Impact of Georegistration Accuracy on Wide Area Motion Imagery Object Detection and Tracking

Noor Al-Shakarji[1], Ke Gao[1], Filiz Bunyak[1], Hadi Aliakbarpour[1], Erik Blasch[2],
Priya Narayaran[3], Guna Seetharaman[4], Kannappan Palaniappan[1]
[1]Electrical Engineering & Computer Science Department
University of Missouri, Columbia, MO, USA 65211
[2]U.S. Air Force Research Laboratory, USA
[3]U.S. Army Research Laboratory, USA
[4]U.S. Naval Research Laboratory, Washington, D.C. USA
{nmahyd, kegao, bunyak, aliakbarpourh, palaniappank}@missouri.edu;
erik.blasch.1@us.af.mil, priya.narayanan.civ@mail.mil, guna.seetharaman@nrl.navy.mil

***Abstract*—Advances in sensor technologies and embedded low-power processing provide new opportunities for using Wide Area Motion Imagery (WAMI) across a spectrum of mapping and monitoring applications covering large geospatial areas for extended time periods. While significant developments have been made in video analytics for ground or low-altitude aerial videos, methods for WAMI have been limited due to lack of benchmarking datasets, data format complexities, lack of labeled training videos, and high data processing requirements. This paper aims to help advance the broader use of WAMI by evaluating the georegistration accuracy and its impact on downstream video analytics using two benchmark datasets (CLIF 2007, ABQ 2013). In addition to the current intensified interest in using deep learning for aerial object recognition and tracking, this paper motivates the need for further development of more robust and fast georegistration algorithms for multi-camera WAMI systems.**

***Index Terms*—Wide area motion imagery, aerial video, georegistration, object detection, object tracking**

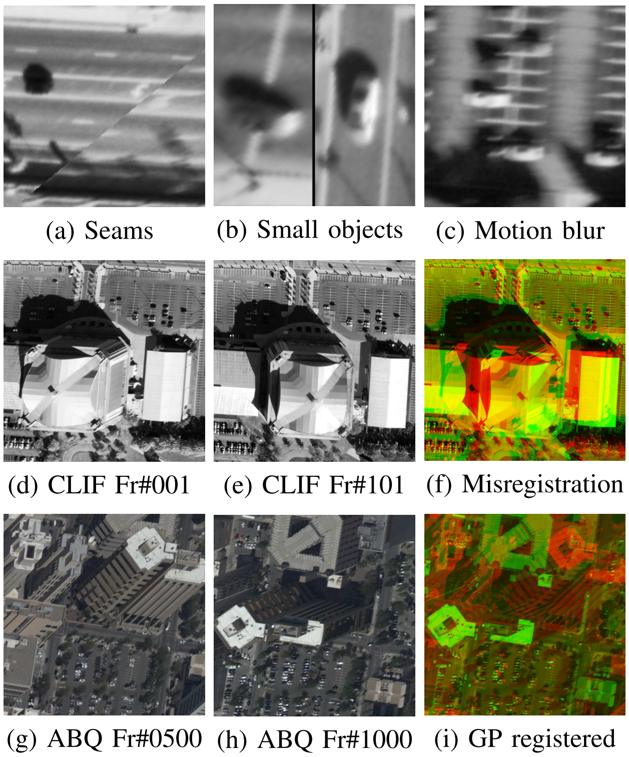
I. Introduction

There has been an exponential increase in aerial motion imagery due to advances in airborne sensor technologies, the increased adoption of manned and unmanned aerial vehicles (UAVs), and the emergence of new applications including aerial delivery, environmental monitoring, smart cities, search and rescue, disaster relief, and precision agriculture. Society is seeing a growing need for robust aerial imagery and video analytics capabilities to take full advantage of data fusion and to meet such application needs [1]. Novel methods, particularly those using artificial intelligence/machine learning (AI/ML), coupled with rapid advances in computational hardware (more powerful, lighter weight, lower energy, lower computing cost) are revolutionizing image processing, pattern recognition, and information fusion (e.g,, WAMI fusion applications [2]).

Wide area motion imagery (WAMI) is characterized by large ground coverage of a few square miles, many objects of interest, and high-altitude oblique viewing geometries. WAMI platforms equipped with orientation sensors circle above a region of interest at a constant altitude, adjusting steadily the orientation of the camera array pointing to a narrow area of interest [3] within the region being imaged. Once georegistered and stabilized, these videos provide a virtual nadir (i.e., downward) view of the region being monitored [3] and enable large-scale surveillance and monitoring for extended periods of time. WAMI exploitation pipelines for object recognition and multitarget tracking have unique challenges such as large camera motion, low frame rate, small object sizes, multi-camera arrays, hundreds to thousands of moving objects per frame, oblique viewing angles, motion blur, parallax effects, shadows, etc., in addition to regular sensor resolution and weather challenges. While significant advancements have been made in the areas of object detection [4]–[6], single-object tracking [7]–[9], and multi-object tracking [10]–[13] thanks largely to benchmarking datasets and challenges [14]–[20], WAMI video analytics remains to be challenging. Datasets and more importantly associated annotations are still limited for WAMI data, which adversely affects AI/ML guided approaches; particularly the data-driven deep learning approaches that require adequate data size and diversity to provide generalized solutions.

WAMI data presents unique challenges for video data fusion, object detection, and object tracking. Some of these challenges are illustrated in Figure 1 (a-c) such as small object sizes, object shape distortions, fast motion due to low frame rates, partial or full occlusions which can drastically affect object detection and tracking performance. Processing of urban scene WAMI videos is further challenging due to high densities of similar objects (i.e., parking lots full of cars, busy intersections with hundreds of vehicles and pedestrians), parallax, and occlusions due to tall buildings.

WAMI Georegistration Challenges: Georegistration and stabilization of sequential video frames are often the first steps (e.g., sub-object assessment) in WAMI video analytics, particularly in moving object detection pipelines. Video registration is often performed by estimating a frame-to-frame (piece-wise) perspective transformation (homography) which maps points of an observed dominant plane in the scene from one image's retinal plane to another. While the estimation-based methods for obtaining homography transformations may work well for general cases, it becomes very challenging for WAMI of urban scenery [3]. Figure 1 (d-f) illustrate registration errors, and Figure 1 (g-i) show high-fidelity 3D buildings viewed from different angles causing high levels of parallax motion. In case of WAMI videos, frame-to-frame homography estimation methods often fail to smoothly stabilize the whole sequence of frames and result in fragmentations [21]. That is because the conventional homography estimation methods only use the information available from 2D feature correspondences and ignore (or unable) to utilize the 3D information in underlying scenes. Thus the large parallax which exists in most of urban WAMI scenarios distract these methods and cause them to fail to register the frames over a long sequence.


