Fully Convolutional Deep Neural Networks for Persistent Multi-Frame Multi-Object Detection in Wide Area Aerial Videos

Rodney LaLonde, Dong Zhang, Mubarak Shah
Center for Research in Computer Vision, University of Central Florida

Abstract

Multiple object detection in wide area aerial videos, has drawn the attention of the computer vision research community for a number of years. A novel framework is proposed in this paper using a fully convolutional deep neural network, which is able to detect all objects simultaneously for a given region of interest. The network is designed to accept multiple video frames at a time as the input and yields detection results for all objects in the temporally center frame. This multi-frame approach yield far better results than its single frame counterpart. Additionally, the proposed method can detect vehicles which are slowing, stopped, and/or partially or fully occluded during some frames, which cannot be handled by nearly all state-of-the-art methods. To the best of our knowledge, this is the first use of a multiple-frame, fully convolutional deep model for detecting multiple small objects and the only framework which can detect stopped and temporarily occluded vehicles, for aerial videos. The proposed network exceeds state-of-the-art results significantly on WPAFB 2009 dataset.

1 Introduction

Object detection in wide area aerial videos has drawn the attention of the computer vision research community for a number of years [14, 18, 20, 23, 26]. Numerous applications exist for both civilian and military domains. In the field of urban planning, applications include automatic traffic monitoring, with potentially real-time traffic optimizations and map updates, driver behavior analysis, and road verification for assisting both scene understanding and land use classification. Civilian and military security is another large area to potentially benefit with applications including military reconnaissance, detection of abnormal or potentially dangerous behavior, border protection, and surveillance of restricted areas. With the recent increases in the use and affordability of drones and other unmanned aerial platforms, the desire for building a robust system to detect objects in wide-area, low-resolution, aerial videos has developed considerably.

Object detection in aerial videos can be defined as the following. Given an aerial video, taken from a camera or a set of cameras mounted on a moving platform and covering a wide area of the ground, place a bounding box or a single point (which typically corresponds to the center of the object) on every object of interest in every frame. One frame of video taken from the WPAFB 2009 [1] with its ground truth detections are shown in Fig. 1.

Figure 1: An example wide area aerial video frame. The yellow-boxed region is enlarged, then a blue-boxed region is further enlarged to show vehicles to detect. Ground truth annotations are marked with red dots.

Wide area aerial videos pose some unique challenges that makes detecting objects extremely challenging. As a result, most of the state-of-the-art object detection methods used in general object detection do not perform well in this domain. In general object detection, most methods are based on appearance (*e.g.* Fast R-CNN [6], ResNet [7]), which learn how the objects look and search for them throughout the video frames. There are three main reasons why the appearance based object detection methods fail on aerial videos: 1) Objects in the aerial videos are very small (*i.e.* roughly on the average 9×18 pixels in WPAFB 2009 dataset) with high intra-class variation, leaving minimal appearance cues to exploit. 2) The airborne sensor platforms can cover a large area (tens of square kilometers) with hundreds of millions of pixels in each frame. This makes the search space extremely large for the appearance-based object detection. Additionally, due to the first reason, object proposal techniques (*e.g.* Faster R-CNN [19]) do not work well with the aerial videos to reduce the search space, as they cannot generate good object proposals for extremely small objects in this application. 3) The appearance-based object detection methods need a large training set (*e.g.* ImageNet [5]) to in-

crease the performance. For aerial videos, it is very difficult to generate enough annotations since it is quite cumbersome to find the small objects in such large videos.

Due to the aforementioned reasons, the most successful object detection methods for aerial videos are motion-based [17, 23, 24], which use background subtraction or frame differencing to find the objects in the videos. However, the motion-based approaches also suffer from three drawbacks. Obviously, the first drawback is they totally ignore any appearance information. The second drawback is they highly rely on the frame registration, and small errors in frame registration can induce large failure in the final results. The third drawback is they do not have a mechanism to use information from multiple frames efficiently. Due to these reasons and drawbacks of the state-of-the-art methods, it is intuitive to ask the following question: How can one use multiple video frames and combine the appearance and motion cues to improve the object detection results on aerial videos?

With the above analyses and reasoning in mind, we propose a multi-frame fully convolutional deep network based method for multiple object detection in aerial videos. This new method can leverage both appearance and motion cues, and process multiple video frames efficiently. The network is designed to accept several video frames at a time as the input and yields detection results for all objects simultaneously. The proposed deep network is a fully convolutional neural network, which means the spatial information for the input is kept throughout the network. The network input consists of several consecutive video frames (*e.g.* 5 frames), and the output is the detection map of the center frame (*e.g.* the 3^{rd} frame). The deep network is trained in end-to-end fashion, and during the testing stage, the deep network is applied to every temporal window in order to generate detection results in every video frame. Since multiple frames are input together into the deep network, appearance and motion cues are automatically combined and learnt by the training process. Also, since it is a fully convolutional network, the searching is automatically done by the convolutions, which significantly improves the computational efficiency. While some work has been done on using multiple video frames as input to neural networks [3, 8, 10], our method is very different from them since they are mostly focused on obtaining better features for action recognition. Our method has very different configurations, and is designed for simultaneous multiple object detection rather than classification.

The contribution of the proposed method can be summarized as follows: 1) First use of a multi-frame, deep-learning model for multiple object detection in aerial videos. 2) The first object detection method for aerial videos which can handle slowing, stopped, and even occluded vehicles for persistent detections. 3) The proposed method significantly exceeds all state-of-the-art results [4, 17, 23, 24] on WPAFB 2009 dataset.

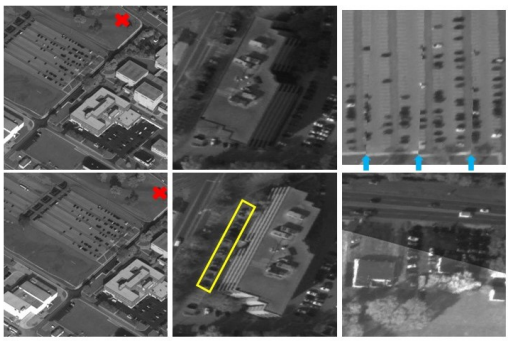


Figure 2: Examples of difficulties present in wide area aerial videos. **Left:** Two consecutive frames of video showing the very low frame rate, illustrated by the red x placed at the same real-world coordinates in each frame. **Center:** Dramatic motion parallax effects. Vehicles in the yellow box are occluded for periods of time due to motion parallax. **Right:** Several difficulties including mosaic seams (shown by the blue arrows), camera gain differences, blurred/unclear object boundaries, etc.

2 Related Work

Due to the difficulties involved in aerial videos, as discussed in the introduction, majority of literature has reported that building appearance feature based or machine learning based classifiers is quite difficult [18, 21, 23, 24]. Thus majority of the state-of-the-art methods fall into two categories: frame differencing and background subtraction. Eleven state-of-the-art methods in this area are compared in [23] and all fall within these two categories.

Both frame differencing and background subtraction methods require video frames to be registered, also known as global platform motion compensation, to a single coordinate system. This is usually achieved via some version of a point-matching based algorithm. Reilly *et al.* [18] detect Haris corners in two frames, computes the SIFT features around those corners and match the points using descriptors. A frame-to-frame homography is then fit, using RANSAC or a similar method, and used to warp images to a common reference frame.

Frame differencing is the process of computing pixel-wise differences in intensities between consecutive frames. Both two-frame and three-frame differencing methods have been proposed in literature with a number of variations [11, 16, 21, 23, 25]. These methods suffer from issues caused by both characteristics of wide area aerial video. The large camera motion combined with a low frame rate, low resolution, and single channel data makes registration methods imprecise. Errors in the registration are then directly transferred to frame differencing methods in the form

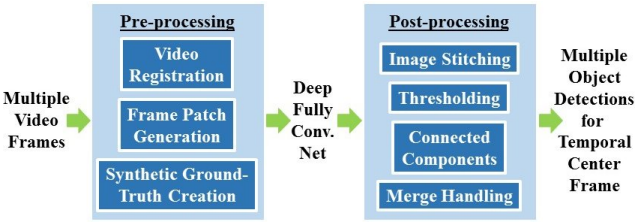


Figure 3: Multi-frame, multi-object detection framework

of false positives. Strong motion parallax effects combined with moving camera mosaic seams (*i.e.* where the multiple cameras are stitched together to form a single sensor) cause additional false positives. Some of these difficulties are illustrated in Fig. 2.

