remote sensing

MDPI

Letter

Local Region Proposing for Frame-Based Vehicle Detection in Satellite Videos

Junpeng Zhang, Xiuping Jia * and Jiankun Hu

School of Engineering and Information Technology, University of New South Wales, Canberra 2612, Australia; junpeng.zhang@student.unsw.edu.au (J.Z.); J.Hu@adfa.edu.au (J.H.)

* Correspondence: x.jia@adfa.edu.au

Received: 31 August 2019; Accepted: 8 October 2019; Published: 12 October 2019

Abstract: Current new developments in remote sensing imagery enable satellites to capture videos from space. These satellite videos record the motion of vehicles over a vast territory, offering significant advantages in traffic monitoring systems over ground-based systems. However, detecting vehicles in satellite videos are challenged by the low spatial resolution and the low contrast in each video frame. The vehicles in these videos are small, and most of them are blurred into their background regions. While region proposals are often generated for efficient target detection, they have limited performance on satellite videos. To meet this challenge, we propose a Local Region Proposing approach (LRP) with three steps in this study. A video frame is segmented into semantic regions first and possible targets are then detected in these coarse scale regions. A discrete Histogram Mixture Model (HistMM) is proposed in the third step to narrow down the region proposals by quantifying their likelihoods towards the target category, where the training is conducted on positive samples only. Experiment results demonstrate that LRP generates region proposals with improved target recall rates. When a slim Fast-RCNN detector is applied, LRP achieves better detection performance over the state-of-the-art approaches tested.

Keywords: satellite videos; region proposals; convolutional neural networks; tiny and dim target detection; component mixture model

1. Introduction

As one of the most promising developments in remote sensing imagery, the satellite videos captured by Skybox and JL-1, have facilitated several emerging research and applications, including super resolution [1,2], video encoding [3,4] and target tracking [5,6]. They expand the earth observation capacity to rapid motion monitoring, such as vehicle and ship tracking [5,7,8]. To reveal these rapid motions, targets of interests need to be located throughout the satellite video first, and the extracted targets in each frame are then associated to construct the trajectories of targets of interest. Therefore, target detection in satellite videos is a fundamental and critical step for target tracking and motion pattern analysis.

Detecting objects of interest in a video can be achieved by the motion-based detectors, which search the changed pixels in a sequence of images by comparing with an estimated background model [9,10]. Various algorithms, such as Frame-Difference [5,11,12], Median Background [13], Gaussian Mixture Model (GMM) [14,15] and Visual Background Extractor (ViBe) [7,16,17], were developed for moving object detection. However, these approaches are prone to the inadequate background modelling and affected by the problem of parallax caused by the motion of the camera.

Alternatively, the image-based object detectors can extract objects of interest from a video frame by frame [18], whose performance is less affected by the parallax motion. By taking the advantage of the discriminative learning methods, these approaches employ a classifier to scan over possible

locations of targets in an image by sliding window [19–21]. To reduce the number of the candidate locations to examine, region proposals, which refer a sparse set of potential target locations, are introduced to replace sliding windows over the entire image. For common computer vision tasks, generating region proposals are commonly guided by the object saliency, such as the edges [22–24], or based on superpixels [25–29] or segmentation masks [30,31]. In aerial videos, the coherent regions extracted by Maximally Stable Extremal Regions (MSER) [32,33] or Top-hat-Otsu [34] are also adopted for region proposal generation. Due to the weak contrast between targets and background in satellite videos, saliency-based approaches result in degraded region proposal performance —either generating too many region proposals or producing a low target recall rate. These approaches also lack the mechanisms for quantifying the region proposals' likelihood of being a target, and place the entire burden of handling a large number of region proposals in the target recognition stage. Convolutional Neural Networks were applied for searching region proposals in recent years. These approaches can provide the confidence score for each region proposal, and a significant portion of false alarms in the region proposals are removed before the recognition state [35–38]. However, they heavily rely on the training of a reliable region proposal network using a large amount of training samples.

To improve the region proposal performance to handle dim and small target detection in satellite video, we propose a Local Region Proposing (LRP) approach with three steps in this study. Our observation is that vehicles in satellite videos appear small and dim globally. Therefore we propose to perform segmentation at a coarse scale to form semantic region first. Possible locations of small targets in each semantic region are then extracted. To reduce the false alarm further and alleviate the computation burden on further target recognition stage, a discrete Histogram Mixture Model (HistMM) is proposed to quantify their likelihoods towards the target category. HistMM presents little difficulty in cooperating with most detectors, as it is estimated separately and only positive samples are required for estimating the model.

The remaining part of this paper is structured as follows. Section 2 presents the proposed local region proposal approach, after which the experimental results are presented in Section 3. We conclude this paper in Section 4 with remarks on the promising direction for future study.

2. Local Region Proposing

Figure 1 shows the Local Region Proposing approach (LRP) developed in this study is composed of three steps. First semantic regions are extracted by coarse-scale segmentation, then possible target locations are searched in each extracted region. The Histogram Mixture Model is developed for removing obvious false alarms from the region proposals.

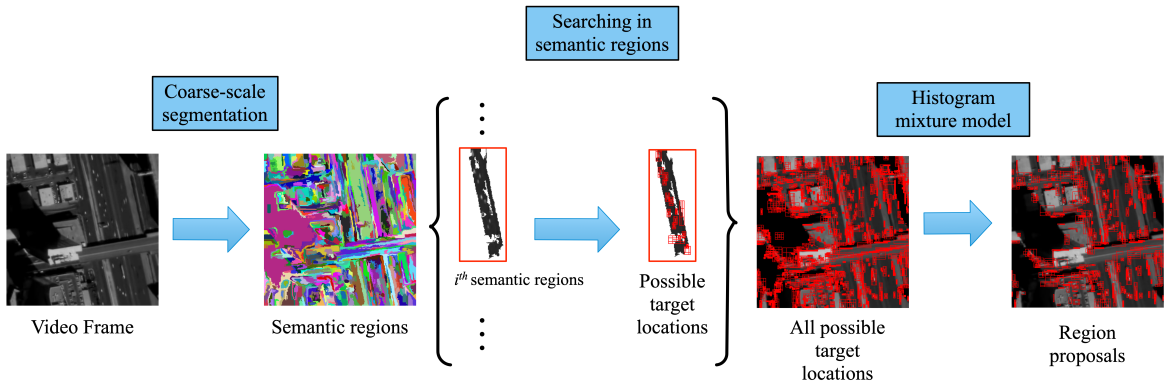


Figure 1. Overview of the proposed region proposal algorithm.

2.1. Semantic Region Extraction

Extracting semantic regions from a video frame can be by segmentation at a coarse scale, and the majority of pixels in each extracted region are more likely from a single land cover type.

The Felzenszwalb's graph-based segmentation approach [39] is a typical method for extracting the semantic regions.

