Appearance and Motion Based Persistent Multiple Object Tracking in Wide Area Motion Imagery

Lars Sommer[1] Wolfgang Krüger[1] Michael Teutsch[2]

[1]Fraunhofer IOSB
Fraunhoferstrasse 1
76131 Karlsruhe, Germany

[2]Hensoldt Optronics GmbH
Carl-Zeiss-Strasse 22
73447 Oberkochen, Germany

firstname.lastname@iosb.fraunhofer.de; michael.teutsch@hensoldt.net

Abstract

Wide Area Motion Imagery (WAMI) data acquired by an airborne sensor for ground observation offers great potential for various applications ranging from the protection of borders and critical infrastructure to city monitoring and surveillance. Persistent multiple object tracking, which is a prerequisite for these applications, is generally based on moving object detection, as the characteristics of existing WAMI datasets, e.g. *weak appearance of objects, impede the usage of appearance based features. Complex and computationally expensive strategies such as exploiting multiple trackers in parallel or classifier-based local search are typically utilized to detect slow and stopping vehicles that are missed by moving object detection. In this paper, we propose a novel and much simpler tracking-by-detection approach for persistent tracking in WAMI data, which avoids such strategies. To overcome limitations caused by image quality of existing WAMI datasets, our proposed tracker was developed on self-acquired WAMI data recorded with a state-of-the-art industrial camera. The improved image quality enables appearance based object detection by Convolutional Neural Networks (CNNs) in WAMI, which we fuse with motion detection to compensate for missed detections in image regions with partial occlusion or shadows. Our proposed tracker is an extension of Deep SORT with modified track management and data association, which is able to yield high recall even in such difficult image regions as well as for slow or stopping vehicles, outperforming state-of-the-art on our self-acquired dataset.*

1. Introduction

Wide Area Motion Imagery (WAMI) data is typically acquired by an airborne sensor for ground observation. The goal is to achieve a large ground coverage of several square kilometers at a detail level that enables the detection and tracking of all relevant objects on the ground such as vehicles or even pedestrians. Potential applications that can be supported with this kind of data range from the protection of borders and critical infrastructure to city monitoring and surveillance and even to enabling smart city capabilities such as adaptations to traffic flow in real-time. A WAMI sensor is usually mounted on a blimp in order to be quasi-stationary.

Persistent multiple object tracking [47] is generally based on moving object detection due to specific characteristics of WAMI data: the weak appearance of objects in existing WAMI datasets such as WPAFB [65] or CLIF [64] primarily hinders the usage of appearance based object detection methods [26, 30, 66]. Though recently proposed spatio-temporal Convolutional Neural Networks (CNNs) clearly outperform conventional moving object detection methods such as frame differencing or background subtraction [25, 30], trackers solely relying on motion detections become unreliable when objects slow down or stop [14, 47]. To achieve persistent tracking, different strategies have been proposed in the literature, *e.g.* the usage of an additional regression tracker in parallel [4, 47] or the classifier-based search in a small local context by utilizing hand-crafted appearance features [25, 62, 70].

In this paper, we propose a novel tracking-by-detection approach for persistent tracking in WAMI data, which avoids expensive additional trackers and the usage of hand-crafted appearance features for local search. Leveraging the impressive results of CNNs for object detection in images [35, 37, 50, 52], tracking-by-detection [7, 10, 68, 69] has become a popular state-of-the-art approach for visual tracking [5, 20] and has demonstrated its potential on various tracking benchmarks [38, 39], including drone-based video [19, 74]. This motivates us to investigate whether these promising results can be transferred to WAMI.

To benefit from recent advances in imaging technologies that are able to overcome limitations of existing WAMI datasets, in particular the weak appearance of objects due to noisy gray-value images, our proposed tracker is developed on self-acquired WAMI data recorded with an industrial state-of-the-art aerial camera. We demonstrate that the improved image quality facilitates appearance based object detection by CNNs for persistent tracking in WAMI. To compensate for missed detections in image regions with partial occlusion or shadows, we fuse appearance based detection with motion based detection, yielding high recall in such difficult regions as well as for slow or stopping vehicles.

We extend the popular multiple object tracker *Simple Online and Realtime Tracking with a deep association metric* (Deep SORT) [69] to show that motion and appearance based object detections can be successfully combined by a much simpler data association than multiple hypothesis tracking [25, 27, 62] or by leveraging multiple trackers in parallel [4, 14, 47]. The visual descriptor needed for data association is directly pooled from features already computed by the CNN-based detector. We further perform adaptions to account for the characteristics of WAMI data, *e.g.* the low frame rate.

In quantitative experiments, we demonstrate that our proposed tracker achieves favourable tracking results for all of the evaluated image sequences and outperforms the state-of-the-art WAMI tracker proposed in [25]. These results also demonstrate the excellent transferability of CNN-based motion detection, which was originally trained on the WPAFB dataset.

The main contributions of our work can be summarized as:

- We propose a novel combination of appearance based and motion based detections for persistent tracking in WAMI.
- We extend the popular tracker Deep SORT in order to account for the characteristics of WAMI and integrate the combined detections.
- Our proposed tracker outperforms state-of-the-art on our self-acquired dataset.
- We demonstrate the excellent transferability of CNN-based motion detection trained on the WPAFB dataset to visually very different WAMI data.

The remainder of this paper is organized as follows. In Section 2, we first give an overview about appearance based object detection and motion based object detection in WAMI. Then, existing approaches for multiple object tracking in WAMI are summarized. Our proposed tracking pipeline is presented in Section 3. In Section 4, we first introduce the self-acquired WAMI dataset followed by a quantitative and qualitative evaluation of our proposed approach. Finally, we conclude in Section 5.