Background subtraction methods focus on obtaining a background model for each frame, then subtracting each video frame from its corresponding background model. Reilly *et al.* computed a background image model by computing the median of ten consecutive images for each frame of video. Using the median as opposed to the mean uses far fewer images and leaves fewer artifacts [18].

A major drawback of both approaches is the incapability of detecting stopped or occluded vehicles. Slowing vehicles also cause a major problem as they are prone to causing split detections in frame differencing [24] while registration errors and parallax effects are increased in background subtraction models which use more frames than frame differencing. Additionally, sudden and dramatic changes in camera gain cause illumination changes which cause problems for background modeling and frame differencing methods which assume consistent global illumination [21].

3 Proposed Method

We propose a multi-frame multi-object detection method based on a fully convolutional neural network (FCNN) (shown in Fig. 3). The proposed method consists of three important stages: 1) Pre-process the videos and annotation for our deep network formulation; 2) Multi-frame multi-object detection by fully convolutional neural network. 3) Post-processing the neural network output for final results. We first introduce our main contribution: the multi-frame multi-object detection neural network in Section 3.1, 3.2, and 3.3, then we discuss other two stages in Section 3.4.

3.1 Single-frame Deep Network

In order to develop our multi-frame multi-object detection neural network, we start from a simpler case: single-frame multi-object detection. Traditional approaches were focused on training an object detector and using sliding-window techniques to find the objects from the frame. Recently, object proposal based approaches (*e.g.* Faster R-CNN [19]) dominate the research and they generate object candidates to reduce the search space. However, these two approaches have some drawbacks in multi-object detection for aerial images, as already discussed in Introduction.

Fully convolutional neural networks (FCNN) are good candidates to tackle this problem; however, several design choices must be carefully considered to boost the performance. A single frame can be input to a FCNN and the location (or bounding boxes) of the objects can be used as the ground-truth. However, this approach makes the neural network very difficult to converge, since the sizes of objects are usually only a few pixels. In order to enable the neural network to perform better, the heat-map formulation can be employed, which has already been demonstrated useful for other computer vision applications (*e.g.* human pose estimation [15]). In this formulation, a heat-map can be generated according to the ground-truth object locations to guide the neural network output. We apply small 2D Gaussians at the locations of objects as the ground-truth and back-propagate them to train the neural network (see Fig. 4).

3.2 Multi-frame Deep Network

The information from multiple video frames can be used simultaneously for object detection in aerial videos. If formulated correctly, both appearance information and motion information can be combined together to improve the performance. State-of-the-art generic object detection approaches mostly focus on appearance information, but for object detection in aerial videos, the most successful methods are purely based on motion information. From these two traditional groups of methods, it is not immediately clear how to best combine the appearance and motion cues.

We propose to employ a multi-frame FCNN to exploit the appearance and motion information from the frames simultaneously (see Fig. 4). The input to the neural network is several consecutive video frames and the output is the object detection heat-map described in Section 3.1. The network can learn the temporal information without explicitly needing to create a difference image or compute the background subtraction. Features are combined from all the frames within the deep network for a given size temporal window. The network uses this information to create a single heat-map for predicted vehicle locations in the temporally center frame. The temporal window is then moved forward one step in time and results are obtained for each frame of video in this manner. To the best of our knowledge, this is the first use of multi-frame fully convolutional neural networks for object detection.

3.3 Deep Network Architecture

Our FCNN structure can be seen in Fig, 4. ReLU layers were used on every convolution with a dropout of 50%

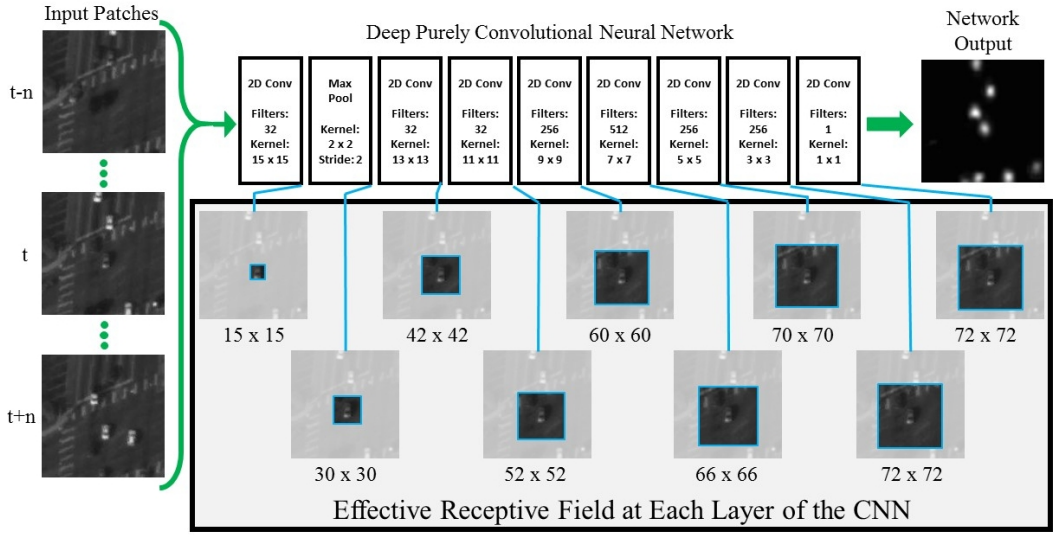


Figure 4: The proposed multi-frame fully convolutional neural network for multiple object detection. The blue-boxed regions demonstrate the effective receptive field for each pixel in the activation maps of each layer. For example after the first convolutional layer, each pixel in the output activation map "sees" a 15×15 region of the input image. By the end of the 8^{th} convolutional layer, each pixel in the output heat-map gathers information from a 72×72 region of the input image.

added to the sixth and seventh convolutional layers. Networks were designed to handle any number of input images and output a single vehicle detection map. These maps were then post-processed as described in Section 3.4 to obtain a single x,y point for each proposed object. Activation maps were sampled from the FCNN to help illustrate how both appearance and temporal information was being learned by the network. These sample activation maps can be seen in Fig. 5. Areas where vehicles are moving appear as blurred out regions in the early activation maps. Classic appearance features such as edges and corners are also present in these early layers. In later layers, the earlier blurred regions begin to have very strong activations and are combined to the appearance features to produce quite impressive results. Clearly the network is successfully combining both sources of information.

3.4 Data Processing for Deep Network

In order for the deep network to perform well, some preprocessing of the video frames and post-processing of the results are needed.

3.4.1 Frame Registration

For all state-of-the-art approaches, the motion registration of video frames over time is absolutely necessary. Video frames are fed into the neural network only after global camera motion has been removed. Video motion compensation was performed following the method proposed by Reilly *et al.* [18] where Harris corners were detected in each frame, SIFT features were extracted around each corner, frame-to-frame homographies were computed and then fit using RANSAC, and finally images were warped to a common coordinate system. Due to large camera motion, parallax, poor resolution, drift in the homographies, etc., these initial alignments were only roughly correct and second-pass local alignment was performed for each area-of-interest (AOI).