By this graph-based segmentation approach, the scale of the generated superpixels can be controlled by a parameter k. Increasing k would lead to more coarse-scale superpixels, and these superpixels tend to present regions from different land cover types. The semantic regions are allowed to be larger than the target size on purpose. Decreasing k would generate fine-scale superpixels. However, it is often difficult to make superpixels to associate with small targets in satellite videos, due to the low spatial resolution and the low contrast of targets, for example, vehicles, to the background in satellite videos.

2.2. Searching Possible Locations in Semantic Regions

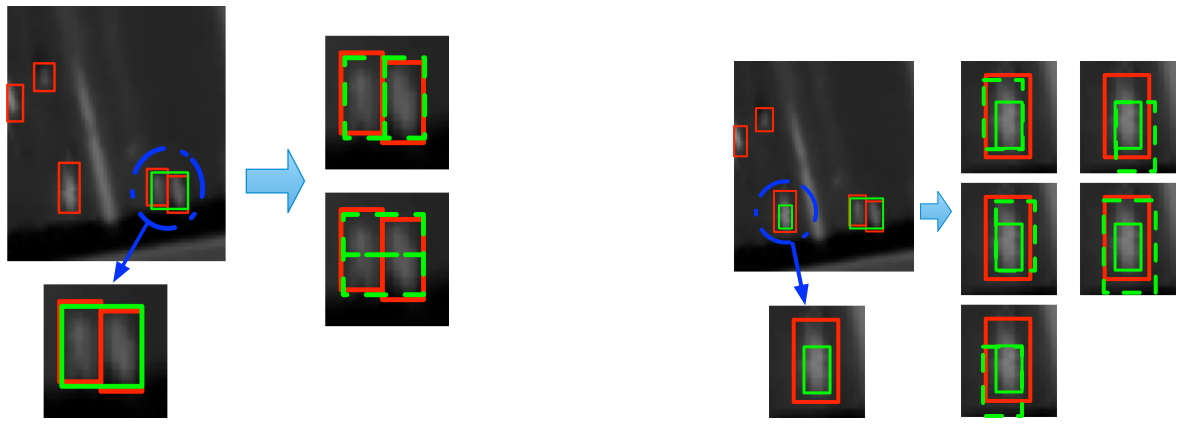
Unlike most dominating saliency object-based approaches, such as Selective Search [26,40], which merge superpixels to form region proposals, the proposed LRP searches region proposals inside semantic regions, where an adaptive threshold is introduced to accommodate the statistics of individual regions.

Note the set of extracted semantic regions as $\mathcal{R}$, for a semantic region that contains m pixels, the set of the pixels' coordinates is noted as $r = \{(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)\} \in \mathcal{R}$. The intensity of a pixel at location (x, y) is referred to $I(x, y)$. The blobs with high local saliency are constructed by the pixels with intensities over a threshold thr_r, $I(x, y) > thr_r, (x, y) \in r$. The threshold thr_r is defined by

$$thr_r = \mu_r + f * \sigma_r, \tag{1}$$

where μ_r and σ_r are the mean and standard deviation of pixel intensities in this local region r. The factor f is the expected saliency against the backgrounds. For each extracted blob, a corresponding boundary box is extracted as a possible location.

In the complex scenarios of satellite videos, this searching strategy may be affected by the presence of crowded vehicles and the blurred boundaries of vehicles, which results in merged proposals or incomplete proposals within an original boundary box extracted. We handle these cases by generating multiple proposals. The large boxes should be divided into sub regions to match the target size approximately and the small boxes should be expanded by half of the target size in each direction as a conservative treatment. Figure 2a shows an example where 4 region proposals are generated. To address those incomplete proposals, as shown in Figure 2b, the given bounding box is expanded in each directions.



(**a**) (**b**)

Figure 2. Generating multiple region proposals from a possible location. The red box refers to the groundtruth, green solid box refers to the extracted possible location, and green dash boxes refer to the generated region proposals. (**a**) and (**b**) illustrate two examples of generating region proposals by splitting and expanding original region proposals, respectively.

2.3. Histogram Mixture Model

2.3.1. Histogram Mixture Model for Removing Obvious False Alarms

The proposed Histogram Mixture Model (HistMM) measures the likelihoods of the generated region proposals towards their corresponding target category, so that obvious false alarms could be removed at an early stage. The HistMM is a mixture model built on a set of histograms, and training or estimating HistMM depends only on positive training samples.

Note the entire set of initial region proposals on a video frame as $\mathcal{X}_{rp} = \{x_0, x_1, \ldots, x_{n_{rp}}\}$, and n_{rp} is the number of initial region proposal on a given frame. For a region proposal $\forall x \in \mathcal{X}_{rp}$, it is marked as either target or background. We decide if x belongs to the target category (T) or the background category (B) by a Bayesian decision function,

$$R = \frac{p(T|x)}{p(B|x)} = \frac{p(x|T)p(T)}{p(x|B)p(B)}, \tag{2}$$

in which R measures the membership rate of x belonging to the target category versus belonging to the background category. $R \geq 1$ implies x is a target. The corresponding decision function for x that belongs to T can be simplified as

$$p(x|T) \geq c_t, \tag{3}$$

where c_t is a threshold.

The $p(x|T)$ refers to the likelihood of a region proposal x to the target category. We model it by a mixture model composed by a set of $n_{\mathcal{H}}$ histograms, $\mathcal{H} = \{h_1, h_2, \ldots, h_{n_{\mathcal{H}}}\}$. In this paper, we assume that each histogram contributes equally to the likelihood $p(x|T)$, therefore, the possibility of a proposal r that belongs to T is defined as,

$$p(x|T) = \frac{1}{n_{\mathcal{H}}} \sum_{i=1}^{n_{\mathcal{H}}} p(x|h_i). \tag{4}$$

The decision function in Equation (3) can be then interpreted as

$$p(x|T) = \frac{1}{n_{\mathcal{H}}} \sum_{i=1}^{n_{\mathcal{H}}} p(x|h_i) \geq c_t \Rightarrow \exists h \in \mathcal{H}, p(x|h) \geq c_t, \tag{5}$$

which means the likelihood to at least one histogram $\hat{h}_i$ in $\mathcal{H}$ is larger than c_t. On the contrary, a region proposals is a background when all likelihoods toward histograms in $\mathcal{H}$ are less than the threshold c_t, as

$$p(x|h) < c_t, \forall h \in \mathcal{H}. \tag{6}$$

For a given pair of a region proposal x and a histogram in $h \in \mathcal{H}$, we appropriate $p(x|h)$ by the Intersection of Histogram (IoH) between the histogram h and the histogram extracted from the region proposal x. For simplicity, we employ the Histogram of Color (HoC) for calculating $p(x|h)$, as

$$p(x|h) = IoH(h, HoC(x)) = \sum \min(h, HoC(x)), \tag{7}$$

which sums up the minimum values in all pairs of corresponding bins from h and $HoC(x)$. As shown in Figure 3, the IoHs on HoCs are distinct for distinguishing targets and backgrounds, although less information is provided due to the dim appearance of the vehicles.