(a) Seams (b) Small objects (c) Motion blur

(d) CLIF Fr#001 (e) CLIF Fr#101 (f) Misregistration

(g) ABQ Fr#0500 (h) ABQ Fr#1000 (i) GP registered

Fig. 1: Illustration of challenges in WAMI: (a) Seams in multi-camera stitching from georegistration errors. (b) Small vehicle objects with drastic appearance change due to relative viewpoint. (c) Uncorrected motion blur. (d), (e) Two frames from CLIF 2007. (f) Composite pseudocolor image with (d) in red and (e) in green channels, showing ground-plane misalignment. (g), (h) Two frames from ABQ 2013 dataset. (i) Composite image showing parallax of tall buildings.

In this paper, we discuss the challenges and opportunities in WAMI image analytics and present benchmarking efforts to improve the state-of-the-art in WAMI object detection and tracking. The subsequent parts of this paper are organized as the following. Section II describes some single and multi-camera WAMI datasets. Section III presents processing steps involved in a typical WAMI image analytics pipeline including georegistration, vehicle detection, vehicle tracking and sample algorithms associated with these steps. Section IV describes the evaluation metrics used in this paper. Experimental results are presented in Section V followed by conclusions.

II. WAMI Datasets

This section briefly describes some WAMI datasets of interest. Table I gives a short description of sample publicly available WAMI datasets of interest. This paper focuses on two of these datasets, a multi-camera Columbus Large Imagery Format (CLIF) 2007 [27] dataset, and a single-camera Albuquerque, New Mexico (ABQ) 2013 [29] dataset (further described below). Video and annotation details for these datasets are summarized in Table II.

a) Columbus Large Image Format (CLIF) 2007 Dataset: CLIF 2007 dataset [27] consists of several hours of imagery collected from a large format electro-optical (EO) platform by AFRL Sensors Directorate on October 2007 over the Ohio State University (OSU) campus. The data is collected using a matrix of six cameras at approximately 2 frames per second.

b) ABQ 2013 Dataset: Aerial urban imagery dataset collected by TransparentSky [29] using a large format camera mounted on a gimbal with on-board GPS and IMU, with a circular data collection flight path 1.5km above ground level over downtown Albuquerque, NM on September 3, 2013 [28], [30]. Imaging was done at frame rate of 4Hz and 2.6km orbit radius. This dataset contains 1071 raw high resolution images (6600×4400) with nominal ground resolution of 25cm. Ground-truth for the dataset consists of manually marked bounding boxes and track IDs for all the moving vehicles (139 vehicle tracks in total) in a 2000×2000 region of interest extracted from 200 consecutive frames.

Prior to benchmarking moving object detection and tracking algorithms, the datasets and associated annotations need to undergo a set of processing steps including coordinate transformations and georegistration as described below.

III. Algorithms for Georegistration, Vehicle Detection and Tracking in WAMI

Georegistration, object detection, and object tracking are core processing steps in a WAMI video analytics pipeline. This section briefly describes sample methods associated with these steps. As baseline for georegistration and object tracking, our group's recent works on analytical homography estimation in 3D [31], [32] and multi-cue multi-target tracking [10], [11] are described. Video object detection [33]–[39]. typically relies on two types of approaches, appearance-based and change or motion-based. Fusion of appearance, change, and motion cues have also been used for more robust performance [11], [40], [41]. In this paper, as baseline for WAMI object detection, we describe and evaluate one appearance-based approach (YOLO network [5]), and one fused multi-cue approach extending YOLO network [5] with our novel change/motion detection scheme Persistent Flux (PFlux).

Name	Sensors	Scene	Ground-truth	Targets
UNICORN 2008 [22]	Visible (6-cameras) SAR	Wright-Patterson Air Force Base	Manual+GPS Partial (4 million labels)	Moving vehicles, radar reflectors, calibration targets
WPAFB 2009 [23], [24]	Visible (6-cameras) SAR	Wright-Patterson Air Force Base	1,537 stitched images; GT: 1/3 training; 1/3 self-test	All moving vehicles for two-thirds of the frames
MAMI 2013 [25], [26]	Aerial (5-color+1-gray) Ground (4-color) cameras	Wright-Patterson Air Force Base	Manual - Partial	Variety of objects in the scene
CLIF 2007 [27]	Visible (6-cameras)	Ohio State University campus	Manual (3,502,401 labels)	All moving vehicles
ABQ 2013 [28]	Single camera	Albuquerque, NM Downtown	Manual	All moving vehicles in an ROI

TABLE I: Five WAMI dataset collections and their characteristics.

Dataset	ABQ 2013 [29]	CLIF 2007 [27]
Frame per second	4	~ 2
Raw frame size	6600×4400 pixels	4016×2672 pixels
Registered frame size	≈ 12000× 12000 pixels	31744×29696 pixels
ROI size with GT	2000×2000 pixels	Full frame
# Frames with GT	200	6,343
# Object instances	139	3,502,401

TABLE II: Video and annotation details for the ABQ 2013 [29] and CLIF 2007 [27] WAMI datasets.