3.4.2 Ground-truth Preparation

For most of the publicly available aerial video datasets (*e.g.* WPAFB 2009 dataset [1]) with object detections, the annotations come in the form of (x, y) coordinates with vehicle ID numbers and other metadata. Since it would be nearly impossible to train a deep network using only a single point as the ground truth, the (x, y) ground-truth locations are processed to a heat-map format. These heat-maps are created by centering a Gaussian filter with set variance, σ, at each (x, y) coordinate, with the highest score being in the center. Each frame of video has a corresponding ground truth heat-map with all vehicles marked by these Gaussian filters, thus one heat-map can have hundreds or even thousands of these Gaussian spots (see Fig. 6). The loss for supervised learning is then computed as the Euclidean distance (Eq. 1) between the network output and the ground-truth heat-map as follows,

$$L_i = \frac{1}{2N} \sum_{j=1}^{N} ||x_j^1 - x_j^2||_2^2 \quad (1)$$

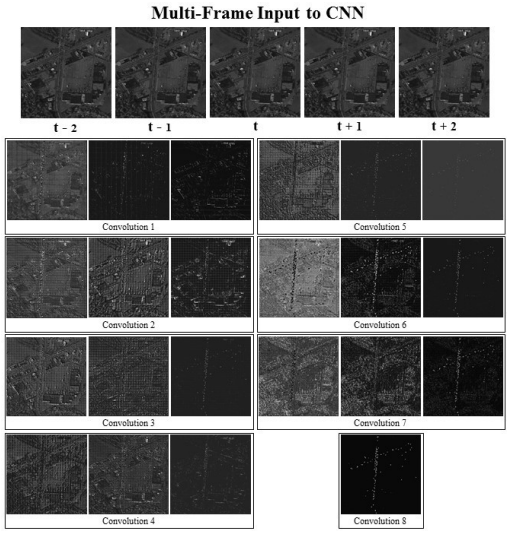


Figure 5: A few activation maps for a multi-frame input are sampled from each layer. In early layers, moving vehicles appear as blurred out regions in some activation maps while others find classical appearance features such as edges and corners. In later layers, the earlier blurred regions begin to have very strong activations. When combined with the more appearance feature-based activation maps also found in later layers, the output of the final convolutional layer produces quite impressive results.

where the loss for a single output i is calculated by taking the squared difference of x_j^1 and x_j^2, the pixel intensities for the j^{th} output heat-map pixel and ground-truth heat-map pixel, calculated over all N pixels. An additional approach was investigated using binary ground-truth maps instead of these Gaussian heat-maps. Here the problem is formulated as a binary segmentation problem (BinSeg) where segmentation masks are created for each vehicle location. Loss is computed using a softmax with cross-entropy loss function (Eq. 2) instead of a Euclidean distance,

$$L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e_j^f}\right) \tag{2}$$

where f_j are the pixel intensity of the output heat-map i, and f_{y_i} is the ground-truth.

3.4.3 Post Processing for Deep Network

As just discusssed, pre-processing the data turns the original ground-truth detections, given as (x, y) coordinates, into heat-maps for training. On the output side of the FCNN we must now turn our output heat-maps back into single (x, y) coordinates. First the output of the network is thresholded to remove weak responses in the heat-map and create a binary map. Connected components are then found and

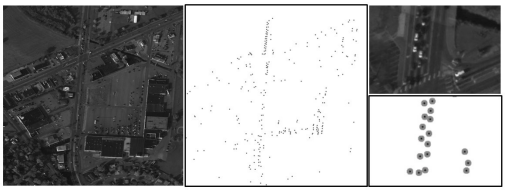


Figure 6: AOI 01 with its corresponding ground-truth Gaussian heat-map. A small section has been enlarged at right. Note: The heat-map has inverted colors and the slightly increased σ for display purposes.

components with very few pixels (*i.e.* < 100) are removed as these are considered just another form of weak response by the network. Next is the merged handling component. Since we trained on fairly large σ Gaussian spots when we created our training heat-maps, (this was done to try to help the network converge) our output spots sometimes overlap for very densely packed vehicles. To solve this, for larger than normal connected components (*i.e.* > 900 pixels), a bounding box is placed around the component and a circle finding algorithm is ran to find the center of each spot within the merged detection. If this method fails to return any circles, the merge is assumed to be only two objects and is split into two points, otherwise the centroid of each circle is returned. Centroids of the connected components not too small or too large are then taken as positive detections and added to those returned by the merge handling component.

3.5 Implementation Details

Since the frame size of the aerial videos is extremely large (over 5 MP for the cropped AOIs and over 315 MP for the entire video frame), the deep network can not process the frames due to the memory limit. Thus, the frames are first divided into small patches, fed through the network, and recombined on the other side. We divide the video frames into 128×128 pixel patches and each 5 temporally consecutive patches at the same location are combined as one patch image. Models were trained from scratch using Caffe [9]. The solvers used Adam to update the weights with a based learning rate of 0.00001. A batch size of 32 was used for both training and testing the network on a single Titan X GPU. Obtaining results for each frame of video using the preceding methods on a single GPU takes roughly 3.5 seconds.

4 Experimental Setup

4.1 Dataset

Our method was evaluated using the WPAFB 2009 dataset [1], one of the only publicly available datasets for wide area aerial video with ground-truth vehicle annotations. The video is taken from a single sensor, comprised of six slightly-overlapping cameras, covering an area of over 19

sq. km., at a frame rate of roughly 1.25 Hz. The average vehicle in these single-channel images make up only approximately 9×18 out of the over 315 million pixels per frame, with each pixel corresponding to roughly $\frac{1}{4}$ meter. With almost 2.4 million vehicle detections spread across only $1,025$ frames of video, there averages out to be well over two thousand vehicles to detect in every frame. After registering the frames to compensate for camera motion, eight AOIs were cropped out in accordance to those used is testing other state-of-the-art methods [4, 17, 23, 24], allowing for a proper comparison of results. AOIs $01 - 04$ are 2278×2278, covering different types of surroundings and varying levels of traffic. AOI 34 is 4260×2604. AOI 40 is 3265×2542. AOI 41 is 3207×2892. AOI 42 is simply a sub-region of AOI 41 but was included to test our method on persistent detections where slowing and stopped vehicles were not removed from the ground truth. All cropped AOIs are shown with their ground-truth heat-maps (one example frame and heat-map each) in the supplemental materials.

4.2 Data Creation

Ground-truth heat-maps and segmentation maps were created for each of the eight areas of interest. We obtained AOIs 34, 40, and 41 with their ground-truth annotations already redacted to only moving vehicles, since none of the state-of-the-art methods can handle detecting stopped vehicles, with the exception of [17] which using a tracking method to obtain persistent detections. AOIs $01 - 04$ were created ourselves by the above stated methods and only vehicles which did not move a distance greater than 15 pixels (or $\frac{2}{3}$ a car length) over the course of five frames were removed. This was done to show our method is more robust to slowing and stopped vehicles and could achieve higher results even with this more stringent exclusionary criteria, allowing more vehicles into the ground-truth to require detections. Data was then split into training and testing splits in the following way. For training, only tiles which contain vehicles were included. The splits were as follows: AOIs 02, 03, and 34 were trained on AOIs 40, 41, and 42; AOIs 01 and 40 were trained on AOIs 34, 41, and 42; and AOIs 04, 41, and 42 were trained on 34 and 40.

Binary extension images were created with varying channel depths and stored in lmdbs for training with Caffe. Single frame training used a single frame stored in a one channel image. Five frame images were combined in order with the frame at time $t - 2$ was in the first channel and at $t+2$ in the 5^{th} channel. For background subtracted images, these included 2 copies of the frame at t_0 and a background subtracted image filing a three channel input image. For all of these, the ground-truth map was a single channel binary extension image of the objects at frame t_0.

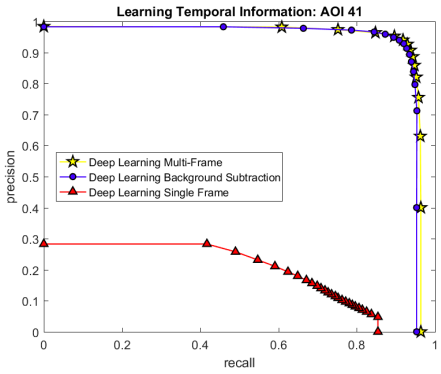


Figure 7: Results on AOI 41 testing the ability of the deep FCNN to learn explicitly or implicitly given temporal information, and its necessity.