Our HistMM removes obvious false alarms by the threshold c_t. A larger c_t tends to removal more possible false alarms, whereas it also risks abandoning some target instances. A smaller c_t may improve the coverage of targets in the region proposals, but the remaining number of proposals would be high. The detailed effects of different parameter settings are discussed in Section 3.2.

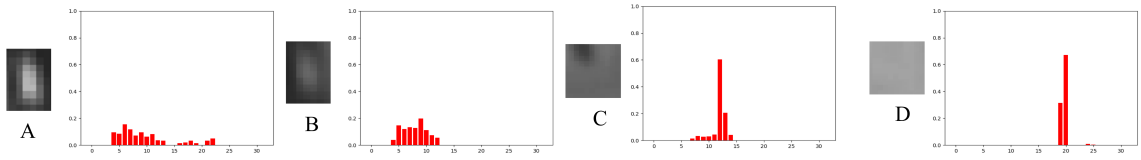


Figure 3. Histogram of Color can distinguish targets from backgrounds. Region proposal A and B are vehicles, whereas the region proposal C and D are obvious false alarms. For the four selected region proposals, their corresponding HoC are extracted, as shown in the right part of the figure. For A and B, the IoH is high, while both C and D have low IoH due to the extremely low similarities.

2.3.2. Estimating Histogram Mixture Model

For a set of n_{rp} possible region proposals $\mathcal{X}_{rp}$ on a video frame, we predict a region proposal $x \in \mathcal{X}_{rp}$ as a target or a background by Equation (6), as summarized in Algorithm 1. The complexity for predicting region proposals by HistMM grows linearly with the size of $\mathcal{X}_{rp}$, $\mathcal{O}(n_{\mathcal{H}} \times n_{rp})$. Therefore, our proposed HistMM is computationally feasible and scalable for the case with a large number of region proposals.

Algorithm 1 Removing Obvious False Alarms by Histogram Mixture Model (HistMM)

Input: $\mathcal{X}_{rp} = \{x_0, x_1, \ldots, x_{n_{rp}}\}$, $c_t > 0$, and $\mathcal{H} = \{h_1, h_2, \ldots, h_{n_{\mathcal{H}}}\}$
Output: $\mathcal{X}_{rp}$

```
for x ∈ X_rp do
    if ∀h ∈ H, p(x|h) ≤ c_t then
        Remove x from X_rp.
    end if
end for
return X_rp
```

HistMM is estimated by a recursive learning algorithm on the positive samples of groundtruths [14,41]. Note the estimated set of histograms by $\hat{\mathcal{H}} = \{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_{n_{\mathcal{H}}}\}$, and all the positive samples in the groundtruths is denoted by $\mathcal{X}_{gt}$. For a groundtruth $x_{gt} \in \mathcal{X}_{gt}$, a histogram $\hat{h}_m$, $m \in \{1, \ldots, n_{\mathcal{H}}\}$, is updated by

$$\begin{aligned} \hat{\pi}_m &\leftarrow \hat{\pi}_m + o_m(x_{gt}) \\ \hat{h}_m &\leftarrow \frac{\hat{h}_m \times \hat{\pi}_m + HoC(x_{gt}) \times o_m}{\hat{\pi}_m + o_m}, \end{aligned} \tag{8}$$

where $\hat{\pi}_m$ counts the updates of estimated histogram $\hat{h}_m$, and, as $\hat{\pi}_m$ increases, the lower fraction of the new samples are taken into $\hat{h}_m$. $o_m(x_{gt})$ defines the x_{gt}'s ownership of an estimated histogram $\hat{h}_m$ as

$$o_m(x_{gt}) = \begin{cases} 1, & p(x_{gt}|\hat{h}_m) \geq c_t \quad \text{and} \quad m = \underset{i \in \{0,1,\ldots,n_{\mathcal{H}}-1\}}{\arg\max} \; p(x_{gt}|\hat{h}_i) \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

by which $o_m(x_{gt}) = 1$ indicts that the new sample x_{gt} updates the histogram $\hat{h}_m$ by Equation (8). Otherwise, $o_m(x_{gt}) = 0$ means no nearby histogram component exists for this sample x_{gt}, and a new histogram component $\hat{h}_{n_{\mathcal{H}}}$ is added to $\hat{\mathcal{H}}$. $\hat{\pi}_{n_{\mathcal{H}}}$ is then initialized as 1 and the added histogram component $\hat{h}_{n_{\mathcal{H}}}$ is initialized by $HoC(x_{gt})$. This update procedure continues until it finishes iterating over the groundtruth set $\mathcal{X}_{gt}$, as summarized in Algorithm 2.

Algorithm 2 Training procedure of Histogram Mixture Model (HistMM)

Input: $\mathcal{X}_{gt} = \{x_1, \ldots, x_{n_{gt}}\}$, $c_t > 0$
Output: $\hat{\mathcal{H}}$

1: **for** $x \in \mathcal{X}_{gt}$ **do**
2: **if** $\exists \hat{h} \in \hat{\mathcal{H}}, p(x|\hat{h}) \geq c_t$ **then**
3: Find the updating histogram $\hat{h}_m$ and the ownership $o_m(x)$ by Equation (9).
4: Update $\hat{h}_m$ by

$$\hat{\pi}_m \leftarrow \hat{\pi}_m + o_m(x)$$

$$\hat{h}_m \leftarrow \frac{\hat{h}_m \times \hat{\pi}_m + HoC(x) \times o_m}{\hat{\pi}_m + o_m}.$$

5: **else**
6: Initialize a new component by $HoC(x)$, and add it to $\mathcal{H}$.
7: **end if**
8: **end for**
9: **return** $\hat{\mathcal{H}}$

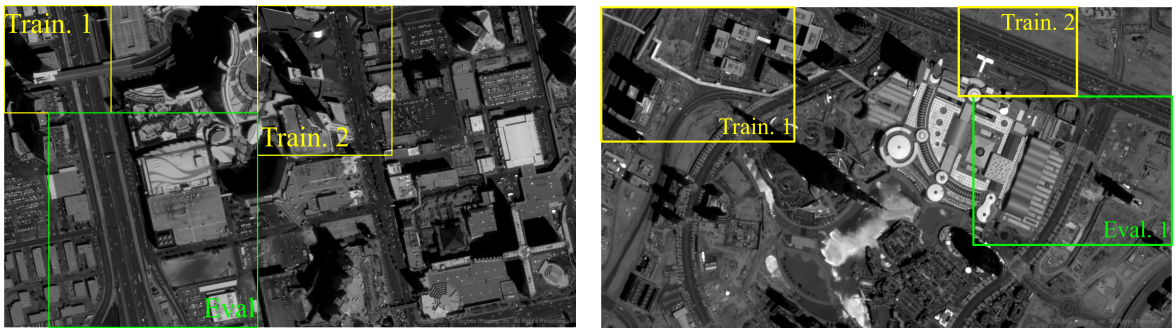
3. Experimental Results

3.1. Datasets

Two satellite video datasets, SkySat-Las Vegas dataset and SkySat-Burj Khalifa dataset, were used for experimental evaluation of the proposed method for efficient region proposal. For both datasets, the satellite videos were collected by SkySat, which recorded 1800 frames with 30 frames per second. The spatial resolution of each frame in this video is 1.5 m and the frame size is 1920 × 1080 pixels.