2. Related Work

2.1. Appearance based object detection

Deep learning based detection methods that solely rely on spatial and appearance information achieve state-of-the-art results in numerous fields of application. The most prominent of these methods are Faster R-CNN [52], SSD [37], YOLO [49] and their variants, which exploit multiple feature maps [11, 21, 35, 36, 50]. While these methods are typically not applied on WAMI so far due to poor image quality or inappropriate annotations [30], deep learning based detection methods have been widely adopted for object detection in single aerial imagery [1, 17, 18, 23, 48, 53, 54, 58, 59, 57, 60, 67]. Several adaptations, *e.g.* appropriate feature map resolutions and anchor box scales, have been proposed in order to account for the characteristics of aerial imagery, in particular the small object dimensions [1, 53, 54, 58, 60]. To further improve aerial object detection, common procedure is the exploitation of multiple feature maps [17, 18, 23, 48, 59, 67]. A more detailed overview about deep learning based detection methods for aerial imagery is given in [32].

2.2. Motion based object detection in WAMI

Conventional methods for moving object detection in WAMI are based on either frame differencing [27, 45, 55, 63, 70] or background subtraction [28, 34, 40, 45, 51, 56]. A comprehensive survey on these conventional methods is provided in [61]. Recently, LaLonde *et al.* [30] demonstrated the potential of spatio-temporal CNNs for moving object detection, outperforming conventional methods by a large margin on the WPAFB dataset [65]. The proposed FoveaNet, which comprises a sequence of convolutional layers, takes multiple adjacent frames as input and outputs a heatmap for predicted object locations. To reduce the search area and thus, the computational effort, a region proposal network termed ClusterNet is applied prior to FoveaNet. In [44], the authors adopted FoveaNet for moving object detection in satellite videos. By integrating sublayers with different kernel sizes, Heo *et al.* [26] modify FoveaNet for moving object detection in oblique images, yielding improved detection results. Similar to FoveaNet, Hartung *et al.* [25] employs a spatio-temporal CNN, which takes five consecutive frames as input to detect moving objects in WAMI. Canepa *et al.* [12] propose a spatio-temporal network for real-time detection of small moving objects. For this, three consecutive frames are passed pairwise through two separate CNNs rather than using all frames as input. Instead of using multiple frames as input, Vella *et al.* [66] generate a background context frame, which is passed together with the current frame through a sequence of convolutional layers to predict vehicle locations. An alternative approach is the usage of conventional background subtraction to generate region proposals, which are classified by applying a small CNN to suppress false alarms caused by parallax or registration artifacts [73]. Li *et al*. [33] combine conventional moving object detection and appearance based detection by using the resulting maps from two-frame differencing and the original RGB image as input for Faster R-CNN to detect weak moving objects in remote sensing videos. In [2], the authors apply flux tensor spatio-temporal filtering to detect vehicles in aerial videos. To reduce the number of false detections, appearance based detections that are generated using a deep learning based detection method are fused with the motion based detections.

2.3. Multiple object tracking in WAMI

In the literature, there exist multiple approaches for multiple object tracking in WAMI. Perera *et al*. [43] use nearest neighbor association to form short tracklets, which are linked to tracks by applying the Hungarian algorithm. Reilly *et al*. [51] adopt the Hungarian algorithm for association of detections as well. Saleemi and Shah [55] propose an alternative tracking approach based on an object-centric association method, which allows the sharing of detections among tracks. Keck *et al*. [27] propose the combination of three-frame differencing and multiple hypothesis tracking for WAMI. Chen and Medioni [13] propose a tracker based on motion propagation detection association by iteratively propagating motion information and optimizing an objective function at each frame. In [66], the authors apply a Kalman filter to generate tracks from detections obtained by a CNN for moving object detection. Pelapur *et al*. [41] introduce a track-before-detect approach based on fusing multiple sources of information about the target and its environment. In [42], the authors propose an approach to update the target appearance model within a tracking scheme comprised of a rich feature set and a motion model. Al-Shakarji *et al*. [3] propose a two-step data association scheme for robust multiple object tracking in WAMI. Spatial information is used to generate reliable short-term tracklets, which are linked globally using discriminative features and tracklets history. Zhou and Maskell [73] propose the usage of a Gaussian Mixture Probabilistic Hypothesis filter for tracking and a regression CNN to predict the positions of moving objects. However, the reliance solely on moving object detection impedes the handling of slow or stopping vehicles. Xiao *et al*. [70] attempt to track stopping vehicles by using appearance and shape templates. To improve the association, road and spatial constraints are considered, which is costly in the applied Hungarian algorithm. Prokaj and Medioni [47] propose to run two trackers in parallel: a detection based tracker for initialization and reacquisition and a regression tracker based on target appearance templates to overcome missed detections. Basharat *et al*. [4] propose the combination of a data association based tracker and an appearance based tracker, which is applied when the data association tracker fails or a track becomes very slow. In [14], the authors combine a detection based tracker with a local context tracker to handle missing motion detections. To avoid the additional complexity of two parallel trackers, Spraul *et al*. [62] applies a classifier based detector to recover missing motion detections within a multiple hypothesis tracker. Hartung *et al*. [25] improves the multiple hypothesis tracker proposed in [62] by replacing the background subtraction based moving object detection with a spatio-temporal CNN.

3. Methodology

In the following section, we will describe our proposed pipeline for persistent tracking in WAMI, which is schematically given in Figure 1. First, we introduce the applied appearance based detector and the moving object detector. Then, we discuss the functional principle of Deep SORT, which is used as base tracker, and the performed adaptions to account for the characteristics of WAMI.