5 Experimental Results

To be consistent with literature [23] detections were considered true positives (TP) if they fell within 20 pixels, or roughly 5 meters, of a ground truth coordinate. If multiple detections are within this radius of a ground truth coordinate, the closest one is taken and the rest, if they do not have any other ground truth coordinates with 20 pixels are marked as false positives (FP). Any detections that are not within 20 pixels of a ground truth coordinate are also marked as false positives. Ground truth coordinates which have no detections within 20 pixels are marked as false negatives (FN). These three statistics were used to generate precision-recall curves as well as F_1 scores for each experiment where

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{and } F_1 = 2 * \frac{precision * recall}{precision + recall}. \tag{4}$$

5.1 Learning Temporal Information

The first experiment run was looking to examine three key questions. These questions needed to be answered before the proposed method could be considered a candidate for further exploration. 1) Can deep learning improve the results of background subtraction (BS) methods when used together? In other words, can the network incorporate this temporal information with the appearance information of the video frames, when explicitly given it. To create these BS images for testing this hypothesis, we followed the method proposed by Reilly *et al.* [18] and trained the same deep FCNN as the proposed multi-frame method with three inputs: two copies of a single frame at time t_0 and one of its background subtraction image. 2) If the network can

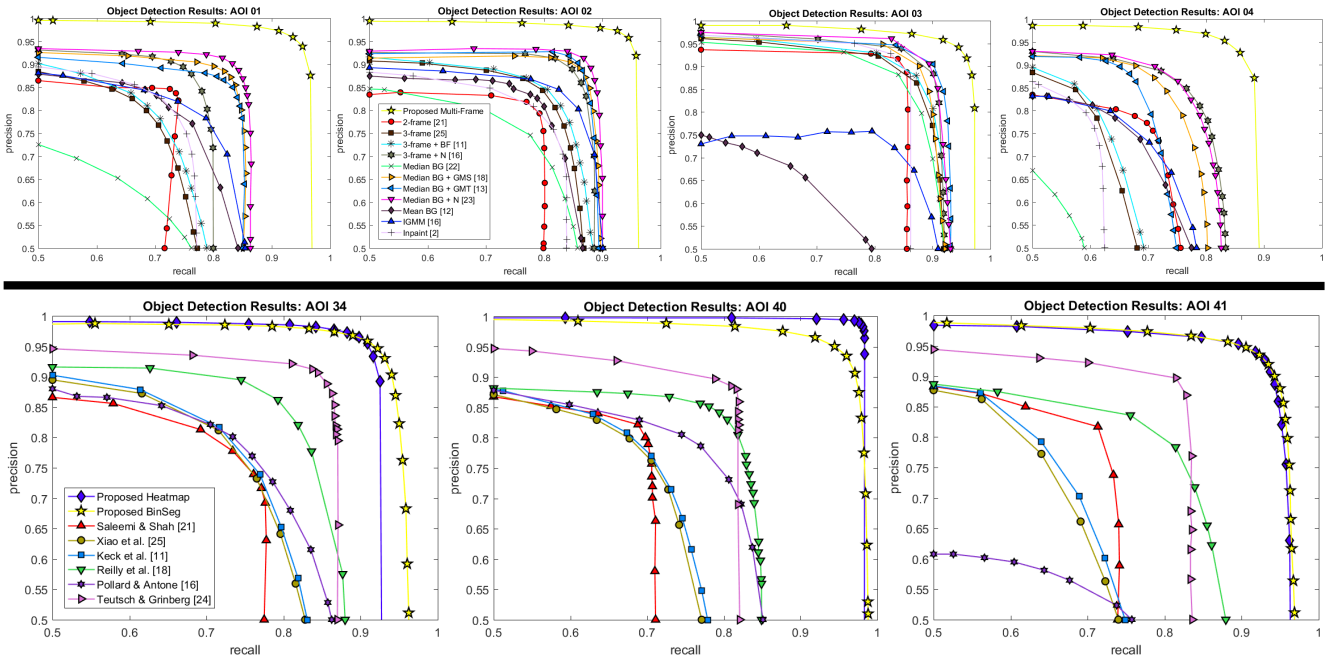


Figure 8: Moving object detection results on seven cropped AOIs with comparisons to state-of-the-art approaches. Note that the top and bottom rows each have their own separate keys. As one can see, our results for both the heat-map method and BinSeg method significantly outperform the current state-of-the-art.

Comparison of F_1 Scores on Eight Crop and Aligned Sections of the WPAFB 2009 Dataset

Method	01	02	03	04	34	40	41	42
Sommer *et al.* [23]	0.866	0.890	0.900	0.804	x	x	x	x
Shi [22]	0.645	0.760	0.861	0.575	x	x	x	x
Liang *et al.* [13]	0.842	0.880	0.903	0.760	x	x	x	x
Kent *et al.* [12]	0.767	0.807	0.668	0.711	x	x	x	x
Aeschliman *et al.* [2]	0.764	0.795	0.875	0.679	x	x	x	x
Pollard & Antone (3-frame + N) [16]	0.816	0.868	0.892	0.805	x	x	x	x
Saleemi & Shah [21]	0.783	0.793	0.876	0.733	0.755	0.749	0.762	x
Xiao *et al.* [25]	0.738	0.820	0.868	0.687	0.761	0.733	0.700	x
Keck *et al.* [11]	0.743	0.825	0.876	0.695	0.763	0.737	0.708	x
Reilly *et al.* [18]	0.850	0.876	0.889	0.783	0.826	0.817	0.799	x
Pollard & Antone (IGMM) [16]	0.785	0.835	0.776	0.716	0.766	0.778	0.616	x
Teutsch & Grinberg [24]	x	x	x	x	0.874	0.847	0.854	x
Prokaj & Medioni [17]	x	x	x	x	x	x	x	0.631
Proposed Multi-Frame	**0.947**	**0.951**	**0.942**	**0.887**	**0.933**	**0.983**	**0.928**	**0.927**

Table 1: F_1 scores of state-of-the-art methods. If precision-recall or F_1 values were not reported in the original work, the values reported in [23] and/or [24] were used. The proposed method outperforms all state-of-the-art methods by a significant margin on all AOIs. Note that AOI 42 is results on persistent detection (no vehicles removed from ground truth) and is compared with one of the only other persistent detection methods currently in literature.

learn to combine temporal information and appearance information effectively when it is explicitly given, can it learn the temporal information when only given implicitly? More concretely, given multiple frames as input, can the network *learn* the motion information? If the multi-frame method can perform on-par with the BS method, then we can remove the large computational overhead of first computing all of the median and then background subtracted images for every frame, in addition to limiting the errors the crop up in them from alignment defects. 3) Assuming both of

the above are successful, is any of it really necessary? It has been stated in numerous recent works [18, 21, 23, 24] that appearance based classifiers and machine learning classifiers almost universally fail due to the difficulties presented in wide area aerial video. Can we validate or refute this claim? Our experimental results shown in Fig. 7 answer all three of these questions strongly in the affirmative.

6 Comparison with State-of-the-Art

As stated previously, nearly all state-of-the-art methods can only detect moving vehicles, thus we dedicate a large section of experiments to detecting only moving vehicles before testing the robustness of our method on slowing and stopped vehicles as compared with one of the only other persistent detection methods currently proposed in literature [17].

6.1 Moving Object Detection

Experiments were performed on seven of the cropped and aligned AOIs previously discussed. For AOIs 34, 40, and 41, the results using both the binary segmentation approach (BinSeg) and the Gaussian heat-map approach (heat-map) are reported. For the rest of the curves, the multi-frame method is always using the Gaussian heat-map approach. Quantitative results of these experiments are shown in Fig. 8 and Table 1.

6.2 Persistent Object Detection

AOI 42 tests persistent detection rates. In this AOI, none of the ground-truth vehicles were removed to see if our method was more robust to slowing, stopped, and partially or temporarily occluded vehicles. The results far surpass the only other persistent method reporting results on this dataset and none of the other state-of-the-art methods included are able to detect these stopped vehicles. A qualitative example of our results can be seen in Fig. 9 and quantitative results in Fig. 10 and Table 1.

7 Conclusion

We have proposed a novel fully convolutional neural network based method for persistent multi-frame multi-object detection in aerial videos. In our method, we successfully taking advantage of both appearance and motion cues and integrate them into a single detection network, trained end to end. We have shown comparisons with many state-of-the-art methods, and the performance improvements are relatively 5 to 16% on moving objects for multiple videos in the WPAFB 2009 dataset as measured by F_1 score and nearly 50% relative improvement on persistent detections compared to [17]. Additionally, while detections are considered true positives if they fall within 20 pixels of the ground-truth, the proposed method's mean distance from ground truth annotations, averaged over all true positive detections, was roughly 2 pixels (0.5 m), compared to 5.5 pixels reported in [24]. We further demonstrated that the proposed method can handle stopped vehicles well, which is often a failure case in other methods. Future work can be, but not limited on, object detections with unaligned frames, region proposals for whole video frames, etc.