A. WAMI Georegistration and Stabilization

In order to deal with WAMI georegistration challenges (described in Section I), we utilize the recent georegistration works [31], [32] that are able to analytically estimate the homography matrices in 3D space resulting in robust and global registrations. ABQ dataset have been stabilized using MU BA4S [32], [42] georegistration through a direct analytical homography from camera 3D poses (location and orientation) as described in [6]. In MU BA4S, the registration has been carried out by applying a homography transformation between each image plane and the ground dominant plane of the scene. Such homography transformations are analytically obtained using camera parameters, i.e. their rotation matrices and translation vectors (Table III). The camera 3D poses are obtained through efficient Bundle Adjustment [42]. For a 3D point $\mathbf{X}$ lying on a dominant ground plane in the scene, π, its Z component is zero, a homogeneous image point $\tilde{\mathbf{x}}$ can be computed as follows:

$$\tilde{\mathbf{x}} = \mathbf{K}(\mathbf{R}\mathbf{X} + \mathbf{t}) \tag{1}$$

$$\tilde{\mathbf{x}} = \mathbf{K}\begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix}{}^{\pi}\tilde{\mathbf{x}} \tag{2}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are respectively the first and second columns of the rotation matrix $\mathbf{R}$ between the world coordinate reference and the corresponding camera, and $\mathbf{K}$ represents a 3×3 matrix of the camera intrinsic parameters. One can consider the term $\mathbf{K}[\mathbf{r}_1 \;\; \mathbf{r}_2 \;\; \mathbf{t}]$ as a 3×3 homography transformation matrix which maps any 2D point from π onto the camera image plane as $\tilde{\mathbf{x}} = \mathbf{H}_{\pi \to c}{}^{\pi}\tilde{\mathbf{x}}$. Likewise, a homogeneous image point $\tilde{\mathbf{x}}$ can be mapped on π as ${}^{\pi}\tilde{\mathbf{x}} = \mathbf{H}_{c \to \pi}\,\tilde{\mathbf{x}}$, where $\mathbf{H}_{c \to \pi}$ is the inverse of $\mathbf{H}_{\pi \to c}$ and is equal to

$$\mathbf{H}_{c \to \pi} = \begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix}^{-1}\mathbf{K}^{-1}. \tag{3}$$

Assuming $\mathbf{T} = \begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix}$, f as the focal length in pixel, and (u, v) as the camera image principal point, (3) after simplification can be expressed as:

$$\mathbf{H}_{c \to \pi} = \frac{1}{\lambda}\begin{bmatrix} m_{11} & -m_{21} & [-m_{11} \quad m_{21} \quad m_{31}]\,\mathbf{v} \\ -m_{12} & m_{22} & [\;\; m_{12} \; - m_{22} \; - m_{32}]\,\mathbf{v} \\ r_{13} & r_{23} & -\mathbf{r}_3^{\intercal}\mathbf{v} \end{bmatrix} \tag{4}$$

where $\mathbf{v} = \begin{bmatrix}u & v & f\end{bmatrix}^{\intercal}$ and λ is a scalar defined as $\lambda = f\mathbf{r}_3^{\intercal}\mathbf{t}$, and m_{ij} is the $minor(i,j)$ of matrix $\mathbf{T}$. Note that λ in (4) can be omitted as a homography matrix is up-to-scale.

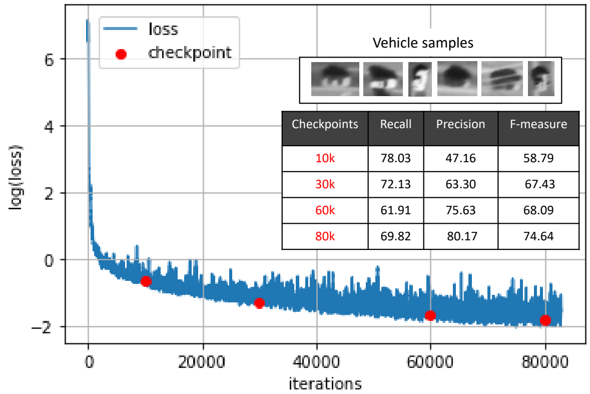
W	world coordinate system
π	dominant ground plane
$C_1, C_2 \ldots C_n$	n airborne cameras
$\mathbf{X} = [x \;\; y \;\; z]^{\intercal}$	image homogeneous coordinate of a 3D point from the world reference system W projected on the image plane of camera C
$\mathbf{K}$	calibration matrix (intrinsics)
$\mathbf{R}$	rotation matrix
$\mathbf{t}$	translation vector from W to C
${}^{\pi}\tilde{\mathbf{x}} = [x \;\; y \;\; 1]^{\intercal}$	2D homogeneous coordinates of the 3D point $\mathbf{X}$ on π [43].

TABLE III: Notation used for georegistration.

The introduced mathematical model for image stabilization works well for stabilization of parts of the image which lie on the ground dominant plane (on-the-plane). However for off-the-plane points (any non-flat objects such as buildings, cars etc.), their homographic projections introduce significant spurious motions, which can be very distractive for motion detection algorithms.

B. Appearance-based WAMI Vehicle Detection Using YOLO

Recent advances in AI/ML particularly in deep learning have revolutionized object detection. Deep learning-based object detectors can be divided into two main categories: region proposal based two-stage detectors (e.g. Faster R-CNN [44], R-CNN [45]), and single-stage detectors (e.g. YOLO [5], and SSD [46]), which do not require a separate region proposal process, making them more computationally efficient. We explored YOLO single-stage detector [5] (specifically YOLOv3 network) as one of the WAMI baseline object detectors. YOLO was used since it has: (a) significant speed advantages over two-stage detectors while maintaining high detection accuracies, and (b) better generalization capabilities allowing the network to make reasonably accurate detections on unseen images visually different from the training data. We demonstrated appearance based detection performance using two YOLO networks, one trained using the CLIF dataset [27] described in Section II and one using the Vehicle Detection in Aerial Imagery (VEDAI) datase [47]. VEDAI consists of 1200 satellite images collected during Spring 2012, over Utah,



Checkpoints	Recall	Precision	F-measure
10k	78.03	47.16	58.79
30k	72.13	63.30	67.43
60k	61.91	75.63	68.09
80k	69.82	80.17	74.64

Fig. 2: Loss for appearance training phase in YOLOv3 on CLIF 2007 . The red dots on the curve correspond to (recall, precision, and f-measure) values listed in the right sub-figure.

USA with an image resolution of 12.5cm × 12.5cm per pixel. Figure 2 shows training loss versus iterations trained using the CLIF dataset and a few image patches from the training set.

During both training and inference, the very large WAMI images are partitioned into 500×500 non-overlapping image patches and fed to the network patch by patch. This process prevents loss of image resolution caused by image resizing, a critical problem for WAMI datasets with small targets.