3.1. Appearance based object detection

We employ Faster R-CNN [52] with Feature Pyramid Network (FPN) [35] as appearance based object detector. Faster R-CNN is comprised of two stages: an initial stage referred to as Region Proposal Network (RPN) generates a set of region candidates, which are classified in the second stage. Both stages share the convolutional layers of the base network and use the output of the last convolutional layer as feature map. The RPN predicts for each feature map location confidence scores about the presence of an object, which are often termed objectness scores, and corresponding coordinates via bounding box regression. For this, a set of pre-defined anchor boxes is used as bounding box reference. Then, a fixed number of region candidates with the highest objectness scores are passed to the classification stage. For each region candidate, corresponding features are extracted via Region of Interest (RoI) pooling and passed through a sequence of fully connected layers, which outputs confidence scores for each object category and refined coordinates. We attach an FPN to the base network, which yields semantically rich features due to the additional top-down pathway. Instead of using a single feature map, multiple pyramid levels are exploited as feature maps.

To train our appearance based object detector, we use the xView dataset [31], which comprises images from WorldView-3 satellites with a Ground Sampling Distance (GSD) of 0.3 meters per pixel. Hence, the xView dataset exhibits characteristics similar to WAMI. The xView dataset

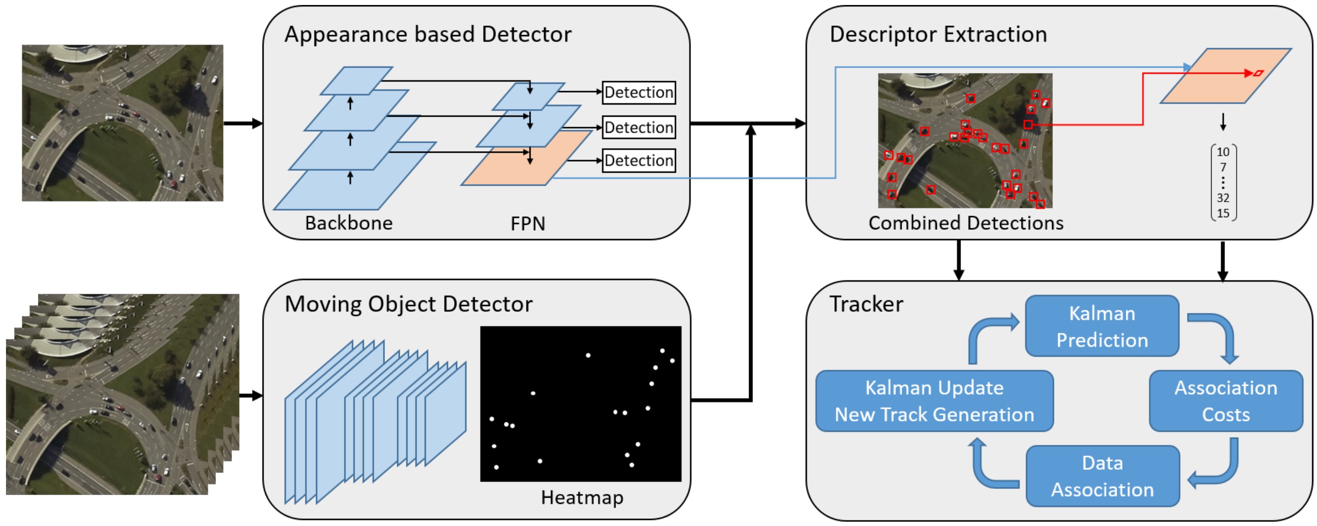

Figure 1. Schematic illustration of our proposed tracking pipeline. Multiple consecutive frames are used as input into a spatio-temporal CNN to detect moving objects. In addition, Faster R-CNN with FPN is applied to further detect slow and stopping vehicles based on appearance features. The moving object and appearance based detections are combined and a descriptor is extracted for each detection by using the features of the lowest pyramid level. The detections are used as input for an extended Deep SORT, which follows the popular tracking-by-detection paradigm.

comprises 60 classes, which can be summarized into 8 meta classes, *i.e.* fixed-wing aircraft, passenger vehicle, truck, railway vehicle, maritime vessel, engineering vehicle, building and others. While our appearance based detector is trained on all 8 meta classes, only the classes passenger vehicle and truck are considered during inference.

In general, detections of small objects such as vehicles in WAMI are provided by the lowest pyramid level [37, 72]. Hence, we exploit the features of the lowest pyramid level to compute the appearance descriptor used for data association in tracking as shown in Figure 1. For this purpose, each detection is projected onto the lowest pyramid level and the corresponding features are extracted via ROI pooling. The output width and height of the ROI pooling is set to 1, yielding a vector with a fixed-length of 256.

3.2. Motion based object detection

We use the CNN-based approach from Hartung *et al.* [25] to detect moving objects in WAMI. Motion based object detection processes a stack of five consecutive aligned images and outputs a heatmap, in which detected objects are represented by Gaussian peaks. Non-maximum suppression in a 3×3 neighborhood and thresholding is sufficient to localize the detected objects at peak centers. The corresponding heatmap intensity is used as a detection score. The detector was trained with annotated sequences from the WPAFB dataset [65].

To compute the visual appearance descriptors needed for data association in tracking, we use fixed-sized square bounding boxes centered on each detected peak. Motion based object detection is executed before appearance based object detection, which receives the fixed-sized bounding boxes as additional input and generates descriptors in the same way as described in the last section. The fixed-sized bounding boxes are also used in the multiple object tracker introduced in the next section.

3.3. Multiple object tracking

We follow the popular tracking-by-detection paradigm and extend Deep SORT [69] for multiple object tracking in WAMI. Deep SORT uses frame-by-frame data association between object detections and existing tracks. Each object detection has to provide a bounding box with confidence score and a visual appearance descriptor that is employed to guide data association. New tracks are initialized from detections, for which there are no associations to existing tracks in the current frame.

Deep SORT uses a Kalman filter with constant velocity motion model to estimate center, size, and aspect ratio of target bounding boxes.

