HAL open science

Enhanced GM-PHD filter for real time satellite multi-target tracking

Camilo Aguilar, Mathias Ortner, Josiane Zerubia

To cite this version:

Camilo Aguilar, Mathias Ortner, Josiane Zerubia. Enhanced GM-PHD filter for real time satellite multi-target tracking. ICASSP 2023 - IEEE International Conference on Acoustics, Speech, and Signal Processing, Jun 2023, Rhodes, Greece. hal-04029072

HAL Id: hal-04029072

https://inria.hal.science/hal-04029072v1

Submitted on 14 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ENHANCED GM-PHD FILTER FOR REAL TIME SATELLITE MULTI-TARGET TRACKING

Camilo Aguilar[⋆] *Mathias Ortner*[†] *Josiane Zerubia*[⋆]

[⋆] Inria, Université Côte d'Azur, Sophia Antipolis, France
[†] Airbus Defense and Space, Toulouse, France

ABSTRACT

We present a real-time multi-object tracker using an enhanced version of the Gaussian mixture probability hypothesis density (GM-PHD) filter to track detections of a state-of-the-art convolutional neural network (CNN). This approach adapts the GM-PHD filter to a real-world scenario to recover target trajectories in remote sensing videos. Our GM-PHD filter uses a measurement-driven birth, considers past tracked objects, and uses CNN information to propose better hypotheses initialization. Additionally, we present a label tracking solution for the GM-PHD filter to improve identity propagation given target path uncertainties. Our results show competitive scores against other trackers while obtaining real-time performance. Code is available at https://github.com/Ayana-Inria/RFS-filters-for-satellite-videos.

***Index Terms*—** CNN, GM-PHD, MTT, Remote Sensing

1. INTRODUCTION

Multi-target tracking (MTT) estimates the number of objects and their states from observations. Sample applications include the areas of security surveillance, autonomous driving, and remote sensing videos. For example, the Chinese Jilin-1 satellite constellation aims to improve urban planning [1] using ground videos spanning 15 kms^2 with a 1m/pixel resolution. Similarly, the French company Airbus Defense and Space aims to improve border security [2] with the Geostationary telescope GO-3S that captures 10000 kms^2 with 3ms/pixel resolution at 5 Frames Per Second (FPS).

Despite significant advancements in computer vision, detecting and tracking objects in large ground areas presents several challenges. Targets span between 5 to 15 pixels in width and often lack discriminative features. Furthermore, objects move at varying directions and speeds; a target in a high-speed road reduces its velocity drastically when approaching intersections. Additionally, satellite videos present sources of artifacts such as clutter detections and misdetections. For example trees, roofs, containers, or small structures could become false positive detections.

Thanks to BPI France for providing financial support (LiChiE contract) and to the OPAL infrastructure from Université Côte d'Azur for providing computing resources and support.

Traditional object detectors and trackers rely on a combination of features such as appearance, object motion, or background subtraction. For example, [3, 4] uses the 3-frame difference algorithm paired with morphological filters to extract objects, and an ad-hoc track-by-detection approach to recover tracks. The works [5, 6] improve the small object detection by using a combination of the 3-frame difference with a window-based CNN. Sequentially, they recover tracks using the GM-PHD filter. These methods show promising results; however, the 3-frame difference step is prone to parallax artifacts caused by the intrinsic imaging system motion.

Recent works address object detection with deep learning-based techniques. Anchor-free methods [7, 8, 9, 10] obtain remarkable scores on the WPAFB benchmark [11]. However, when extended to object tracking, purely deep learning methods perform frame-to-frame-based tracking and fail to recover trajectory uncertainties. Additionally, they rely on ad-hoc data association methods such as the Hungarian Algorithm [12] or Simple Online Real-time Tracking (SORT) [13] method.

The Random Finite Set (RFS) [14] framework proposes a Bayesian tool to perform MTT while modeling target uncertainties. Among the RFSs filters, the Generalized Labelled Multi Bernoulli (GLMB) [15] filter obtains state-of-the-art results at the cost of large computational demands. The simplified yet robust GM-PHD [16] filter performs multi-object tracking while preserving real-time performance. However, the original GM-PHD filter fails to recover object unique track identities. Likewise, both GLMB and GM-PHD filters rely on prior knowledge of hypotheses' "birth" locations and the birth hypotheses' initial state.

In this paper we employ a state-of-the art CNN object detector coupled with an improved version of the GM-PHD filter. Particularly, we propose three improvements to original GM-PHD filter in order to better track noisy CNN-based detections. First, we improve the traditional label tracking for the GM-PHD filter. Second, we adapt hypothesis birth distribution by using a measurement-driven and a history-based birth. Finally, we improve the birth initialization using deep learning-based information. We show that our approach obtains better tracking scores than competing tracking or joint tracking-detection methods for satellite images while preserving real-time performance.