The SkySat-Las Vegas dataset refers to the satellite video captured over Las Vegas, USA in March 2014. As illustrated in Figure 4a, two sub-regions were selected for training and one sub-region was selected for evaluation.

The SkySat-Burj Khalifa dataset refers to the satellite video, which is captured over Burj Khalifa, United Arab Emirates on April, 2014. This video is 60 seconds long, which counts up to 30 frames per second. As shown in Figure 4b, 3 sub-regions were selected from the original video, two of which were for training and the remaining one for evaluation.



(**a**) SkySat-Las Vegas dataset (**b**) SkySat-Burj Khalifa dataset

Figure 4. Two typical frames from the two satellite video datasets used. (The regions surrounded by the rectangle in yellow color are for training, while the regions in green color are for testing.)

For both datasets, vehicles on five frames from each datasets were annotated, and their corresponding boundary boxes were provided as labelled samples. As we can see in Table 1, the average target sizes are very small.

Table 1. Detailed information for the datasets.

Dataset	Region	Size	Average Vehicle Size
SkySat-Las Vegas	Train. 1	360 × 360	7.09 × 5.12
	Train. 2	580 × 1070	6.27 × 5.03
	Eval	720 × 700	7.54 × 6.00
SkySat-Burj Khalifa	Train. 1	300 × 400	6.52 × 5.11
	Train. 2	450 × 650	7.07 × 5.28
	Eval	500 × 670	6.97 × 5.80

3.2. Parameter Discussion

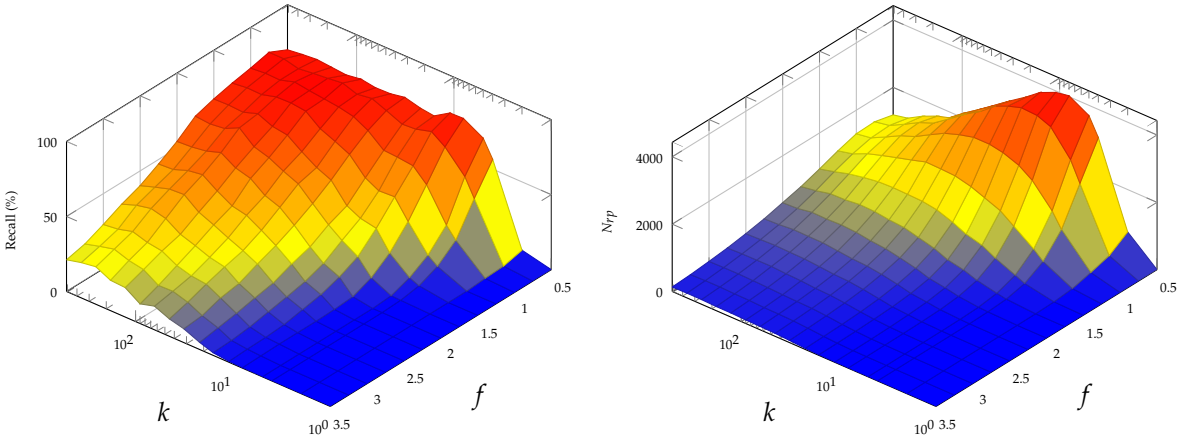
The LRP approach is mainly controlled by 3 parameters: the local region scale k, the threshold factor f and the threshold c_t in HistMM. The effect of each of them is discuss below. Their performance were evaluated in terms of the coverage of targets (recall), where a targets is recalled if there is at least 50% of IoU between any proposals and the ground-truth bounding box. These evaluations were conducted by the Leave-One-Out Cross Validation (LOOCV) strategy on training set of the SkySat-Las Vegas dataset.

- Semantic region Scale k controls size of the semantic regions generated. A larger k is preferred as it will generate a coarse segmentation as required. The semantic regions are allowed to be larger than the target size on purpose. As presented in Figure 5, reducing k gives fine-scale segmentation and leads to an increased number of region proposals with lower recall rate, while with increasing k, LRP generates fewer region proposals with improved recall rate.
- Threshold Factor f controls the segmentation threshold in each semantic region. Selecting a large f would result in fragmented region proposals and decrease recall scores. As illustrated in Figure 5, increasing f from 1.0 to 3.5, the recall scores experience a drop of over 40%.
- HistMM Threshold c_t is the Bayesian decision threshold in the HistMM for removing obvious false alarms as presented Section 2.3. The HistMM model with a smaller c_t tends to keep more obvious false alarms, which leads to unnecessarily more region proposals decreases. On the other hand, increasing c_t would filter out more obvious false alarms from the searched region proposals. As shown in Figure 6, when c_t increases to 0.5, the number of region proposals (N_{rp}) reduces significantly, while the recall scores holds nearly stable about 80%, which presents the most efficient case.

When c_t was set to 0.5 based on the cross validation on using the training data, the number of region proposals are reduced by over 60% by HistMM with almost no decrease in recall rate, las presented in Table 2 and Figure 7, which demonstrates the effectiveness of the proposed HistoMM model.

Table 2. Evaluation on the effectiveness of HistMM.

Dataset	Recall			N_{rp}		
	Before	After	Diff	Before	After	Diff
SkySat-Las Vegas	75.92%	75.10%	−0.82%	30,614	10,100	−67.01%
SkySat-Burj Khalifsa	77.31%	76.83%	−0.48%	17,017	6525	−61.66%



(**a**) Recall (**b**) N_{rp}

Figure 5. Region performance evaluation with different k and f.