C. Tensor-based Motion and Change Detection

Real-world change or motion detection remains a challenging computer vision task due to confounding factors such as rapid illumination change, environmental effects, background/camera motion, shadows, camouflage effects and degraded environmental conditions (like weather, smoke, dust, chaff). For detection-based multi-object trackers, it is important to robustly detect true object motion and structural changes in the scene as opposed to changes caused by artifacts such as illumination changes [28]. Furthermore, being able to consistently detect objects that temporarily stop for a short period of time will also improve the tracking results. To address these problems, we have developed PFlux (Persistent Change with Flux Tensor) system, inspired by our earlier work [35], [48]–[50]. While these earlier results concentrated on motion dynamics, the proposed system includes new modifications and extensions to handle robust detection of longer-term, persistent changes. PFlux detection relies on visual cues generated by two tensors, 3D structure tensor and our Flux tensor [49]. The PFlux tensor is used for motion estimation and modeling, combining the 3D structure tensor with flux tensor to model structural (static) and dynamic (moving) edges.

1) 3D Color Structure Tensors: The 3D color structure tensor matrix $\mathbf{J}(\mathbf{x}, \mathbf{W})$ for a spatio-temporal photometric volume centered at $(\mathbf{x}, t)$ with smoothing and scale filter $W(\mathbf{x})$, uses first partial derivative optical-flow information of the light-field projected on the camera focal plane. The trace of the color structure tensor, $\mathbf{J}$, captures the magnitude of local static and dynamic (i.e. moving edges) orientation gradients [35], [49],

$$\textbf{trace}(\mathbf{J}(\mathbf{x}, \mathrm{W})) = \int_{\mathbf{\Omega}} ||W * \nabla \mathbf{I}||^2 d\mathbf{y} \tag{5}$$

2) Color Flux Tensors: The 3D color flux tensor, $\mathbf{J_F}(\mathbf{x}, \mathbf{W})$, uses the temporal derivatives of the 3D color structure tensor $\mathbf{J}$, to discriminate between moving and stationary salient image structures of the scene [35], [49]. The trace of the color flux tensor matrix, measuring the photometric changes in the image plane correlated with strong edge motion,

$$\textbf{trace}(\mathbf{J_F}(\mathbf{x}, \mathrm{W})) = \int_{\mathbf{\Omega}} ||\frac{\partial}{\partial t} W * \nabla \mathbf{I}||^2 d\mathbf{y} \tag{6}$$

is used to directly classify moving and non-moving regions without an eigenvalue decomposition of $\mathbf{J_F}$.

3) PFlux persistent change detection: PFlux detection module fuses information from 3D color structure and flux tensors to robustly identify dynamic regions and to differentiate moving edges/corners from static edges/corners that belong to scene structures. Persistent Flux (PFlux) detection steps:

(i) Build edge, color, and motion models for the scene using 3D color structure and color flux tensors:

$$E_S = \textbf{trace}(\mathbf{J}) - \alpha \cdot \textbf{trace}(\mathbf{J_F}) \tag{7}$$

(ii) Measure difference between scene model and incoming images:

$$\begin{aligned} D_E(x, y, t) =& w_{e1} \times |E_S(x, y, t) - E_{BG}(x, y)| \\ &- w_{e2} \times M_F(x, y) \end{aligned} \tag{8}$$

where E_S denotes the static edge features for the incoming frame; E_{BG} denotes the static edge features for the learned scene model; M_F is the motion evidence obtained from flux tensor; and w_{e1}, w_{e1} are the two weighting factors.

(iii) Identify persistent change by accumulating change in time using a sliding window of k frames.

(iv) Report regions of motion and persistent change.

PFlux is implemented in C++ with GPU support and runs at approximately 45 FPS on high definition videos (1280 × 1024 pixels), using a laptop with an 8-core Intel Core i7-7700HQ 2.80GHz CPU and an NVIDIA GeForce GTX 1060 GPU.

D. Object Tracking Algorithms

Object tracking [36], [51], [52] is one of the ultimate goals in WAMI video analytics. In this paper, as baseline WAMI multi-object tracker, we have tested and evaluated M2Track-lite a slim version of our multi-cue multi-target tracker, SCTrack, described in [10], [11]. SCTrack uses multiple visual cues (position, size, shape, color, texture) and multiple association steps to robustly link detections in time. Its main highlights include: (i) multi-cue appearance description; (ii) an efficient multi-step data association pipeline that maintains the identities of the tracked objects by resolving association ambiguities through a sequence of steps that are increasing in complexity; and (iii) explicit tracklet linking, merge/split handling, and occlusion handling modules for robust and persistent tracking. M2Track-lite uses a part of this pipeline (excluding tracklet linking and occlusion handling modules) with spatial proximity for temporal data association.

IV. Evaluation Metrics

Georegistration Metrics: Georegistration accuracy was assessed using four manually tracked points on the dominant ground plane for ABQ 2013 (4 Hz, 200 frames) and CLIF 2007 (∼2 Hz, 100 frames) to quantify pixel drift errors.

Detection Metrics: Detection accuracy was evaluated using object level recall and precision measures. Note that appearance-based object detectors (i.e. YOLO) tend to over detect due to identifying both stationary and moving objects (i.e. parked and moving vehicles).

MOT Metrics: Different evaluation metrics have been proposed for multi-object tracking [18], [19], [53]. For the evaluation of WAMI multi-object tracking results, we adopted Multi-Object Tracking (MOT) challenge evaluation metrics summarized below and described in [53], [54]. The toolkit for MOT benchmark evaluation provided in [55] was used to evaluate the WAMI tracking results. Below is a brief description of the MOT evaluation metrics used in this work:

1) **MOTA**: Multiple Object Tracking Accuracy, the most popular metric, is calculated by combining three types of errors on a per frame basis, and can be negative:

$$\mathrm{MOTA} = 1 - \frac{\sum_t (FN_t + FP_t + IDS_t)}{\sum_t GT_t} \tag{9}$$

where t is the frame number, FN_t is the number of undetected (missed) objects (false negatives), FP_t is the number of extra detected objects (false positives), and GT_t is the number of ground-truthed objects in frame t.

2) **IDS**: Identity switches, counts number of identity mismatches by considering the ID mapping in frame t and $t-1$. The *IDS* metric describes the number of times that the matched identity of a tracked trajectory changes.

3) **FRAG**: Fragmentation metric is the number of times that trajectories are fragmented. Both *IDS* and *FRAG* metrics reflect the accuracy of tracked trajectories.

4) **MT**: Mostly tracked metric computes percentage of trajectories (with respect to the number of ground-truth trajectories) tracked accurately for more than 80% of the trajectory duration.