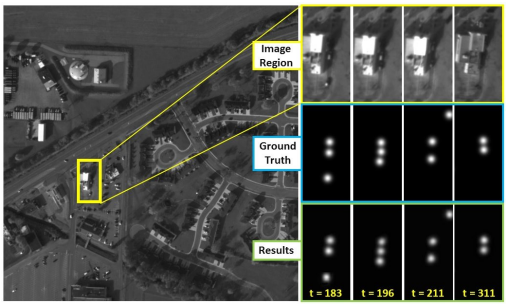


Figure 9: Persistent detection results for AOI 42. The video frame number is marked at the bottom of each column in yellow. **Top Row:** Highlighted image region at each of the four times. **Middle Row:** Ground-truth heat-map. **Bottom Row:** Output heat-map without any post-processing. Notice in the first frame the black car in the shadow of the building is nearly invisible to the naked eye. Additionally due to motion parallax, in the last column the white vehicle is nearly completely occluded by the building, but detection is maintained.

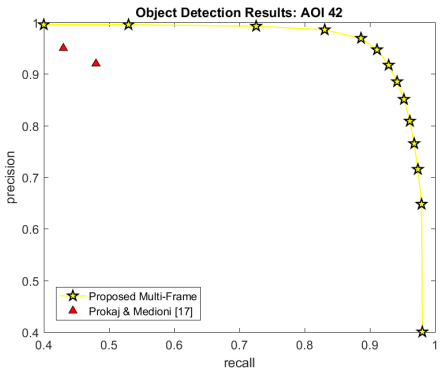


Figure 10: Precision-Recall curve for AOI 42 on persistent detection. No ground-truth coordinates were removed in these results.

Acknowledgement

The authors would like to acknowledge Lockheed Martin for the funding of this research.

References

[1] AFRL, Wright-Patterson Air Force Base (WPAFB) dataset. http://sdms.afrl.af.mil/index.php?collection=wpafb2009, 2009. 1, 4, 5

[2] C. Aeschliman, J. Park, and A. C. Kak. Tracking vehicles through shadows and occlusions in wide-area aerial video. *IEEE Transactions on Aerospace and Electronic Systems*, 50(1):429–444, January 2014. 7

[3] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *Proceedings of the Second International Conference on Human Behavior Unterstanding*, HBU'11, pages 29–39, Berlin, Heidelberg, 2011. Springer-Verlag. 2

[4] A. Basharat, M. Turek, Y. Xu, C. Atkins, D. Stoup, K. Fieldhouse, P. Tunison, and A. Hoogs. Real-time multi-target tracking at 210 megapixels/second in wide area motion imagery. In *IEEE Winter Conference on Applications of Computer Vision*, pages 839–846, March 2014. 2, 6

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1

[6] R. Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[8] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, Jan 2013. 2

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, June 2014. 2

[11] M. Keck, L. Galup, and C. Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 441–448, Jan 2013. 2, 7

[12] P. Kent, S. Maskell, O. Payne, S. Richardson, and L. Scarff. Robust background subtraction for automated detection and tracking of targets in wide area motion imagery. volume 8546, pages 85460Q–85460Q–12, 2012. 7

[13] P. Liang, H. Ling, E. Blasch, G. Seetharaman, D. Shen, and G. Chen. Vehicle detection in wide area aerial surveillance using temporal context. In *Proceedings of the 16th International Conference on Information Fusion*, pages 181–188, July 2013. 7

[14] T. T. Nguyen, H. Grabner, H. Bischof, and B. Gruber. Online boosting for car detection from aerial images. In *2007 IEEE International Conference on Research, Innovation and Vision for the Future*, pages 87–95, March 2007. 1

[15] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1913–1921, Dec 2015. 3

[16] T. Pollard and M. Antone. Detecting and tracking all moving objects in wide-area aerial video. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–22, June 2012. 2, 7

[17] J. Prokaj and G. Medioni. Persistent tracking for wide area aerial surveillance. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1186–1193, June 2014. 2, 6, 7, 8

[18] V. Reilly, H. Idrees, and M. Shah. Detection and tracking of large number of targets in wide area surveillance. In *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III*, ECCV'10, pages 186–199, Berlin, Heidelberg, 2010. Springer-Verlag. 1, 2, 3, 4, 6, 7, 8

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 1, 3

[20] R. Ruskone, L. Guigues, S. Airault, and O. Jamet. Vehicle detection on aerial images: a structural approach. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 3, pages 900–904 vol.3, Aug 1996. 1

[21] I. Saleemi and M. Shah. Multiframe many—many point correspondence for vehicle tracking in high density wide area aerial videos. *Int. J. Comput. Vision*, 104(2):198–219, Sept. 2013. 2, 3, 7, 8

[22] X. Shi, H. Ling, E. Blasch, and W. Hu. Context-driven moving vehicle detection in wide area motion imagery. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2512–2515, Nov 2012. 7

[23] L. W. Sommer, M. Teutsch, T. Schuchert, and J. Beyerer. A survey on moving object detection for wide area motion imagery. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016. 1, 2, 6, 7, 8

[24] M. Teutsch and M. Grinberg. Robust detection of moving vehicles in wide area motion imagery. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1434–1442, June 2016. 2, 3, 6, 7, 8

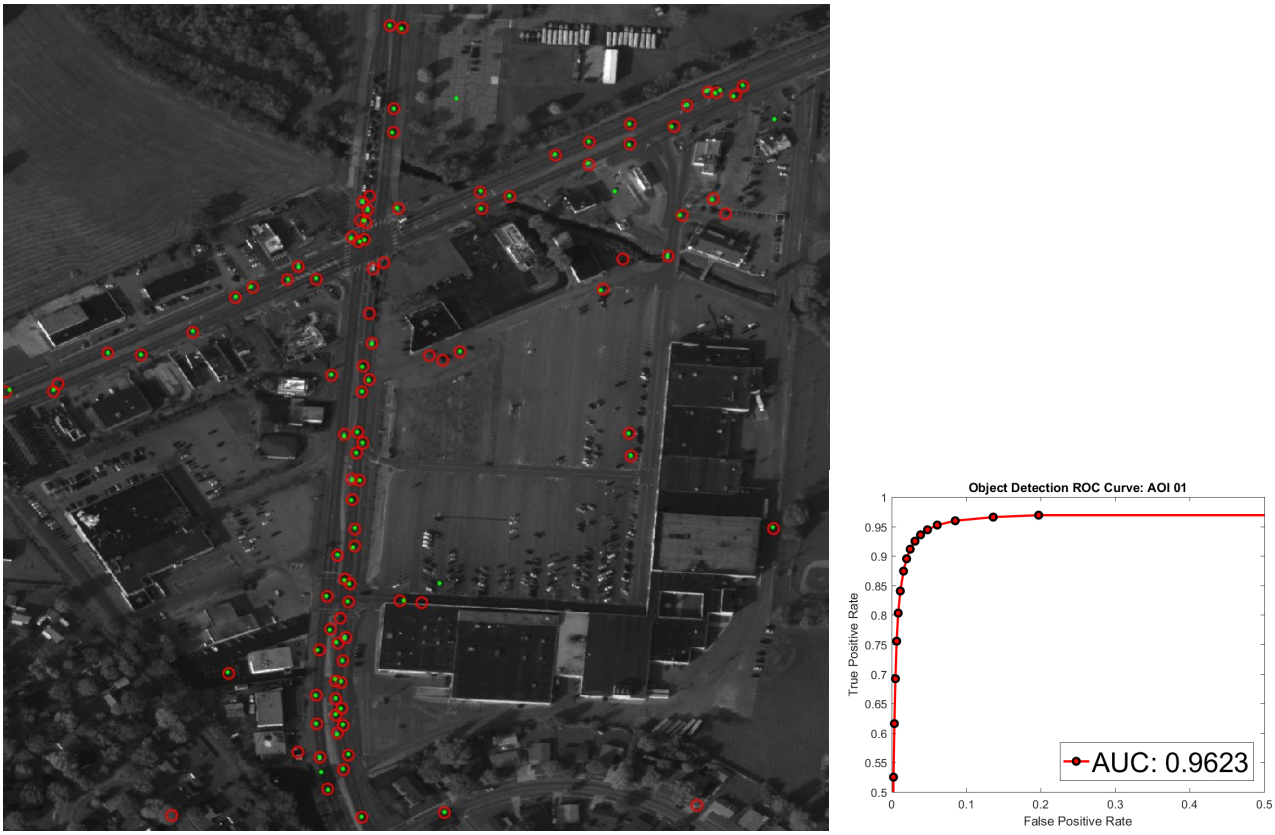
[25] J. Xiao, H. Cheng, H. Sawhney, and F. Han. Vehicle detection and tracking in wide field-of-view aerial video. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 679–684, June 2010. 2, 7