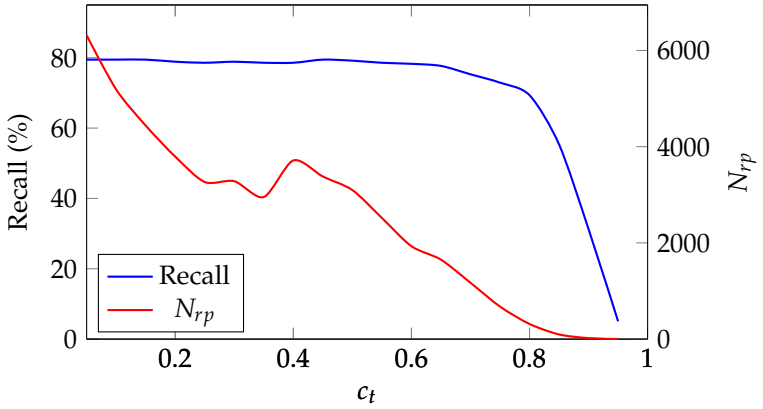
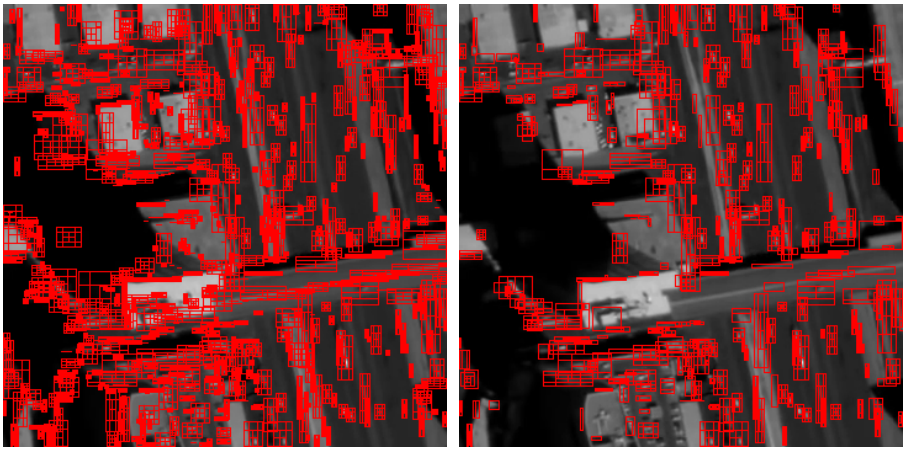


Figure 6. Region proposal performance by different c_t with $k = 81$, $f = 1.25$.



(**a**) Before HistMM (**b**) After HistMM ($C_t = 0.5$)

Figure 7. Visualization on region proposals before and after HistoMM.

3.3. Comparison of Region Proposal Approaches

The region proposal performance was compared with a set of existing region proposals approaches for both common object detection tasks as well as aerial object detection tasks. Inspired by the systematic region proposal evaluation research [42], the proposed region proposal scheme was evaluated against Superpixels (SP) [39,42], Selective Search (SS) [26] and Region Proposal Network

(RPN) [36]. SP generates a region proposal for each extracted superpixel, and SS merges neighboring superpixels as region proposals. For both SS and SP the extraordinarily tiny or large region proposals are considered impossible for vehicles in satellite videos and removed by post-processing. In addition to these well-known region proposals techniques, two approaches for aerial object detection are also included for comparison, which are Maximally Stable Extremal Regions (MSER) [33] or Top-hat-Otsu [34].

Qualitatively, the region proposals generated by our LRP are more concentrated on possible targets, while those saliancy object-based approaches, SS and SP, produce more evenly distributed region proposals, as shown in Figure 8. A similar phenomenon is observed on the results by RPN, as both RPN and our LRP remove those obvious false alarms from the background.

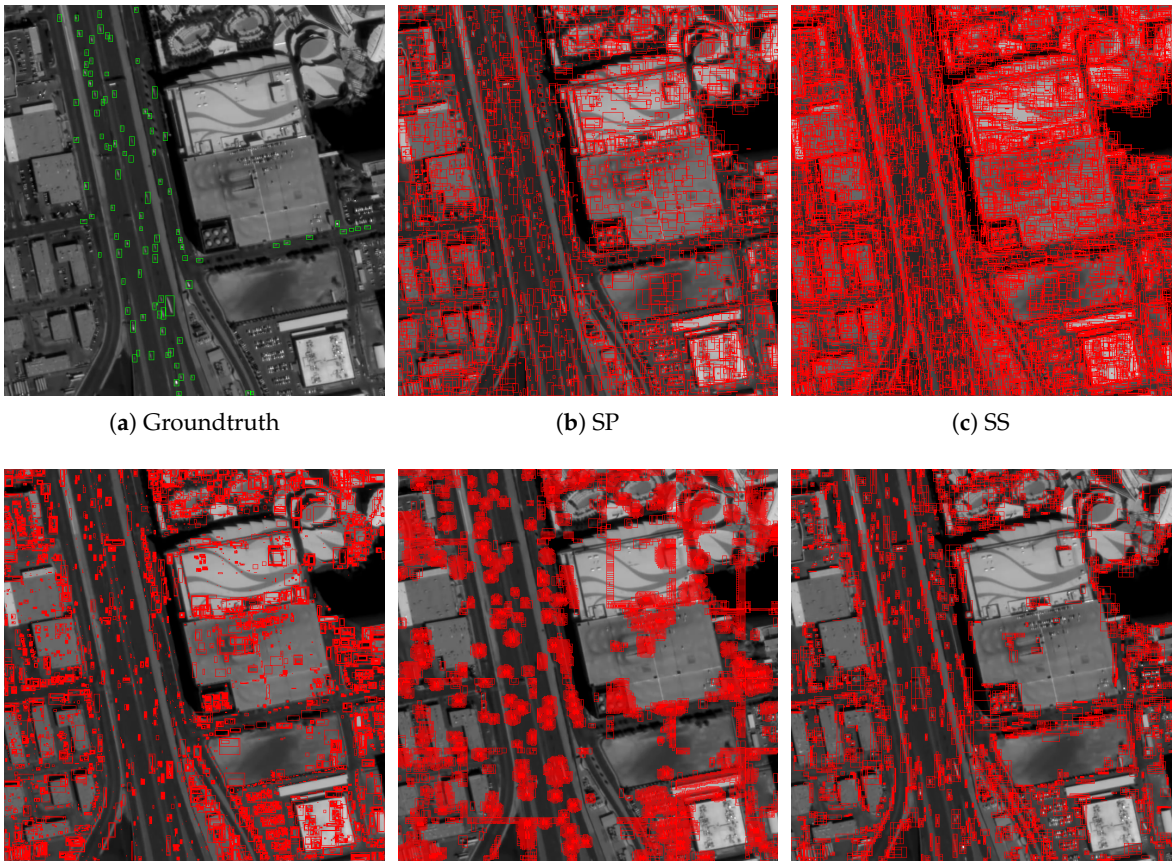
Then quantitative performance evaluation on different approaches was conducted in terms of recall scores. Benefiting from the adopted searching strategy and the HistMM, LRP generates a reasonable number of region proposals with good coverage of the possible targets. As presented in Table 3 and Figure 9, our LRP achieves the highest recall$_{@0.5}$ scores on both evaluation datasets. In term of the number of the generated region proposals, it seems like our LRP generates more region proposals than SP, but it should be noted that more than one region proposals are generated by LRP for most possible targets, as shown in Figure 8. Although RPN generates more region proposals with better recall rates, it takes advantage of the finetune scheme from our Fast R-CNN model.

