762–4769.
29. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Advances in Neural Information Processing Systems. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
30. Kendall, A.; Badrinarayanan, V.; Cipolla, R. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *arXiv* **2015**, arXiv:1511.02680.
31. Stow, E.; Kelly, P.H.J. Convolutional Kernel Function Algebra. *Front. Comput. Sci.* **2022**, *4*, 921454. [CrossRef]
32. Zamir, S.W.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Khan, F.S.; Zhu, F.; Shao, L.; Xia, G.-S.; Bai, X. ISAID: A Large-Scale Dataset for Instance Segmentation in Aerial Images. *arXiv* **2019**, arXiv:1905.12886.
33. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
34. Wei, Y.; Zhang, Y.; Huang, J.; Yang, Q. Transfer Learning via Learning to Transfer. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5085–5094.
35. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
36. Gao, K.; Liu, B.; Yu, X.; Zhang, P.; Tan, X.; Sun, Y. Small Sample Classification of Hyperspectral Image Using Model-Agnostic Meta-Learning Algorithm and Convolutional Neural Network. *Int. J. Remote Sens.* **2021**, *42*, 3090–3122. [CrossRef]
37. Khoshboresh-Masouleh, M.; Akhoondzadeh, M. Improving Weed Segmentation in Sugar Beet Fields Using Potentials of Multispectral Unmanned Aerial Vehicle Images and Lightweight Deep Learning. *JARS* **2021**, *15*, 034510. [CrossRef]
38. Khoshboresh-Masouleh, M.; Shah-Hosseini, R. Deep Few-Shot Learning for Bi-Temporal Building Change Detection. *arXiv* **2021**, arXiv:2108.11262. [CrossRef]
39. Huang, P.-Y.; Hsu, W.-T.; Chiu, C.-Y.; Wu, T.-F.; Sun, M. Efficient Uncertainty Estimation for Semantic Segmentation in Videos; In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany. 2018. pp. 520–535.
40. Sa, I.; Popović, M.; Khanna, R.; Chen, Z.; Lottes, P.; Liebisch, F.; Nieto, J.; Stachniss, C.; Walter, A.; Siegwart, R. WeedMap: A Large-Scale Semantic Weed Mapping Framework Using Aerial Multispectral Imaging and Deep Neural Network for Precision Farming. *Remote Sens.* **2018**, *10*, 1423. [CrossRef]
41. Das, S. Image-Segmentation-Using-SVM. Available online: https://github.com/SIdR4g/Semantic-Segmentation-using-SVM (accessed on 11 January 2023).
42. Trainable Segmentation Using Local Features and Random Forests—Skimage v0.19.2 Docs. Available online: https://scikit-image.org/docs/stable/auto_examples/segmentation/plot_trainable_segmentation.html (accessed on 11 January 2023).
43. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *arXiv* **2017**, arXiv:1707.03237. [CrossRef]

44. Gaj, S.; Ontaneda, D.; Nakamura, K. Automatic Segmentation of Gadolinium-Enhancing Lesions in Multiple Sclerosis Using Deep Learning from Clinical MRI. *PLoS ONE* **2021**, *16*, e0255939. [CrossRef]
45. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *arXiv* **2020**, arXiv:1908.07919. [CrossRef]