[26] T. Zhao and R. Nevatia. Car detection in low resolution aerial image. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 710–717 vol.1, 2001. 1

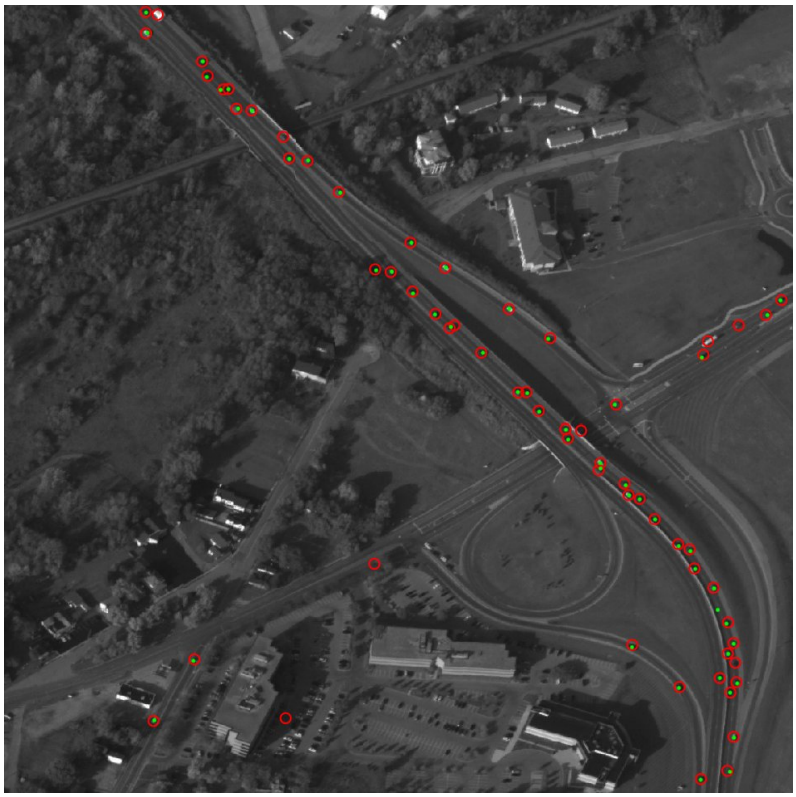
8 Supplemental Materials

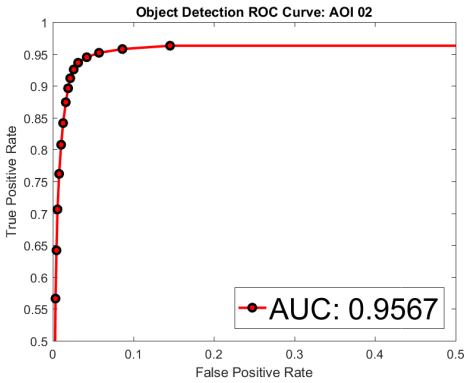
8.1 Qualitative Results and ROC Curves

Figure 11: **Left Column:** Exemplars for the aligned and cropped scenes from the WPAFB 2009 dataset. Red Circles are centered on ground truth coordinates. Green dots are the final predicted object locations by the proposed framework.
Right Column: Receiver operator curves (ROC) to compliment the precision-recall (PR) curves in the main paper.