5) **PT**: Partially tracked are cases not labeled as *MT* or *ML*.

6) **ML**: Mostly lost metric computes the percentage of trajectories tracked accurately for less than 20% of the trajectory duration. *MT* and *ML* metrics determine how much of the trajectories are recovered by the tracker.

V. Experimental Benchmarking Results

Georegistration Results: The mean and standard deviation of the translation errors ($\Delta x, \Delta y$, Euclidean distance) for each tracked point, averaged over all frames, due to shifts (drift) arising from georegistration errors are shown in Table IV. Figure 3 shows the drift with respect to the mean shift (left scatter plot) and with respect to adjacent frame pairs (right line plot). It is evident that CLIF has very large georegistration errors more than one order of magnitude higher than the subpixel errors in ABQ using our BA4S pose refinement [31], [32], [56]. Outliers have up to 50 pixel shift error, with respect to mean position, and over 25 pixel frame-to-frame shift error in CLIF due to inaccuracies in multi-camera georegistration. The former reflects georegistration accuracy, while the latter is indicative of difficulties during data association for tracking.

Dataset	Point A	Point B	Point C	Point D	Mean	StdDev
ABQ-BA4S	0.5620	0.5095	0.4463	0.6607	0.5446	0.3599
CLIF-Conv	6.3970	4.5907	6.7544	6.3670	6.0273	5.0900

TABLE IV: Mean drift error in pixels for four points tracked in each WAMI sequence for 200 and 100 frames respectively in ABQ and CLIF, after georegistration using different methods.

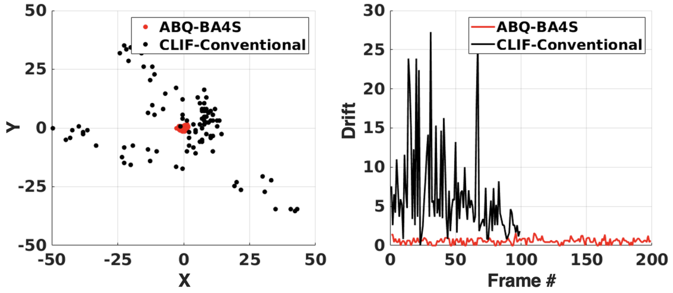


Fig. 3: Drift in pixels for one manually tracked point in each WAMI sequence across 200 (ABQ) and 100 (CLIF) frames.

Detection Results: Figure 4 shows motion detection results from Flux [49], [50] and PFlux on sample frames from two WAMI datasets ABQ 2013 single-camera WAMI, and AFRL CLIF-2007 multi-camera WAMI. For ABQ there are motion responses from both moving objects and building structures due to motion parallax. This makes it difficult to use only the estimated motion areas to aid in detecting and tracking vehicles, despite the fact that the ground plane in the video is well stabilized. By using 3D building structure cues [6] the motion responses due to building motion parallax can be accurately filtered out. Image transformation and mosaicing is another challenge in large scale WAMI datasets. Inaccurate georegistration can result in both unstable ground-plane motion and motion from building parallax which makes filtering buildings more difficult. The detection problem is further compounded by visible seams when mosaicing multiple cameras in WAMI frames, which can lead to motion detection failures, as shown in Figure 4 AFRL CLIF. While both Flux and PFlux methods result in false detections, PFlux can differentiate between changes due to mis-registration (red channel) versus parallax and true object motion (blue channel), to reduce the number of false detections. Here we use PFlux primarily to visualize the influence of georegistration errors.

MOT Results: Table V shows M2Track-Lite multi-object tracker performance on two sample WAMI datasets, single-camera ABQ 2013 [29] dataset with fairly accurate georegistration, and multi-camera AFRL CLIF 2007 [27] dataset with

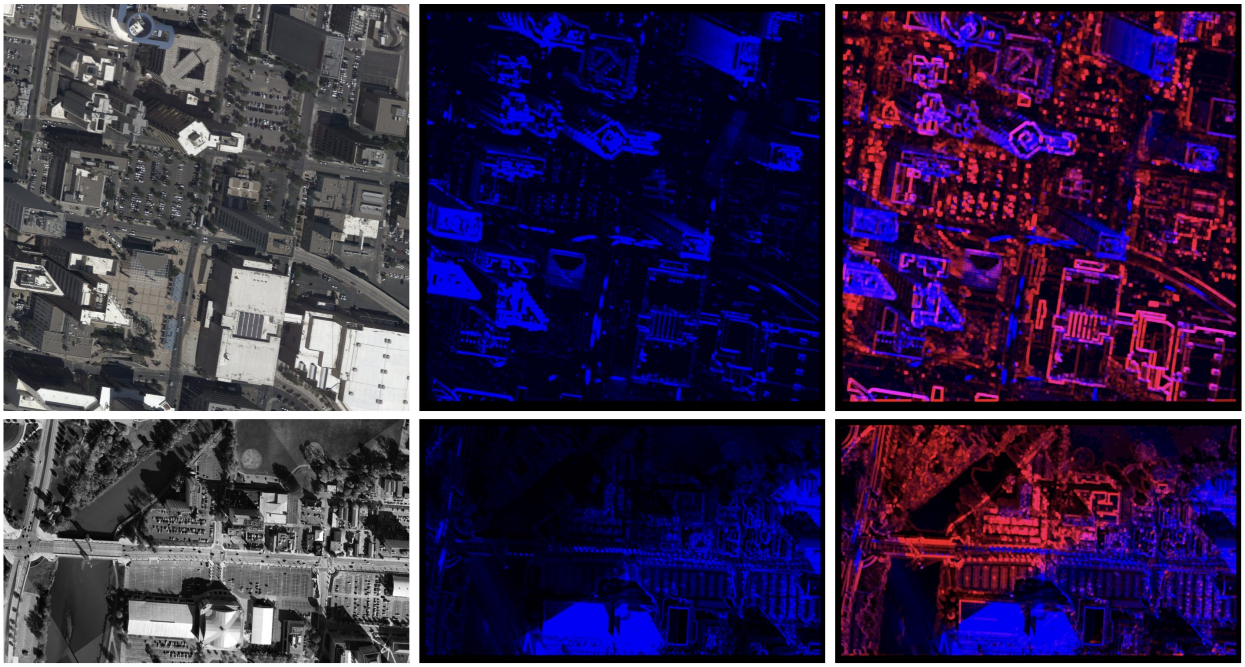


Fig. 4: Motion detection results for sample images from two WAMI datasets: Row 1 is ABQ 2013 and Row 2 is AFRL CLIF 2007 dataset. Col 1 shows the original images. Col 2 shows motion detection results using Flux (blue channel). Col 3 is Flux response (blue channel) and PFlux (red channel) showing persistent change from motion, and jitter or georegistration errors.