Data association is cast as a linear assignment problem that can be solved by the Hungarian algorithm [29]. Deep SORT introduces assignment costs based on motion as well as on appearance information.

The motion based term uses Mahalanobis distance d_M between predicted bounding box measurements $\tilde{\boldsymbol{b}}_i$ for track T_i and bounding box measurements $\boldsymbol{b}_j$ for detection D_j:

$$d_M^2(i,j) = (\boldsymbol{b}_j - \tilde{\boldsymbol{b}}_i)^T \boldsymbol{S}_i^{-1} (\boldsymbol{b}_j - \tilde{\boldsymbol{b}}_i). \tag{1}$$

Covariance $\boldsymbol{S_i}$ considers estimation uncertainty and is provided by the Kalman filter. Gating threshold t_M is used to prevent association of detections that are too far from the predicted box by requiring $d_M(i,j) < t_M$.

The appearance based term uses Cosine distances between visual appearance descriptors. To this end, an appearance descriptor $\boldsymbol{a}_j$ with $\|\boldsymbol{a}_j\| = 1$ is required for each detection D_j and every track T_i stores a history $\mathcal{H}_i = \{\boldsymbol{a}_{i,1}, \ldots, \boldsymbol{a}_{i,N(i)}\}$ of appearance descriptors for the (at most) last $N(i)$ associated detections. Then, the appearance based distance between detection D_j and track T_i is given by the minimum distance between the appearance descriptor for D_j and the descriptors stored in the history of track T_i:

$$d_A(i,j) = \min\{1 - \boldsymbol{a}_j^T \boldsymbol{a}_{i,k} \mid \boldsymbol{a}_{i,k} \in \mathcal{H}_i\}. \qquad (2)$$

Gating threshold t_A is used to select admissible associations by requiring $d_A(i,j) < t_A$.

Like [69], we combine both distance terms by a weighted sum to get the final cost for admissible associations between detections and tracks:

$$c(i,j) = \lambda d_A(i,j)/t_A + (1-\lambda) d_M(i,j)/t_M. \qquad (3)$$

For this, the weighting coefficient $\lambda \in [0,1]$ is introduced.

In order to use Deep SORT for WAMI, we propose to modify track management and to fuse tracking by appearance based detections with tracking by motion based detections. The idea is to initialize tracks only from motion based detections and to use appearance based detections only for persistent tracking.

Track management. Deep SORT distinguishes tracking modes *tentative* and *confirmed*. New tracks start in mode *tentative* and need successful detection associations in each of the first three frames to survive and to switch to mode *confirmed*. For tentative tracks data association is done by intersection-over-union (IoU) between tracking and detection bounding boxes. While IoU is suitable for high frame rate video (*e.g.* 25 Hz), object motion in low frame rate WAMI (*e.g.* 1 - 2 Hz) is generally to large to have overlapping bounding boxes between adjacent image frames, thus preventing track initialization by IoU. Therefore, we use association costs from Eq. 3 also for track initialization.

When using both appearance based and motion based detections, we initialize new tracks from unassociated motion based detections only and require, that the first two associations must be with motion based detections.

In a final post-processing step, we remove the most recently appended object positions in a track, until the last true detection assignment is found.

Fusing detections by preprocessing. We propose to fuse appearance and motion based detections by running a non-maximum suppression for all detection bounding boxes giving priority to motion based detections. Thereby, appearance based detections having sufficiently large IoU with motion based detections are suppressed.

Figure 2. Image example from our self-acquired WAMI data with evaluation region E01.

Fusing appearance- and motion based tracking. During tracking, we wish to utilize appearance based detections only for slow or temporarily stationary objects, i.e. for persistent tracking. Therefore, we adjust association costs between tracks and appearance based detections:

- Association of appearance based detections is only allowed for slow tracks with maximum velocity v_{max}.
- To decrease the risk of incorrect persistent tracking associations in dense traffic, a reduced gating area is used for appearance based detections. We require that the Euclidean distance between admissible bounding box centers must be no greater than position threshold t_{pos}.

In our experiments, we used $v_{max} = 20$ pixel/frame and distance threshold $t_{pos} = 10$ pixel.

4. Experimental Results

In this section, we first introduce the self-acquired WAMI dataset, which facilitates the usage of appearance based object detection. Then, we describe the experimental setup and present evaluation results in quantitative and qualitative manner.

4.1. Data

For our experiments, we use a self-acquired WAMI dataset. Therefore, we took an industrial, off-the-shelf camera that is stabilized and already certified for airborne missions. This camera is able to acquire images at a resolution of 150 megapixels at a frame rate of 2 Hz and with three color channels (Bayer RGB). Each image then has a spatial resolution of $14,204 \times 10,652$ pixels. This camera was mounted on a helicopter. After reaching the desired altitude of about 2,200 meters above ground, the helicopter hovered at the same position for about ten minutes to create one image sequence. Multiple sequences were recorded at different times of the day and with different terrain properties such as urban, rural, or mixed scenarios. The GSD is 0.256 meters per pixel and the Common Operational Picture (COP), *i.e*. the ground coverage, is 9.88 km^2. An example image of the self-acquired dataset is given in Figure 2.

There are multiple differences compared to already existing WAMI datasets such as WPAFB [65] or CLIF [64]: we use RGB color images instead of gray-value images. In this way, we can utilize color information for image processing. As we use a rather new imaging device, we can see an improved image quality in terms of a reduced noise level or an increased detail visibility even though we have nearly the same GSD compared to the WPAFB dataset. We use only one high-resolution camera instead of multiple cameras arranged in a camera array or matrix. In this way, image mosaicking as it was mandatory in the past [46] is not needed anymore. This not only speeds up processing as we can omit this step in the processing pipeline, but also we can avoid mosaic seam artifacts that affected WAMI data processing in the past [27]. One drawback of this approach is the reduced ground coverage of only about 10 square kilometers, but at the same time the sensor hardware setup is optimized in size and weight with less than 20 kg. Such a sensor system can be carried as payload not only by a blimp but even by larger drones.