Table 3. Evaluation on region proposal performance.

Method	SkySat-Las Vegas			SkySat-Burj Khalifsa		
	N_{rp}	Recall	Time (s)	N_{rp}	Recall	Time (s)
SP	4092	37.95%	1.98	7922	51.38%	1.28
SS	18,222	20.00%	588.97	11,728	19.34%	264.00
MSER	15,347	37.73%	0.48	10,569	55.80%	0.34
Top-hat-Ostu	1329	2.01%	0.02	1280	29.28%	0.01
RPN (Finetuned from Fast-RCNN-LRP)	13,288	90.00%	0.72	7908	90.05%	0.48
lLRP	9874	80.00%	4.23	7424	79.56%	3.60

Besides, we also compare the detection performance by using a slim Fast-RCNN detector. This slim Fast-RCNN receives 128 × 128 video frame as input, and it includes two groups of convolutional layers and a branch of fully connected layers for classification, where the branch for boundary box regression are replaced with carefully selected anchor distribution. Each group of convolutional layers contains three layers with kernel in the same size of 3 × 3, and the number of output channels is 16 and 32 for the first and second convolutional layer group, respectively. After each convolutional layer, a non-linear transformation is conducted by a Rectifier Linear Unit (ReLU) [43,44], which is followed by a Batch Normalization (BN) layer [45]. The output size by Roi Pooling is 2 × 2, which is followed by two fully connected layers with 512 and 32 hidden neural units, respectively. A Faster R-CNN model is also included for comparison. Due to the limited number of training samples, directly training a Faster R-CNN model is challenging, therefore, this Faster R-CNN model is finetuned from our Fast R-CNN-LRP. The performance evaluation is based on the PASCAL VOC metrics, where we use Average Precision (AP) instead of Mean Average Precision (mAP), since only one target category is contained in both datasets.

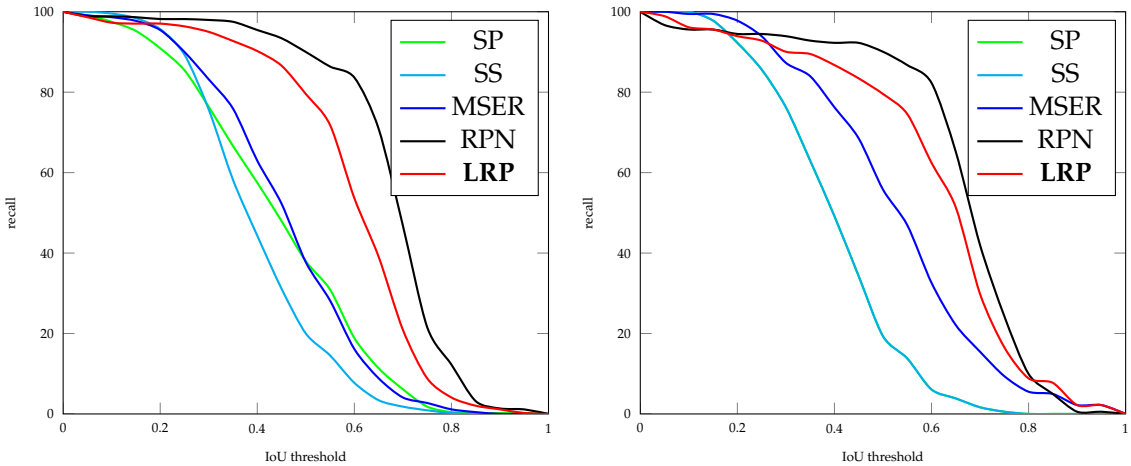
Compared with detection results by SP and SS approaches, our approach recalls most of the targets with the highest AP scores, as presented in Table 4 and Figure 10. Compared with the state-of-the-art Faster-RCNN model, the developed LRP with Fast-RCNN model achieves slightly improved detection performance. As illustrated in Figure 11, fewer false alarms with higher detection scores are produced by the Fast R-CNN model using the proposed LRP approach.



(**a**) Groundtruth (**b**) SP (**c**) SS

(**d**) MSER (**e**) RPN (Finetuned from Fast R-CNN-LRP) (**f**) **LRP**

Figure 8. Visualization on generated region proposals by different approaches on SkySat-Las Vegas Dataset.



(**a**) SkySat-Las Vegas Dataset (**b**) SkySat-Burj Khalifa Dataset

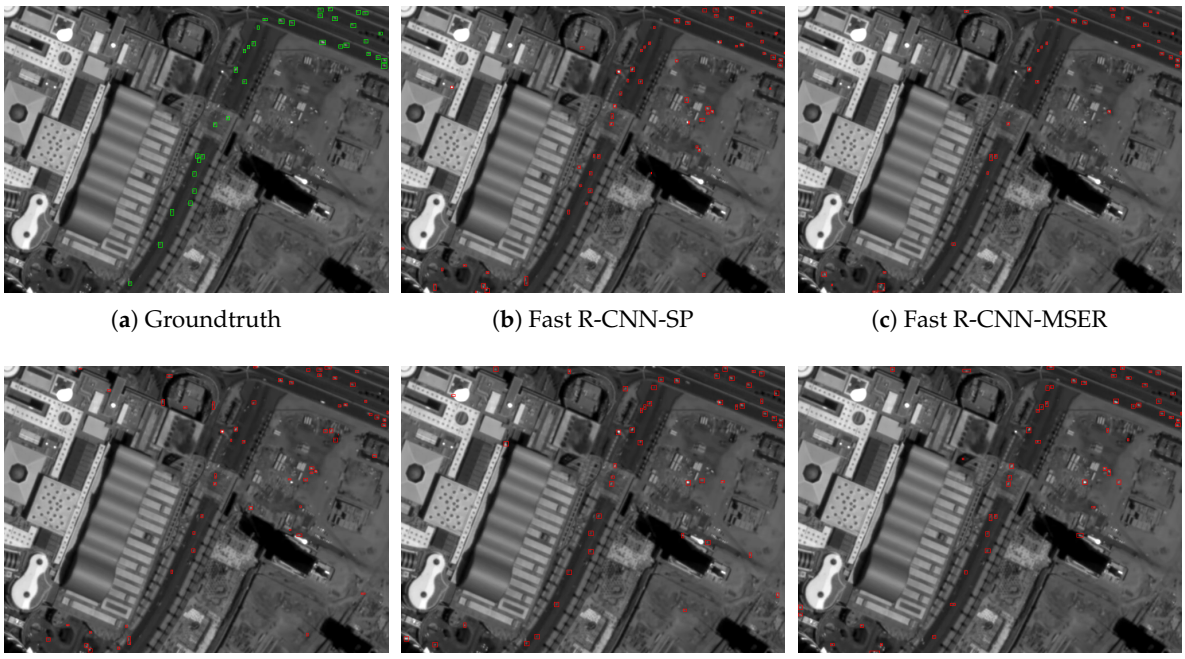
Figure 9. Recall rates over different IoU thresholds.