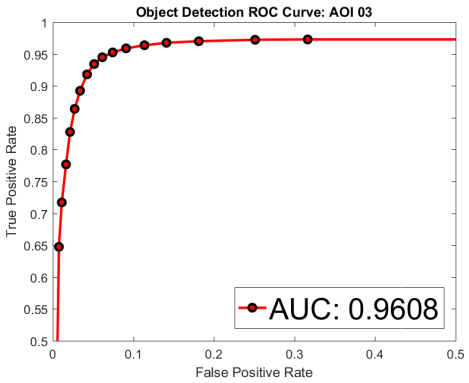


(a) AOI 01: Multi-frame heat-map approach

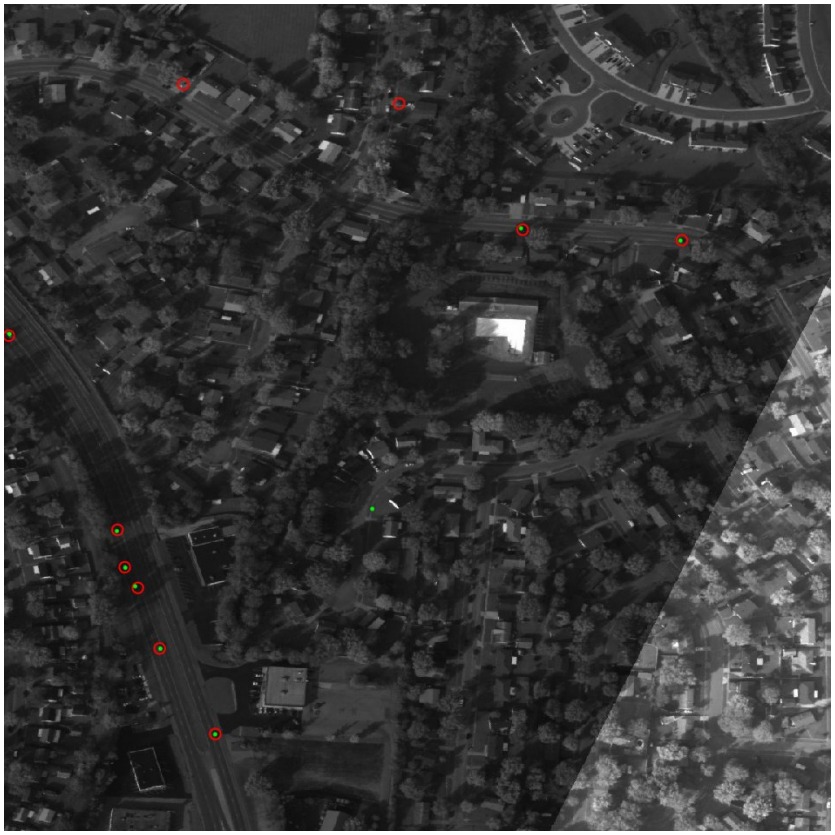


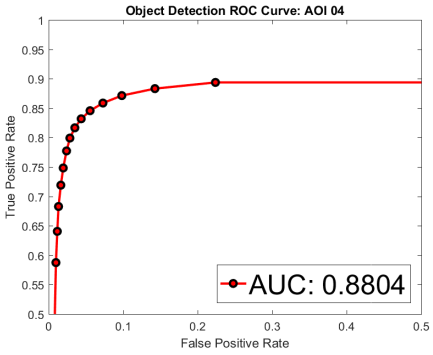


(b) AOI 02: Multi-frame heat-map approach

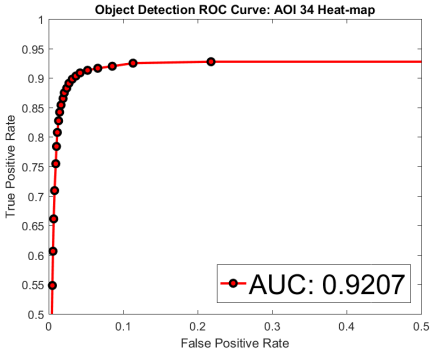


(c) AOI 03: Multi-frame heat-map approach

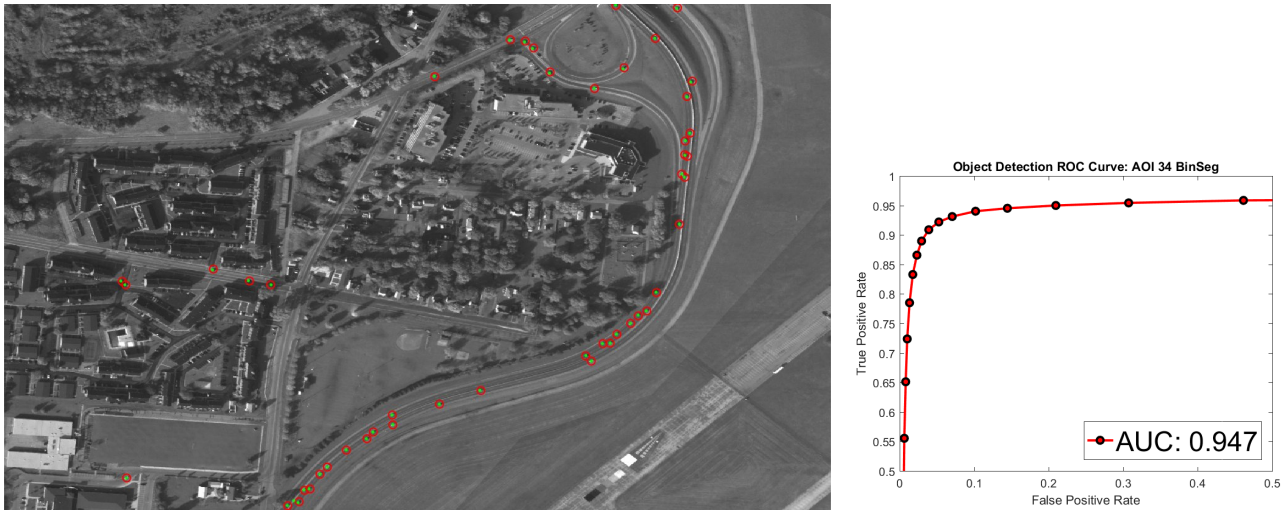




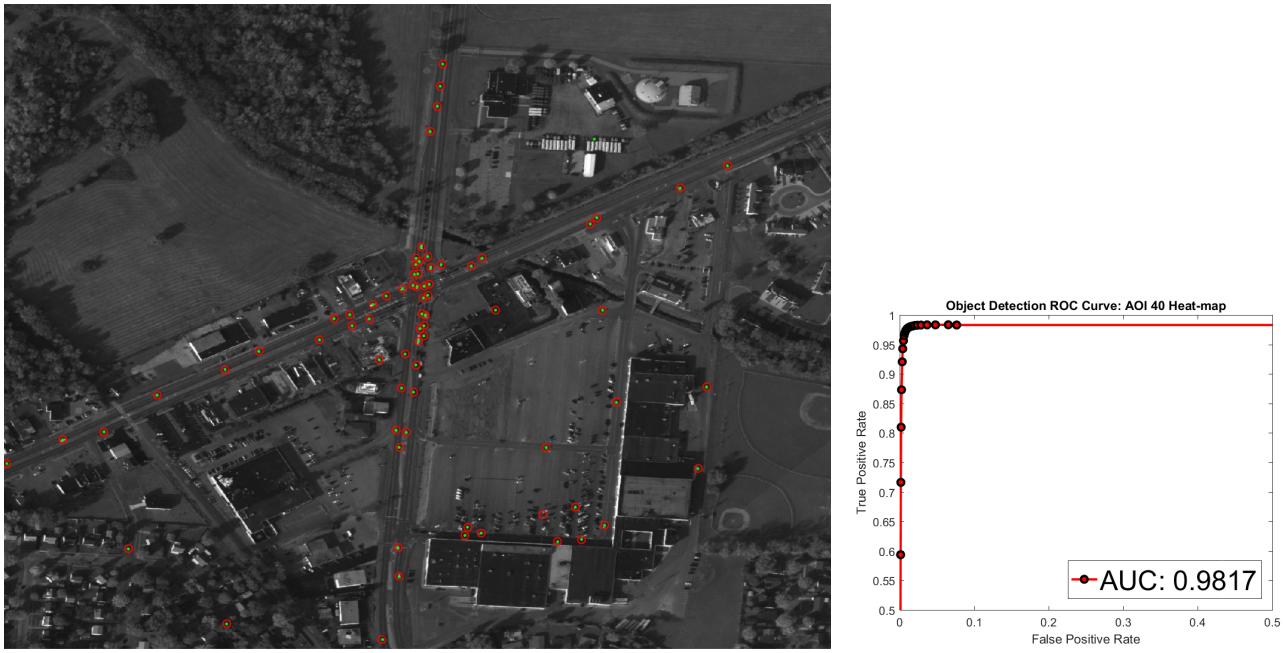
(d) AOI 04: Multi-frame heat-map approach



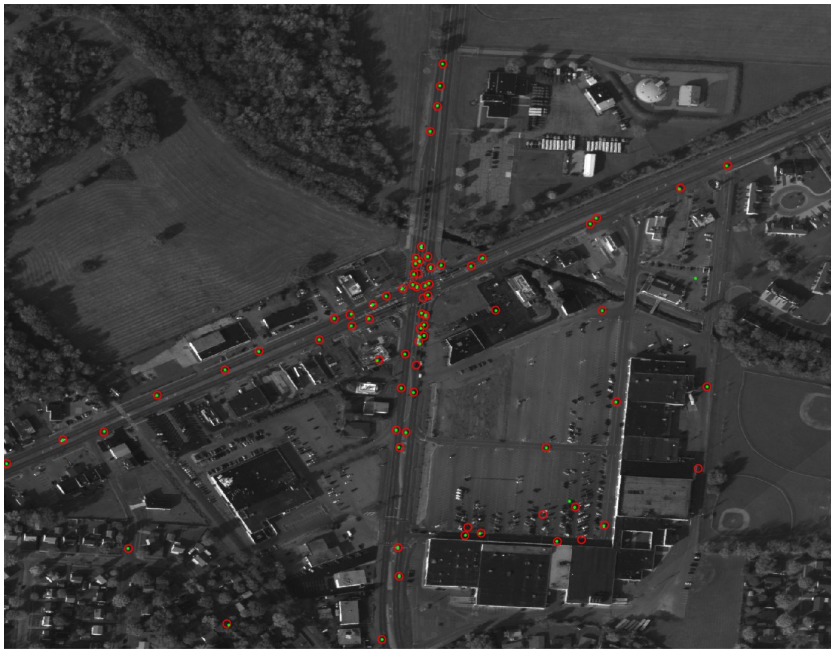
(e) AOI 34: Multi-frame heat-map approach

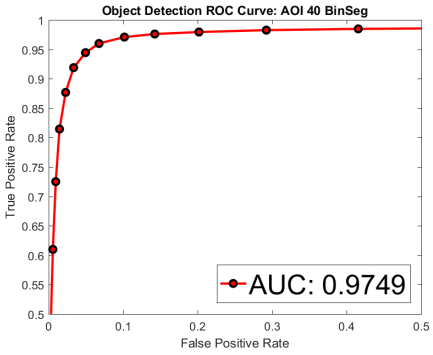


(f) AOI 34: Multi-frame binary segmentation (BinSeg) approach

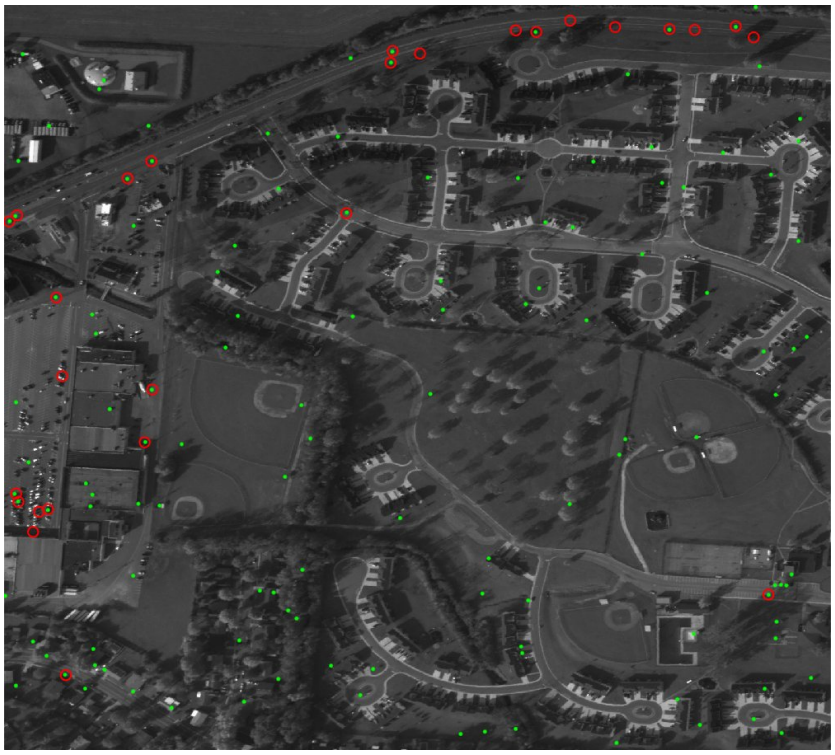


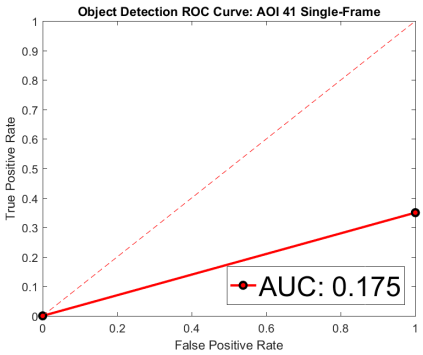
(g) AOI 40: Multi-frame heat-map approach