4.2. Evaluation

To evaluate the proposed tracking methods, we selected three regions of interest and performed image alignment via homography estimation [24] to compensate for camera motion. After warping with the estimated homographies we obtained well aligned image sequences, each consisting of 120 frames with dimensions 1536×1024 pixels. Similar to the evaluation procedure commonly used with the WPAFB dataset [25, 47, 62] we annotated a persistent tracking ground truth, that contains all vehicles that move at least once. An annotated track starts in the frame a vehicle begins to move and remains alive for the rest of the sequence or until the vehicle leaves the region of interest. Therefore, tracks may include *e.g*. vehicles stopping at intersections or parking vehicles, provided that the vehicles had been in motion before. Annotated tracks contain vehicle position (center of object), vehicle identifier, and frame number.

Using existing terrain map data [8, 22] or utilizing semantic segmentation in order to segment road regions [71] are common procedures to identify regions of interest such as streets, roads, or highways. In the following, we use image masks derived from OpenStreetMap [16] to focus the evaluation on traffic areas and to avoid annotation ambiguities, *e.g*. at forested parking areas.

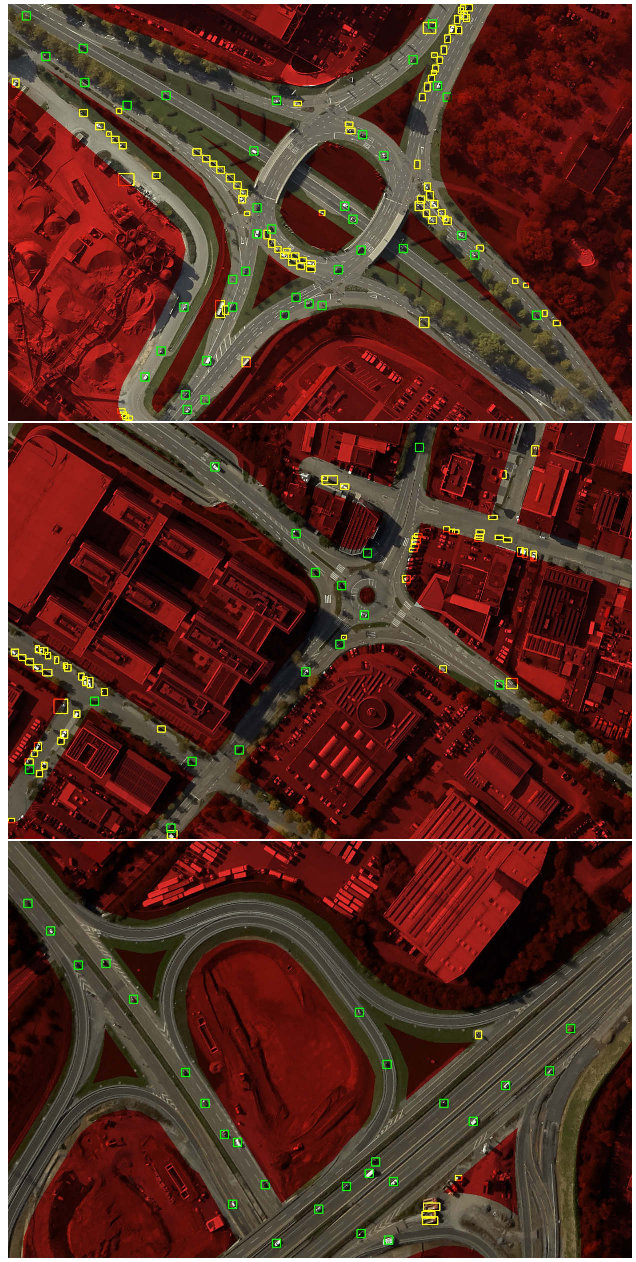

Figure 3. From top to bottom, example frames from evaluation regions E01 to E03 with motion based (green) and appearance based (yellow) object detections after non-maximum suppression giving priority to motion based detections. Masks (red overlay) derived from OpenStreetMap [16] are used to focus evaluation on traffic areas and to avoid annotation ambiguities, *e.g*. at forested parking areas.

Our annotated evaluation sequences are shown in Figure 3. Region E01 contains many slow, stopping or starting vehicles posing substantial challenges for persistent tracking. Thus, moving object detection alone is clearly not sufficient to handle Region E01. In Region E02 we have fewer vehicles needing persistent tracking, but areas in which it is quite difficult to detect all vehicles due to shadows or occlusions. Region E03 shows a highway situation with fast vehicles, no need for persistent tracking, but many overtaking maneuvers.

We use the same evaluation metrics as in [25], *i.e*. precision, recall, F-score, ID/GT (identity switches per number of ground truth tracks), and MOTA (multiple object tracking accuracy [6]). Since the GSD of our evaluation sequences is very similar to WPAFB data, we use the same distance threshold of 20 pixels as in the literature [25, 30, 61, 62, 73] to decide, whether associations between vehicle positions in ground truth and tracking output yield true or false positives.

Tracking output from methods using only appearance based object detections will contain tracks for moving and for stationary vehicles. In order to evaluate w.r.t persistent tracking ground truth, we need an additional PT-filtering (persistent tracking filtering). To describe PT-filtering, we represent a track with length L as a list of pixel positions $((x,y)_1, ..., (x,y)_L)$. In a first filtering step, we strip all positions from the beginning of a track until sufficient motion is encountered, i.e. until $|(x,y)_{i+1} - (x,y)_i| > t_0$. The second step suppresses stationary tracks, i.e. tracks for which all positions $(x,y)_i$ lie inside of an enclosing box with diagonal $d < d_0 + d_1 L$. In our experiments, we use $t_0 = 3, d_0 = 20$, and $d_1 = 0.2$. We apply the same PT-filtering to the output of all evaluated tracking methods as well as to the ground truth.