larger georegistration errors (Table IV, Figure 3) and visible inter-camera seams. Detection and tracking performance evaluations are also presented for a $5,000 \times 3,000$ region of interest (ROI) from the AFRL CLIF 2007 dataset positioned at $(x = 8750, y = 11220)$. This ROI was selected because the region includes multiple busy intersections and persistently stays in the field of view of the cameras resulting in longer ground-truth trajectories. Tracking results are presented for three types of detections corresponding to manually generated ground-truth (GT) detections, appearance based deep network detections (YOLO [57]), and multi-cue detections based on decision fusion of appearance-based and motion/change based detections described in III-C (YOLO+Flux [6]).

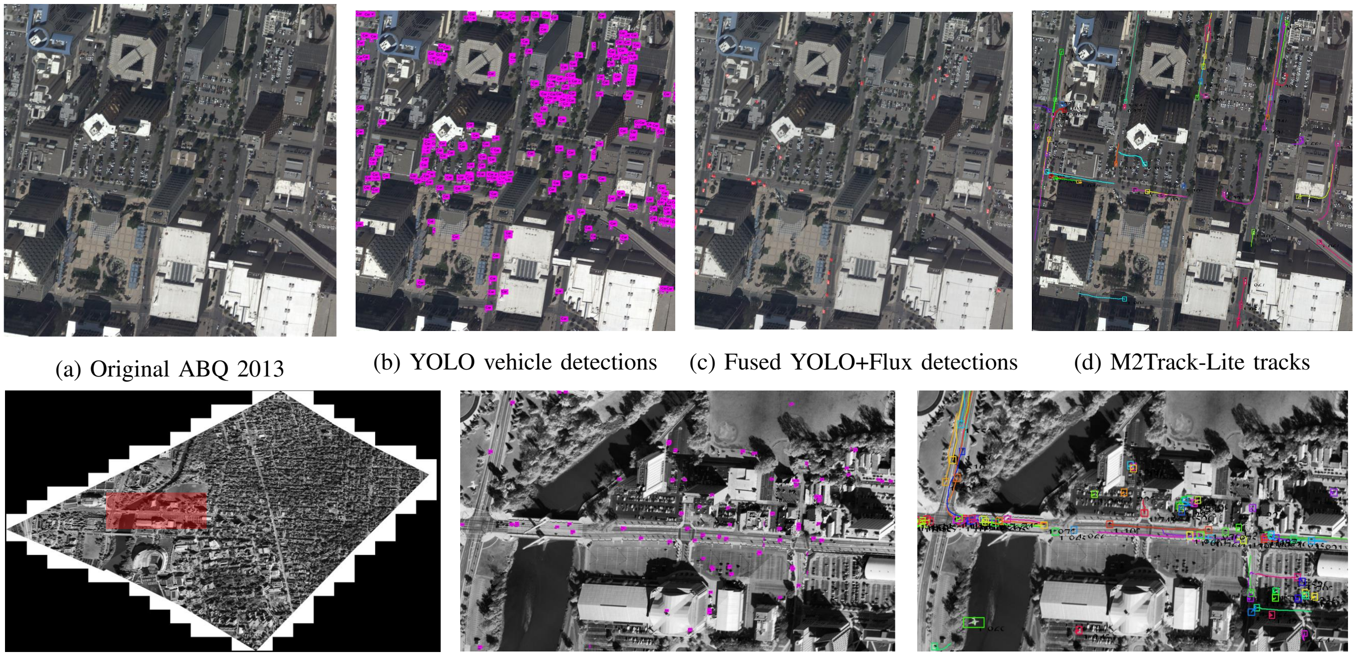
WAMI multi-object tracking performance heavily depends on object detection and georegistration accuracy. Figure 5 illustrates sample detection results (YOLO and YOLO+Flux) and multi-object tracking results using ground-truth detections. Fusion of multiple cues provides better detections leading to improved tracking; using YOLO only versus YOLO+Flux detections MOTA improves from -310 to 73 on ABQ (see Table V). The impact of georegistration accuracy on multi-object detection and tracking was evaluated using CLIF 2007. When large georegistration errors are present, average $\sim$6 pixels in CLIF, accurate estimation of motion and change cues is adversely impacted, decreasing MOTA score from 71.9 for YOLO only, to 68.3 for YOLO+Flux in the cropped CLIF 2007 and from 40.3 to 30.3 for the Full CLIF 2007 dataset (see Table V). These results demonstrate that early upstream processing stages of WAMI video analytics pipelines, especially multi-camera georegistration accuracy, are crucial for accurate performance in small target detection and tracking.

VI. Conclusions

Wide area motion imagery (WAMI) enables mapping and monitoring of large geospatial areas for extended periods of time driving the expanded use of single and multi-platform aerial imagery. While we have seen remarkable advances in video analytics for ground or low-altitude aerial videos, automated or semi-automated analysis of WAMI videos have been limited due to data complexities, lack of labeled training videos, and large data processing requirements. In this paper, we presented an evaluation of how georegistration accuracy effects vehicle detection and multi-object tracking using two WAMI datasets (CLIF 2007 and ABQ 2013) providing a baseline set of results for developing improved techniques to detect and track small objects accurately in aerial imagery. The sensitivity of object recognition and tracking algorithms to georegistration accuracy motivates the further development of more robust and faster georegistration algorithms for multi-platform multi-camera systems that can dynamically adapt to environmental and optical changes. This will facilitate advancing the state-of-the-art in WAMI video analytics, particularly multi-object detection and tracking of small objects.

Acknowledgments

This work was partially supported by awards from U.S. Army Research Laboratory W911NF-1820285 and Army Research Office DURIP W911NF-1910181. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U. S. Government or agency thereof.