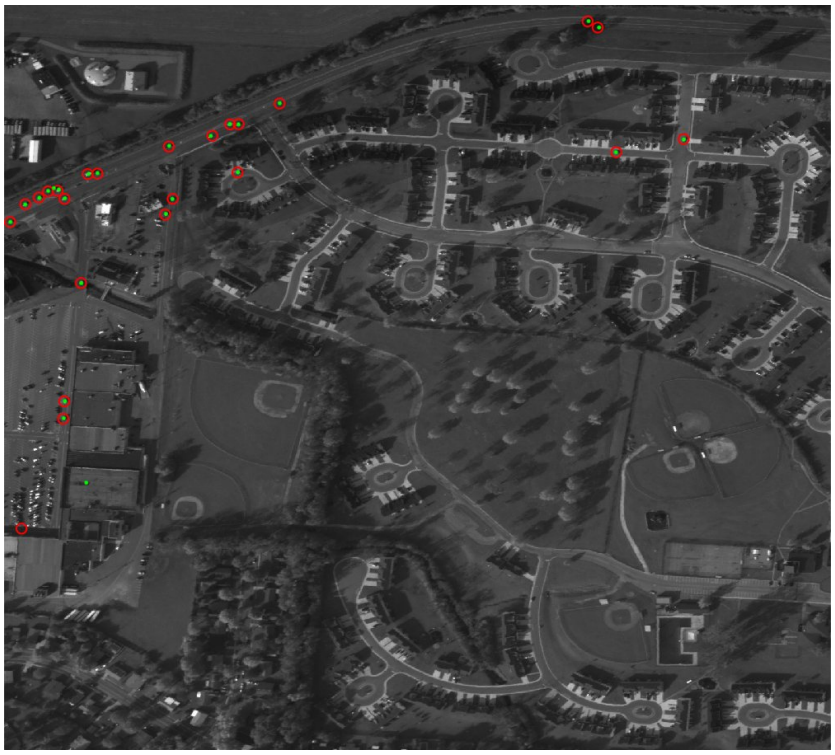


(h) AOI 40: Multi-frame binary segmentation (BinSeg) approach

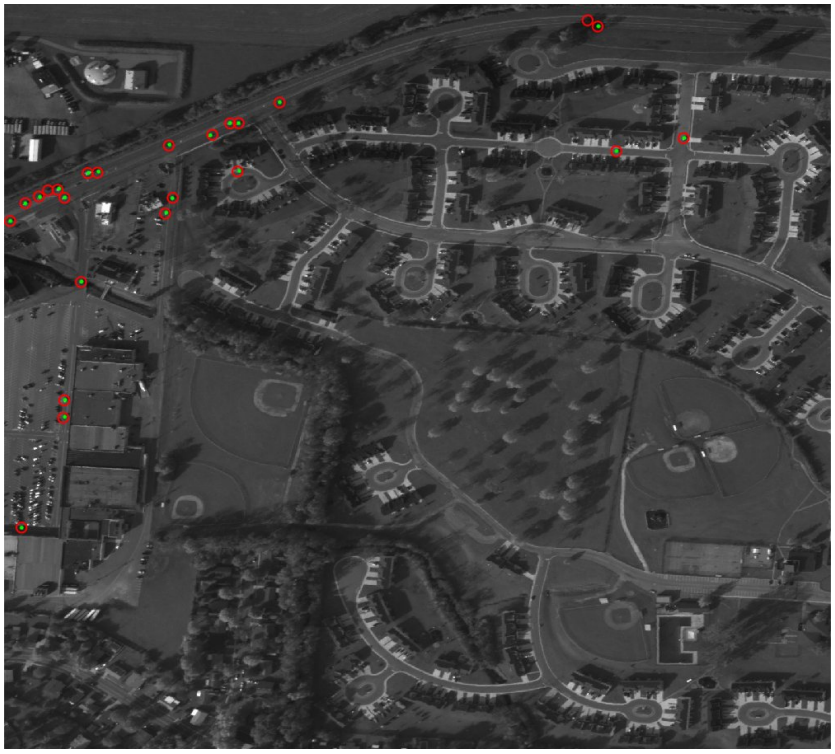




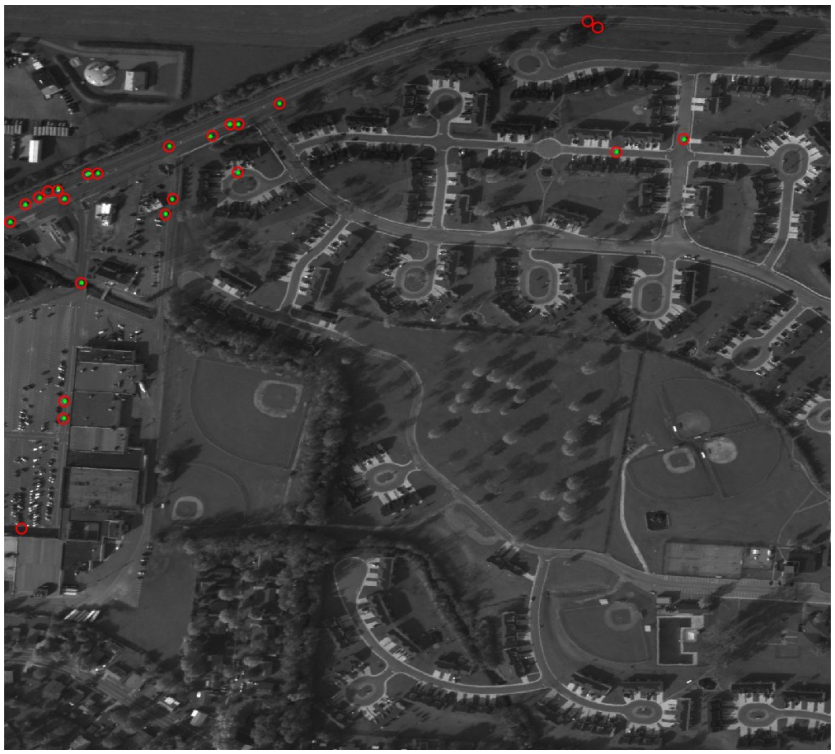
(i) AOI 41: Single-frame heat-map approach

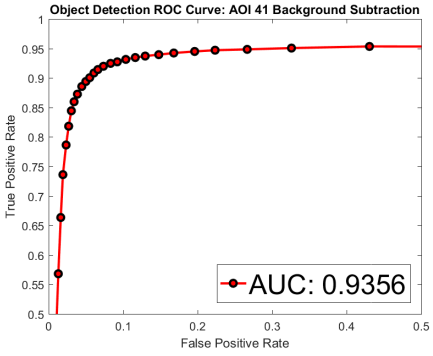


(j) AOI 41: Multi-frame heat-map approach

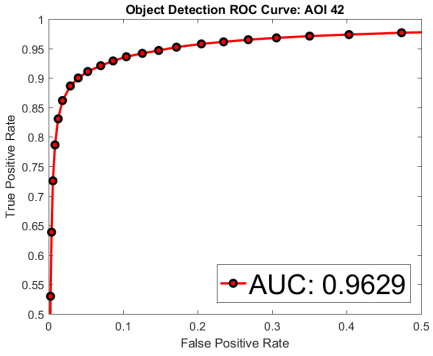


(k) AOI 41: Multi-frame binary segmentation (BinSeg) approach





(l) AOI 41: Multi-frame background subtraction approach



(m) AOI 42: Multi-frame heat-map approach