We evaluate three variants of our tracking method described in Section 3. The first two variants, DSORT-APP and DSORT-MOT, use appearance based object detections only. The difference between these two variants is, that DSORT-APP uses only appearance based association costs ($\lambda = 1$ in Eq. 3) while DSORT-MOT uses only motion based association costs ($\lambda = 0$). The third variant, DSORT-PT, is our proposed persistent tracking method, which implements the combination of appearance based and motion based object detections described in Section 3. For data association, we rely on the visual appearance descriptor ($\lambda = 1$ in Eq. 3).

The evaluation results are shown in Table 1. We see, that the proposed DSORT-PT achieves the best detection performance (F-Score) as well as the best tracking performance (MOTA, ID/GT) among the three DSORT-variants.

When comparing DSORT-MOT and DSORT-APP, the weak results for MOTA and ID/GT in Region E03 clearly show, that a purely motion based association cost (DSORT-MOT) is not sufficient for low frame rate WAMI. In Region E03, a combination of low frame rate and large vehicle speed leads to inter-frame displacements for vehicles that are comparable to intra-frame vehicle distances and thus,

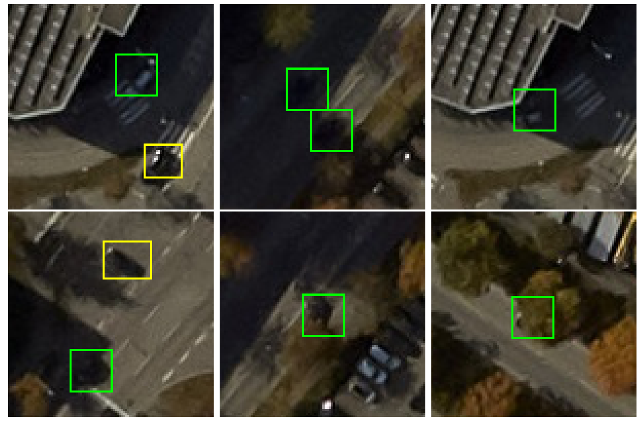

Figure 4. Examples for motion based (green) and appearance based (yellow) object detection from region E02. Appearance based detector is needed to facilitate persistent tracking for vehicles waiting e.g. at intersections (left column). We need motion based detector to find moving vehicles in challenging situations like shadows or occlusion by treetops (middle and right column).

do not facilitate proper track initialization in dense traffic. This is in contrast to multiple object tracking in high frame rate video, where simple bounding box overlap may provide good results [9].

On Region E02, DSORT-PT outperforms both DSORT-MOT and DSORT-APP by a large margin. The reason is low recall of appearance based object detection for vehicles in shadows and vehicles partly occluded by treetops (cf. Figure 4). Thus, we conclude, that currently, even with state-of-the-art detectors, relying on appearance cues only is no solution for multiple object tracking in WAMI and motion based cues are still needed. Combining both cues, proposed tracker DSORT-PT provides favourable results. In this context, we would like to emphasize the excellent transferability of CNN-based motion detection that was trained on the visually very different WPAFB dataset.

We show results for [25] to compare with state-of-the-art for persistent multiple object tracking in WAMI. Hartung *et al*. [25] combined multiple hypothesis tracking with motion detection and a classifier-based detector. They also integrated vehicle-collision tests, clutter handling, and an appearance based similarity measure based on Local Binary Patterns and local variance [15].

Using a much simpler data association, DSORT-PT is able to achieve superior results for F-Score and MOTA on all three evaluation regions. We attribute this performance gain to the advanced appearance based object detector and a more suitable visual descriptor. The strength of tracking multiple hypotheses during data association in [25] is the very low proportion of identity switches on Region E03. On the other hand, results for E03 and E02 are already quite good for DSORT-PT, which is much better on the more

Region	Method	Precision	Recall	F-Score	ID/GT ↓	MOTA
E01	DSORT-MOT	0.928	**0.933**	0.930	0.982	0.842
	DSORT-APP	**0.949**	0.931	**0.940**	0.419	0.873
	DSORT-PT	**0.949**	0.931	**0.940**	**0.389**	**0.874**
	Hartung *et al*. [25]	0.925	0.874	0.899	0.509	0.793
E02	DSORT-MOT	0.858	0.782	0.818	0.690	0.638
	DSORT-APP	0.884	0.776	0.827	0.762	0.658
	DSORT-PT	**0.965**	0.988	**0.976**	0.167	**0.949**
	Hartung *et al*. [25]	0.948	**0.991**	0.969	**0.143**	0.933
E03	DSORT-MOT	0.928	0.976	0.951	2.320	0.792
	DSORT-APP	0.976	0.986	**0.981**	0.150	0.954
	DSORT-PT	0.974	**0.989**	**0.981**	0.095	**0.958**
	Hartung *et al*. [25]	**0.982**	0.974	0.978	**0.015**	0.956

Table 1. Evaluation results for proposed DSORT-PT (combination of appearance based and motion based object detections), variants DSORT-APP and DSORT-MOT using appearance based object detections only, and persistent multiple object tracking from [25]. Smaller values are better for ID/GT (identity switches per number of ground truth tracks).

challenging Region E01.