In addition to aforementioned single-frame-based detection approach, we also compare our approach with three popular background subtraction-based approaches —Gaussian Mixture Model (GMM) [46], GMMv2 [14] and Visual Background Extractor (ViBe) [16] approaches (A post-processing is applied to all these background subtraction-based approaches for removing extremely small or

large blobs.). Their performance are compared in terms of recall, precision and F_1 scores at IoU $= 0.5$. Compared with these background subtraction-based approaches, Fast-RCNN-LRP that uses our region proposals generates better F_1 scores, and the background subtraction-based approaches suffer from poor precision, as shown in Table 5.

Table 4. Detection performance evaluation.

Method	SkySat-Las Vegas				SkySat-Burj Khalifa			
	Rcll	**Prcn**	F_1	**AP**	**Rcll**	**Prcn**	F_1	**AP**
Fast R-CNN-SP	34.32%	35.53%	34.91%	29.20%	46.41%	31.82%	37.75%	35.30%
Fast R-CNN-SS	14.32%	19.57%	16.54%	7.43%	16.02%	12.78%	14.22%	5.90%
Fast R-CNN-MSER	30.45%	31.16%	30.80%	20.21%	41.44%	47.17%	44.12%	33.96%
Fast R-CNN-Top-hat-Ostu	1.82%	8.08%	2.97%	1.15%	26.52%	26.23%	26.37%	13.37%
Fast R-CNN-LRP	58.18%	43.91%	50.05%	**49.48**%	64.09%	42.49%	51.10%	**50.57**%
Faster R-CNN (Finetuned from Fast R-CNN-LRP)	59.32%	55.53%	56.31%	46.46%	62.43%	46.12%	53.05%	45.15%



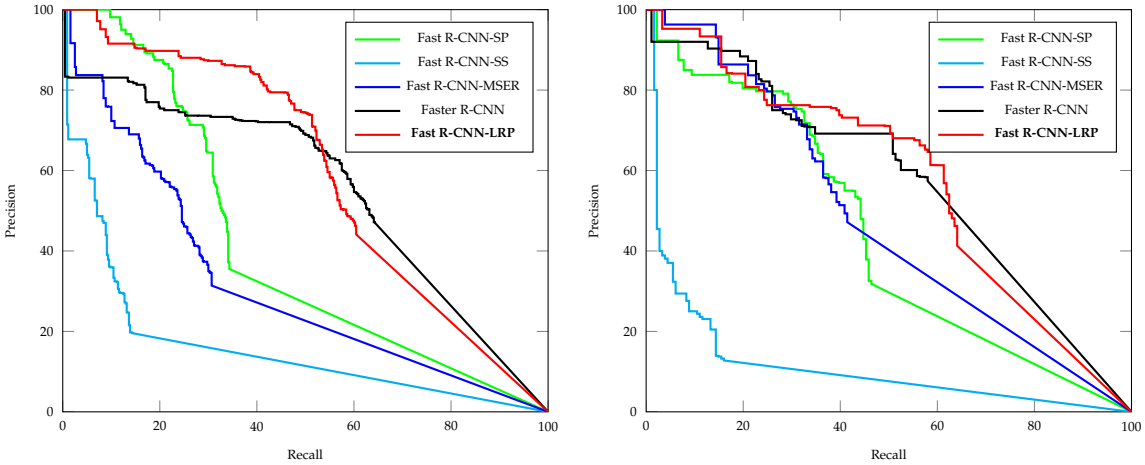
(**a**) Groundtruth (**b**) Fast R-CNN-SP (**c**) Fast R-CNN-MSER

(**d**) Fast R-CNN-SS (**e**) Faster R-CNN (**f**) Fast R-CNN-LRP

Figure 10. Visualization on detection results by selected approaches on SkySat-Burj Khalifsa dataset.

Table 5. Detection results comparisons.

Dataset	Method	Rcll	Prcn	F_1
SkySat-Las Vegas	GMM	45.8%	49.6%	47.6%
	GMMv2	64.7%	26.7%	37.8%
	ViBe	58.0%	16.7%	25.9%
	Fast-RCNN-LRP	58.18%	43.91%	50.05%
SkySat-Burj Khalifa	GMM	33.5%	56.7%	42.1%
	GMMv2	70.1%	37.7%	49.0%
	ViBe	74.6%	22.0%	34.0%
	Fast-RCNN-LRP	64.09%	42.49%	51.10%



(**a**) SkySat-Las Vegas Dataset (**b**) SkySat-Burj Khalifa Dataset

Figure 11. Precision-recall curve.

4. Discussion and Conclusions

Region proposal extraction is a valuable step to make target detection efficient. However, it is challenging to generate a small number of region proposals without missing any targets. This is more difficult when the targets are small and dim, such as those presented in satellite videos, due to their limited spatial resolution.

To address the degraded performance of current region proposal extraction methods for satellite videos, we proposed a novel region proposal approach (LRP), in which possible locations of targets are searched in semantic regions by coarse-scale segmentation and a Histogram Mixture Model (HistMM) is proposed to select region proposals with high likelihood from them.

The proposed LRP achieves improved recall rates of the targets with an acceptable increase in time cost, when compared with saliency object-based region proposal approaches, such as Superpixels (SP), Selective Search (SS), Maximally Stable Extremal Regions (MSER) and Top-hat-Otsu. Although the Region Proposal Network (RPN) recalls more targets with less time cost, it requires sufficient training samples or finetuning from a pre-trained model, such as the one obtained from LRP. Another advantage of the proposed LRP is that its training procedure only relies on positive training samples, even when a limited number of training samples is available.

With the improved recall rates by LRP, the detection performance by it with a slim Fast R-CNN is also superior to other saliency object-based region proposal approaches. The detection results are comparable with those by a finetuned Faster R-CNN model from our Fast R-CNN model. Compared with those background subtraction techniques, the proposal LRP approach outperforms them in term of precision, as fewer false alarms are generated.

As more satellite video data are available, more extensive testing can be conducted in the future study. In addition, the approach proposed in this manuscript is developed and tested on a panchromatic video data without color information. It may be extended to multi-channel data in the future research and improved detection performance can be expected.

Author Contributions: Conceptualization, J.Z.; methodology, J.Z. and X.J.; software, J.Z.; validation, J.Z.; writing–original draft preparation, J.Z. and X.J.; visualization, X.J.; supervision, X.J. and J.H.; project administration, X.J. and J.H.; funding acquisition, X.J.

Funding: This research received no external funding.