(a) Original ABQ 2013 (b) YOLO vehicle detections (c) Fused YOLO+Flux detections (d) M2Track-Lite tracks

(e) Original AFRL CLIF 2007 with ROI (f) YOLO vehicle detections in cropped region (g) M2Track-Lite tracks in cropped region

Fig. 5: Sample detection and tracking results for ABQ 2013 [29] and AFRL CLIF 2007 [27] WAMI showing improvement in detection accuracy when appearance (YOLO) and motion (Flux) are combined. M2Track-Lite tracks using GT detections.

Datasets	**Georegistration**	**Detector Alg**	**M2Track Lite Tracker**							
Metrics			**MOTA↑**	**IDSW↓**	**FRAG↓**	**MT↑**	**PT↑**	**ML↓**	**GT-ID**	**#Detections**
ABQ	BA4S	GT	99.70	24	0	139	0	0	139	8,323
ABQ	BA4S	YOLO [57]	-310.0	53	0	68	11	43	139	35,027
ABQ	BA4S	YOLO+Flux [6]	73.80	248	8	139	0	0	139	10,066
Full CLIF	Conventional	GT	90.42	3,559	676	885	1	0	886	12,413
Full CLIF	Conventional	YOLO [57]	40.30	184	46	571	0	0	571	17,574
Full CLIF	Conventional	YOLO+Flux [6]	30.30	1,058	45	551	15	5	571	16,338
Cropped CLIF	Conventional	GT	99.30	52	19	129	0	0	129	7,544
Cropped CLIF	Conventional	YOLO [57]	71.90	1054	281	100	25	4	129	6,550
Cropped CLIF	Conventional	YOLO+Flux [6]	68.30	622	264	73	48	8	129	5,778

TABLE V: Tracking performance using ground-truth vehicle detections, YOLO and YOLO+Flux object detection in WAMI datasets ABQ 2013, AFRL CLIF 2007, and cropped $5,000 \times 3,000$ pixel region of interest (ROI) from AFRL CLIF 2007 (see Figure 5(e)). M2Track-Lite multi-object tracker was used without tracklet linking. Evaluation metrics include Multi-object Tracking Accuracy (MOTA), ID switches, Fragmentation, Mostly Tracked, Partially Tracked, Mostly Lost, and GT-ID (Number of Tracks) as described in HOTA [54]. ABQ 4 Hz, 200 frames; CLIF ∼2 Hz, 100 frames.

References

[1] H. Ling, Y. Wu, E. Blasch, *et al.*, "Evaluation of visual tracking in extremely low frame rate wide area motion imagery," in *Int. Conf. on Information Fusion*, 2011.

[2] E. Blasch, G. Seetharaman, S. Suddarth, K. Palaniappan, *et al.*, "Summary of methods in wide-area motion imagery (WAMI)," in *Proc. SPIE 9089*, 2014.

[3] K. Palaniappan, R. Rao, and G. Seetharaman, "Wide-area persistent airborne video: Architecture and challenges," in *Distributed Video Sensor Networks: Research Challenges and Future Directions* (B. Banhu *et al.*, eds.), ch. 24, pp. 349–371, Springer, 2011.

[4] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. Computer vision and Pattern Recognition*, pp. 779–788, 2016.

[6] N. Al-Shakarji, F. Bunyak, H. Aliakbarpour, G. Seetharaman, and K. Palaniappan, "Multi-cue vehicle detection for semantic video compression in georegistered aerial videos," in *IEEE Conf. Comp. Vision Pattern Recognition Workshops*, pp. 56–65, 2019.

[7] R. Pelapur, S. Candemir, F. Bunyak, M. Poostchi, G. Seetharaman, and K. Palaniappan, "Persistent target tracking using likelihood fusion in wide-area and full motion video sequences," *Int. Conf. on Information Fusion*, pp. 2420–2427, 2012.

[8] Z. Ma, L. Wang, H. Zhang, W. Lu, and J. Yin, "RPT: learning point set representation for siamese visual tracking," in *European Conf. on Computer Vision*, pp. 653–665, 2020.

[9] N. M. Al-Shakarji, F. Bunyak, G. Seetharamany, and K. Palaniappan, "CS-LOFT: Color and scale adaptive tracking using max-pooling with bhattacharyya distance," in *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–7, 2016.

[10] N. Al-Shakarji, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Robust multi-object tracking with semantic color correlation," in *IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, pp. 1–7, 2017.

[11] N. Al-Shakarji, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Multi-object tracking cascade with multi-step data association and occlusion handling," in *IEEE Conf. Advanced Video and Signal Based Surveil-*

lance, pp. 1–6, 2018.

[12] Y. Chen, E. Blasch, N. Chen, A. Deng, H. Ling, and G. Chen, "Real-time WAMI streaming target tracking in fog," in *Proc. SPIE 9838*, 2016.

[13] R. Wu, B. Liu, Y. Chen, E. Blasch, H. Ling, and G. Chen, "A container-based elastic cloud architecture for pseudo real-time exploitation of wide area motion imagery (WAMI) stream," *J. of Sig. Proc. Sys., 88 (2): 219-231*, 2017.

[14] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," *IEEE Int. Workshop Perf. Eval. Tracking & Surveillance*, pp. 1–6, 2009.

[15] E. Blasch, G. Seetharaman, K. Palaniappan, H. Ling, and G. Chen, "Wide-area motion imagery (WAMI) exploitation tools for enhanced situation awareness," in *IEEE Applied Imagery Pattern Recog. Workshop*, 2012.

[16] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *IEEE Computer Vision and Pattern Recognition Workshops*, 2012.

[17] P. Zhu *et al.*, "VisDrone-VDT2018: The vision meets drone video detection and tracking challenge results," in *European Conf. on Computer Vision (ECCV)*, vol. LNCS 11133, pp. 496—-518, 2019.

[18] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831v2*, 2016.

[19] L. Wen, D. Du, Z. Cai, *et al.*, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Computer Vision and Image Understanding*, vol. 193, p. 102907, 2020.

[20] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, M. Danelljan, L. Č. Zajc, A. Lukežič, O. Drbohlav, *et al.*, "The eighth visual object tracking VOT2020 challenge results," in *European Conf. on Computer Vision*, pp. 547–601, 2020.

[21] M. Linger and A. Goshtasby, "Aerial image registration for tracking," *IEEE Trans. Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2137–2145, 2015.