5. Conclusion

In this paper, we proposed a novel tracking-by-detection approach DSORT-PT for persistent tracking in WAMI data, which avoids expensive additional trackers and the usage of hand-crafted appearance features for local search. To overcome limitations caused by low image quality of existing WAMI datasets, our proposed tracker was developed on self-acquired WAMI data recorded with an industrial state-of-the-art aerial camera. We demonstrated that the improved image quality facilitates appearance based object detection by CNNs for persistent tracking in WAMI. We also showed, that a combination of appearance based detection with motion detection is needed to compensate for missed detections in image regions with partial occlusion or shadows. Our multiple object tracker is an extension of Deep SORT with modified track management and data association and was able to yield high recall even in such difficult regions as well as for slow or stopping vehicles. In quantitative experiments, we demonstrated that our tracker achieves favourable results and outperforms state-of-the-art on our self-acquired dataset. We also demonstrated the excellent transferability of CNN-based motion detection, which was trained on the WPAFB dataset. Regarding future work, we plan to investigate visual appearance descriptors learned from re-identification tasks.

References

[1] Oliver Acatay, Lars Sommer, Arne Schumann, and Jürgen Beyerer. Comprehensive evaluation of deep learning based detection methods for vehicle detection in aerial imagery. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[2] Noor Al-Shakarji, Filiz Bunyak, Hadi Aliakbarpour, Guna Seetharaman, and Kannappan Palaniappan. Multi-cue vehicle detection for semantic video compression in georegistered aerial videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–65, 2019.

[3] Noor M Al-Shakarji, Filiz Bunyak, Guna Seetharaman, and Kannappan Palaniappan. Robust multi-object tracking for wide area motion imagery. In *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–5. IEEE, 2018.

[4] Arslan Basharat, Matt Turek, Yiliang Xu, Chuck Atkins, David Stoup, Keith Fieldhouse, Paul Tunison, and Anthony 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. IEEE, 2014.

[5] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[6] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[7] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[8] Erik Blasch, Chun Yang, Jesus Garcia, Lauro Snidaro, and James Llinas. Context-enhanced information fusion. In Lauro Snidaro, Jesus Garcia, James Llinas, and Erik Blasch, editors, *Advances in Computer Vision and Pattern Recognition (ACVPR)*, pages 73–97. Springer, 2016.

[9] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[10] Erik Bochinski, Tobias Senst, and Thomas Sikora. Extending iou based multi-object tracking by visual information. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[11] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[12] Alessio Canepa, Edoardo Ragusa, Rodolfo Zunino, and Paolo Gastaldo. T-rexnet—a hardware-aware neural network for real-time detection of small moving objects. *Sensors*, 21(4):1252, 2021.

[13] Bor-Jeng Chen and Gérard Medioni. Motion propagation detection association for multi-target tracking in wide area aerial surveillance. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2015.

[14] Bor-Jeng Chen and Gérard Medioni. Exploring local context for multi-target tracking in wide area aerial surveillance. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 787–796. IEEE, 2017.

[15] Mickael Cormier, Lars Sommer, and Michael Teutsch. Low resolution vehicle re-identification based on appearance features for wide area motion imagery. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2016.

[16] Kevin Curran, John Crumlish, and Gavin Fisher. Openstreetmap. *International Journal of Interactive Communication Systems and Technologies (IJICST)*, 2(1):69–78, 2012.

[17] Zhipeng Deng, Hao Sun, Shilin Zhou, Juanping Zhao, Lin Lei, and Huanxin Zou. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, 145:3–22, 2018.

[18] Peng Ding, Ye Zhang, Wei-Jian Deng, Ping Jia, and Arjan Kuijper. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 141:208–218, 2018.

[19] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.

[20] Patrick Emami, Panos M Pardalos, Lily Elefteriadou, and Sanjay Ranka. Machine learning methods for data association in multi-object tracking. *ACM Computing Surveys (CSUR)*, 53(4):1–34, 2020.

[21] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[22] Jianjun Gao, Haibin Ling, Erik Blasch, Khanh Pham, Zhonghai Wang, and Genshe Chen. Pattern of Life from WAMI Objects Tracking based on Context Aware Tracking and Information Network Models. In *Proceedings of SPIE Vol. 8745*, 2013.

[23] Wei Guo, Wen Yang, Haijian Zhang, and Guang Hua. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. *Remote Sensing*, 10(1):131, 2018.

[24] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[25] Christine Hartung, Raphael Spraul, and Wolfgang Krüger. Improvement of persistent tracking in wide area motion imagery by cnn-based motion detections. In *Image and Signal Processing for Remote Sensing XXIV*, volume 10789, page 107890Q. International Society for Optics and Photonics, 2018.

[26] Won Yeong Heo, Seongjo Kim, DeukRyeol Yoon, Jongmin Jeong, and HyunSeong Sung. Deep learning based moving object detection for oblique images without future frames. In *Automatic Target Recognition XXX*, volume 11394, page 1139403. International Society for Optics and Photonics, 2020.

[27] Mark Keck, Luis Galup, and Chris 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. IEEE, 2013.

[28] Phil Kent, Simon Maskell, Oliver Payne, Sean Richardson, and Larry Scarff. Robust background subtraction for automated detection and tracking of targets in wide area motion imagery. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII*, volume 8546, page 85460Q. International Society for Optics and Photonics, 2012.

[29] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[30] Rodney LaLonde, Dong Zhang, and Mubarak Shah. Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4003–4012, 2018.

[31] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018.

[32] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.

[33] Yuxuan Li, Licheng Jiao, Xu Tang, Xiangrong Zhang, Wenhua Zhang, and Li Gao. Weak moving object detection in optical remote sensing video with motion-drive fusion network. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5476–5479. IEEE, 2019.

[34] Pengpeng Liang, Haibin Ling, Erik Blasch, Guna Seetharaman, Dan Shen, and Genshe Chen. Vehicle detection in wide area aerial surveillance using temporal context. In *Proceedings of the 16th international conference on information fusion*, pages 181–188. IEEE, 2013.

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyra-

mid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[37] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[38] Siwei Lyu, Ming-Ching Chang, Dawei Du, Wenbo Li, Yi Wei, Marco Del Coco, Pierluigi Carcagnì, Arne Schumann, Bharti Munjal, Doo-Hyun Choi, et al. Ua-detrac 2018: Report of avss2018 & iwt4s challenge on advanced traffic monitoring. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[39] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831.