Acknowledgments: This work is partially supported by China Scholarship Council. The authors would like to thank Planet Team for providing the data in this research [47].

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

References

1. Luo, Y.; Zhou, L.; Wang, S.; Wang, Z. Video Satellite Imagery Super Resolution via Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2398–2402. [CrossRef]
2. Xiao, A.; Wang, Z.; Wang, L.; Ren, Y. Super-Resolution for "Jilin-1" Satellite Video Imagery via a Convolutional Network. *Sensors* **2018**, *18*, 1194. [CrossRef] [PubMed]
3. Wang, X.; Hu, R.; Wang, Z.; Xiao, J. Virtual Background Reference Frame Based Satellite Video Coding. *IEEE Signal Process. Lett.* **2018**, *25*, 1445–1449. [CrossRef]
4. Xiao, J.; Zhu, R.; Hu, R.; Wang, M.; Zhu, Y.; Chen, D.; Li, D. Towards Real-Time Service from Remote Sensing: Compression of Earth Observatory Video Data via Long-Term Background Referencing. *Remote Sens.* **2018**, *10*, 876. [CrossRef]
5. Du, B.; Sun, Y.; Cai, S.; Wu, C.; Du, Q. Object Tracking in Satellite Videos by Fusing the Kernel Correlation Filter and the Three-Frame-Difference Algorithm. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 168–172. [CrossRef]
6. Zhang, J.; Jia, X.; Hu, J.; Tan, K. Satellite Multi-Vehicle Tracking under Inconsistent Detection Conditions by Bilevel K-Shortest Paths Optimization. In Proceedings of the 2018 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Canberra, Australia, 10–13 December 2018; pp. 1–8.
7. Yang, T.; Wang, X.; Yao, B.; Li, J.; Zhang, Y.; He, Z.; Duan, W. Small moving vehicle detection in a satellite video of an urban area. *Sensors* **2016**, *16*, 1528. [CrossRef] [PubMed]
8. Mou, L.; Zhu, X.X. Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1823–1826.
9. Cristani, M.; Farenzena, M.; Bloisi, D.; Murino, V. Background subtraction for automated multisensor surveillance: A comprehensive review. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 343057. [CrossRef]
10. Piccardi, M. Background subtraction techniques: A review. In Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics, The Hague, Netherlands, 10–13 October 2004; Volume 4, pp. 3099–3104.
11. Reilly, V.; Idrees, H.; Shah, M. Detection and tracking of large number of targets in wide area surveillance. In Proceedings of the European Conference on Computer Vision, lHeraklion, Crete, Greece, 5–11 September 2010; Springer: Berlin, Germany, 2010; pp. 186–199.
12. Xiao, J.; Cheng, H.; Sawhney, H.; Han, F. Vehicle detection and tracking in wide field-of-view aerial video. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 679–684.
13. Sommer, L.W.; Teutsch, M.; Schuchert, T.; Beyerer, J. A survey on moving object detection for wide area motion imagery. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
14. Zivkovic, Z. Improved adaptive Gaussian mixture model for background subtraction. In Proceedings of the Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, UK, 26 August 2004; Volume 2, pp. 28–31.
15. Pollard, T.; Antone, M. Detecting and tracking all moving objects in wide-area aerial video. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012; pp. 15–22.
16. Barnich, O.; Van Droogenbroeck, M. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Trans. Image Process.* **2011**, *20*, 1709–1724. [CrossRef]
17. Xiang, X.; Zhai, M.; Lv, N.; El Saddik, A. Vehicle counting based on vehicle detection and tracking from aerial videos. *Sensors* **2018**, *18*, 2560. [CrossRef]
18. Kang, K.; Ouyang, W.; Li, H.; Wang, X. Object detection from video tubelets with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 817–825.
19. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
20. Zhang, W.; Sun, X.; Fu, K.; Wang, C.; Wang, H. Object detection in high-resolution remote sensing images using rotation invariant parts based model. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 74–78. [CrossRef]

21. Cheng, G.; Han, J.; Guo, L.; Liu, Z.; Bu, S.; Ren, J. Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4238–4249. [CrossRef]
22. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2189–2202. [CrossRef] [PubMed]
23. Cheng, M.M.; Zhang, Z.; Lin, W.Y.; Torr, P. BING: Binarized normed gradients for objectness estimation at 300fps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3286–3293.
24. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin, Germany, 2014; pp. 391–405.
25. Gokberk Cinbis, R.; Verbeek, J.; Schmid, C. Segmentation driven object detection with fisher vectors. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2968–2975.
26. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
27. Manen, S.; Guillaumin, M.; Van Gool, L. Prime object proposals with randomized prim's algorithm. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2536–2543.
28. Rantalankila, P.; Kannala, J.; Rahtu, E. Generating object segmentation proposals using global and local search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2417–2424.
29. Endres, I.; Hoiem, D. Category-independent object proposals with diverse ranking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 222–234. [CrossRef] [PubMed]
30. Carreira, J.; Sminchisescu, C. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1312–1328. [CrossRef] [PubMed]
31. Pont-Tuset, J.; Arbelaez, P.; Barron, J.T.; Marques, F.; Malik, J. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 128–140. [CrossRef] [PubMed]
32. Matas, J.; Chum, O.; Urban, M.; Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis. Comput.* **2004**, *22*, 761–767. [CrossRef]
33. Teutsch, M.; Krüger, W.; Beyerer, J. Evaluation of object segmentation to improve moving vehicle detection in aerial videos. In Proceedings of the 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, Korea, 26–29 August 2014; pp. 265–270.
34. Zheng, Z.; Zhou, G.; Wang, Y.; Liu, Y.; Li, X.; Wang, X.; Jiang, L. A novel vehicle detection method with high resolution highway aerial image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 2338–2343. [CrossRef]
35. Szegedy, C.; Reed, S.; Erhan, D.; Anguelov, D.; Ioffe, S. Scalable, high-quality object detection. *arXiv* **2014**, arXiv:1412.1441.
36. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
37. Zhang, F.; Du, B.; Zhang, L.; Xu, M. Weakly supervised learning based on coupled convolutional neural networks for aircraft detection. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5553–5563. [CrossRef]
38. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
39. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
40. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
41. Zivkovic, Z.; van der Heijden, F. Recursive unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 651–656. [CrossRef] [PubMed]

42. Hosang, J.; Benenson, R.; Dollár, P.; Schiele, B. What makes for effective detection proposals? *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 814–830. [CrossRef]
43. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processing, Atlanta, GA, USA, 16 June 2013.
44. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
45. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
46. KaewTraKulPong, P.; Bowden, R. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-Based Surveillance Systems*; Springer: Berlin, Germany, 2002; pp. 135–144.
47. Team, P. Application Program Interface: In Space for Life on Earth. San Francisco, CA. Available online: https://api.planet.com (accessed on 31 August 2019).