[22] C. Leong, T. Rovito, O. Mendoza-Schrock1, C. Menart, J. Bowser, L. Moore1, S. Scarborough, M. Minardi, and D. Hascher, "Unified Coincident Optical and Radar for Recognition (UNICORN) 2008 Dataset," 2019. https://github.com/AFRL-RY/data-unicorn-2008.

[23] C. Cohenour, F. van Graas, R. Price, and T. Rovito, "Camera models for the wright patterson air force base (WPAFB) 2009 wide-area motion imagery (WAMI) data set," *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, no. 6, pp. 4–15, 2015.

[24] C. Cohenour, R. Price, T. Rovito, and F. van Graas, "Corrected pose data for the wright patterson air force base 2009 wide area motion imagery data set," *J. Applied Remote Sensing*, vol. 9, no. 1, p. 096048, 2015. https://www.sdms.afrl.af.mil/index.php?collection=wpafb2009.

[25] R. Ilin and S. Clouse, "Extraction and classification of moving targets in multi-sensory MAMI-1 data collection," in *National Aerospace and Electronics Conference (NAECON)*, pp. 387–391, 2015.

[26] https://www.sdms.afrl.af.mil/index.php?collection=mami2013.

[27] T. Rovito, J. Patrick, S. Walls, D. Uppenkamp, O. Mendoza-Schrock, V. Velten, C. Curtis, and K. Priddy, "Columbus Large Image Format (CLIF) 2007 Dataset." https://github.com/AFRL-RY/data-clif-2007.

[28] K. Palaniappan, M. Poostchi, H. Aliakbarpour, R. Viguier, J. Fraser, F. Bunyak, A. Basharat, S. Suddarth, E. Blasch, R. Rao, and G. Seetharaman, "Moving object detection for vehicle tracking in wide area motion imagery using 4D filtering," in *IEEE Int. Conf. on Pattern Recognition (ICPR)*, pp. 2830–2835, 2016.

[29] ABQ video. http://www.transparentsky.net.

[30] M. Poostchi, H. Aliakbarpour, R. Viguier, F. Bunyak, K. Palaniappan, and G. Seetharaman, "Semantic depth map fusion for moving vehicle detection in aerial video," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 32–40, 2016.

[31] H. Aliakbarpour, K. Palaniappan, and G. Seetharaman, "Stabilization of airborne video using sensor exterior orientation with analytical homography modeling," in *Machine Vision and Navigation*, pp. 579–595, Springer, 2020.

[32] H. Aliakbarpour, K. Palaniappan, and G. Seetharaman, "Robust camera pose refinement and rapid SfM for multiview aerial imagery—without RANSAC," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2203–2207, 2015.

[33] T. Gautama and M. Van Hulle., "A phase-based approach to the estimation of the optical flow field using spatial filtering," *IEEE Trans. on Neural Networks*, vol. 13, no. 5, pp. 1127–1136, 2002.

[34] M. Farmer, X. Lu, H. Chen, and A. Jain, "Robust motion-based image segmentation using fusion," *IEEE Int. Conf. on Image Processing*, vol. 5, pp. 3375–3378, 2004.

[35] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split Gaussian models," *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 414–418, 2014.

[36] 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," *IEEE Workshop on Applications of Computer Vision*, pp. 839–846, 2014.

[37] R. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Trans. on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 525–534, 2016.

[38] L. W. Sommer, M. Teutsch, T. Schuchert, and J. Beyerer, "A survey on moving object detection for wide area motion imagery," in *IEEE Winter Conf. on Applications of Computer Vision*, pp. 1–9, 2016.

[39] M. Shafiee, B. Chywl, F. Li, and A. Wong, "Fast YOLO: A fast you only look once system for real-time embedded object detection in video," *arXiv:1709.05943*, 2017.

[40] M. Siam, H. Mahgoub, M. Zahran, S. Yogamani, M. Jagersand, and A. El-Sallab, "MODNET: Moving object detection network with motion and appearance for autonomous driving," *Int. Conf. Intelligent Transportation Systems*, 2017.

[41] B. Heo, K. Yun, and J. Choi, "Appearance and motion based deep learning architecture for moving object detection in moving camera," in *IEEE Int. Conf. on Image Processing*, pp. 1827–1831, 2017.

[42] H. AliAkbarpour, K. Palaniappan, and G. Seetharaman, "Fast structure from motion for sequential and wide area motion imagery," in *IEEE Int. Conf. on Computer Vision Workshops*, pp. 34–41, 2015.

[43] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, ISBN: 0521540518, second ed., 2004.

[44] R. Girshick, "Fast R-CNN," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[45] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[46] W. Liu *et al.*, "SSD: Single shot multibox detector," in *European Conf. Computer Vision*, vol. LNCS 9905, pp. 21–37, 2016.

[47] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 187–203, 2016.

[48] F. Bunyak, I. Ersoy, and S. Subramanya, "Shadow detection by combined photometric invariants for improved foreground segmentation," in *IEEE Workshops on Application of Computer Vision*, vol. 1, pp. 510–515, 2005.

[49] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman, "Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking," *J. Multimedia*, vol. 2, no. 4, 2007.

[50] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman, "Geodesic active contour based fusion of visible and infrared video for persistent object tracking," in *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 35–35, 2007.

[51] C. Aeschliman, J. Park, and A. C. Kak, "Tracking vehicles through shadows and occlusions in wide-area aerial video," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 50, no. 1, pp. 429–444, 2014.

[52] X. Zhao, D. Dawson, W. A. Sarasua, and S. T. Birchfield, "Automated traffic surveillance system with aerial camera arrays imagery: Macroscopic data collection with vehicle tracking," *Journal of Computing in Civil Engineering*, vol. 31, no. 3, p. 04016072, 2017.

[53] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," in *Int. Evaluation Workshop on Classification of Events, Activities and Relationships*, pp. 1–44, 2006.

[54] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "HOTA: A higher order metric for evaluating multi-object tracking," *Int. J. Comp. Vision*, vol. 129, no. 2, pp. 548–578, 2021.

[55] MOT tool kit. https://motchallenge.net/devkit/.

[56] H. AliAkbarpour, K. Palaniappan, and G. Seetharaman, "Parallax-tolerant aerial image georegistration and efficient camera pose refinement—without piecewise homographies," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4618–4637, 2017.

[57] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.