[40] Kannappan Palaniappan, Mahdieh Poostchi, Hadi Aliakbarpour, Raphael Viguier, Joshua Fraser, Filiz Bunyak, Arslan Basharat, Steve Suddarth, Erik Blasch, Raghuveer M Rao, et al. Moving object detection for vehicle tracking in wide area motion imagery using 4d filtering. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2830–2835. IEEE, 2016.

[41] Rengarajan Pelapur, Sema Candemir, Filiz Bunyak, Mahdieh Poostchi, Guna Seetharaman, and Kannappan Palaniappan. Persistent target tracking using likelihood fusion in wide-area and full motion video sequences. In *2012 15th International Conference on Information Fusion*, pages 2420–2427. IEEE, 2012.

[42] Rengarajan Pelapur, Kannappan Palaniappan, and Gunasekaran Seetharaman. Robust orientation and appearance adaptation for wide-area large format video object tracking. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 337–342. IEEE, 2012.

[43] AG Amitha Perera, Chukka Srinivas, Anthony Hoogs, Glen Brooksby, and Wensheng Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 666–673. IEEE, 2006.

[44] Roman Pflugfelder, Axel Weissenfeld, and Julian Wagner. On learning vehicle detection in satellite video. *arXiv preprint arXiv:2001.10900*, 2020.

[45] Thomas Pollard and Matthew 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. IEEE, 2012.

[46] Jan Prokaj and Gerard Medioni. Accurate efficient mosaicking for Wide Area Aerial Surveillance. In *Proceedings of the IEEE Workshop on the Applications of Computer Vision (WACV)*, 2012.

[47] Jan Prokaj and Gerard Medioni. Persistent Tracking for Wide Area Aerial Surveillance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1186–1193, 2014.

[48] Xiaoliang Qian, Sheng Lin, Gong Cheng, Xiwen Yao, Hangli Ren, and Wei Wang. Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion. *Remote Sensing*, 12(1):143, 2020.

[49] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[50] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[51] Vladimir Reilly, Haroon Idrees, and Mubarak Shah. Detection and tracking of large number of targets in wide area surveillance. In *European conference on computer vision*, pages 186–199. Springer, 2010.

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[53] Yun Ren, Changren Zhu, and Shunping Xiao. Small object detection in optical remote sensing images via modified faster r-cnn. *Applied Sciences*, 8(5):813, 2018.

[54] Wesam Sakla, Goran Konjevod, and T. Nathan Mundhenk. Deep multi-modal vehicle detection in aerial isr imagery. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 916–923. IEEE, 2017.

[55] Imran Saleemi and Mubarak Shah. Multiframe many–many point correspondence for vehicle tracking in high density wide area aerial videos. *International journal of computer vision*, 104(2):198–219, 2013.

[56] Xinchu Shi, Haibin Ling, Erik Blasch, and Weiming 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. IEEE, 2012.

[57] Lars Sommer, Tobias Schuchert, and Jürgen Beyerer. Deep learning based multi-category object detection in aerial images. In *Automatic Target Recognition XXVII*, volume 10202, page 1020209. International Society for Optics and Photonics, 2017.

[58] Lars Sommer, Tobias Schuchert, and Jürgen Beyerer. Comprehensive analysis of deep learning-based vehicle detection in aerial images. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2733–2747, 2018.

[59] Lars Sommer, Arne Schumann, Tobias Schuchert, and Jurgen Beyerer. Multi feature deconvolutional faster r-cnn for precise vehicle detection in aerial imagery. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 635–642. IEEE, 2018.

[60] Lars Wilko Sommer, Tobias Schuchert, and Jürgen Beyerer. Fast deep vehicle detection in aerial images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 311–319. IEEE, 2017.

[61] Lars Wilko Sommer, Michael Teutsch, Tobias Schuchert, and Jürgen 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. IEEE, 2016.

[62] Raphael Spraul, Christine Hartung, and Tobias Schuchert. Persistent multiple hypothesis tracking for wide area motion imagery. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017.

[63] Michael Teutsch and Michael Grinberg. Robust detection of moving vehicles in wide area motion imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 27–35, 2016.

[64] U.S. Air Force Research Lab. SDMS: CLIF 2007 Dataset. `https://www.sdms.afrl.af.mil/index.php?collection=clif2007`, 2007.

[65] U.S. Air Force Research Lab. SDMS: WPAFB 2009 Dataset. `https://www.sdms.afrl.af.mil/index.php?collection=wpafb2009`, 2009.

[66] Elena Vella, Anee Azim, Han X Gaetjens, Boris Repasky, and Timothy Payne. Improved detection for wami using background contextual information. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–9. IEEE, 2019.

[67] Chen Wang, Xiao Bai, Shuai Wang, Jun Zhou, and Peng Ren. Multiscale visual attention networks for object detection in vhr remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 16(2):310–314, 2018.

[68] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 107–122. Springer, 2020.

[69] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[70] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, and Feng 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. IEEE, 2010.

[71] Fan Yang, Meng Yi, Yiran Cai, Erik Blasch, Hua-mei Chen, Carolyn Sheaff, Genshe Chen, and Haibin Ling. Multitask Assessment of Roads and Vehicles Network (MARVN). In *Proceedings of SPIE Vol. 10641*, 2018.

[72] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *European Conference on Computer Vision*, pages 323–339. Springer, 2020.

[73] Yifan Zhou and Simon Maskell. Detecting and tracking small moving objects in wide area motion imagery (wami) using convolutional neural networks (cnns). In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2019.

[74] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. Vision meets drones: Past, present and future. *arXiv preprint arXiv:2001.06303*, 2020.