The rest of this paper is organized as follows: Section 2 presents the proposed method, first introducing the object detector and then the upgraded GM-PHD filter. Section 3 shows the setup and results for our experiments and Section 4 denotes the conclusion to our work.

2. PROPOSED METHOD

We utilize the RFS framework and model target states and measurements as sets during each image video frame k. We define a single target state as $x_i^k = [p_i^k, v_i^k]$, where $p_i^k \in \mathbb{R}^2$ denotes the object position, and $v_i^k \in \mathbb{R}^2$ denotes the target velocity. We define the collection of targets as the set $\mathbf{X}_k = \{x_1^k, x_2^k \cdots, x_{n_k}^k\}$, where n_k is a random variable and denotes the number of estimated targets. Similarly, we model the jth measurement as $z_j^k = [\bar{p}_j^k]$, where $\bar{p}_i^k \in \mathbb{R}^2$ denotes the measurement position obtained by a CNN object detector [7]. We define the collection of measurements as $\mathbf{Z}_k = \{z_1^k, z_2^k \cdots z_{m_k}^k\}$, where m_k is the number of measurements.

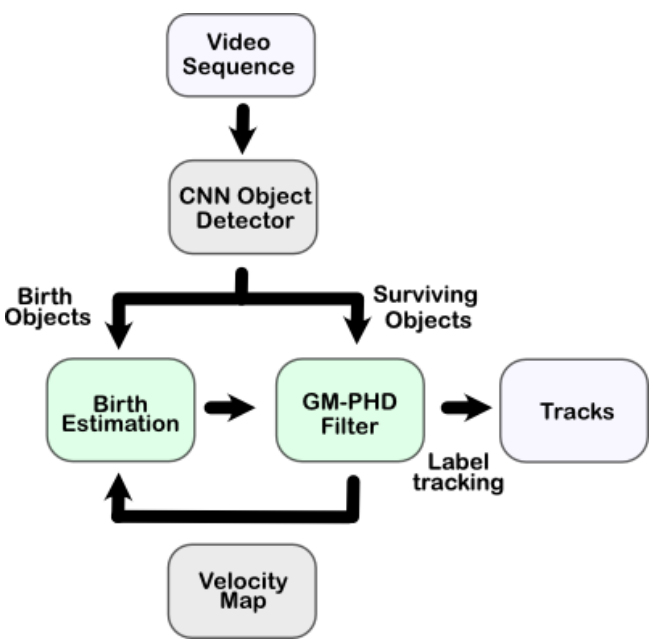


Fig. 1. Block diagram of proposed approach.

2.1. Object Detection

We employ the state-of-the-art CenterTrack (CT) [7] detector. This approach regresses a heatmap of object locations $H_k \in \mathbb{R}^{\frac{w}{s} \times \frac{h}{s} \times 1}$ and a vector map for oriented velocities $V_k \in \mathbb{R}^{\frac{w}{s} \times \frac{h}{s} \times 2}$, where s is a subsampling factor, h and w are the image width and height respectively. We feed 3 images to the network:

$$H_k, V_k = \text{CT}(I_k,\ I_{k-1},\ G_{k-1}) \tag{1}$$

where $I_k \in \mathbb{R}^{w \times h \times 3}$ denotes the current frame, $I_{k-1} \in \mathbb{R}^{w \times h \times 3}$ denotes the previous frame, and $G_{k-1} \in \mathbb{R}^{w \times h \times 1}$ denotes a Gaussian generated heatmap denoting detections at frame $k-1$. CenterTrack obtains object measurement locations $\bar{P}_k = \{\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_{mk}\}$ by using peak extraction in H_k and obtains object velocity approximations $\{\bar{v}_1, \bar{v}_2, \cdots, \bar{v}_{mk}\}$ by evaluating V_k at each point $\bar{p}_j \in \bar{P}_k$, namely $\bar{v}_j = V_k(\bar{p}_j)$.

2.2. The Trajectory GM-PHD Filter

The PHD filter estimates the multi-target posterior RFS $p_{k|k}(\mathbf{X}_k|\mathbf{Z}_{1:k})$ by approximating its first order momentum, namely the PHD function: $D_{k|k}(\mathbf{X}_k|\mathbf{Z}_{1:k})$, often simplified to $D_k(x)$. The GM-PHD approximates the PHD function with a Gaussian mixture of the form:

$$D_k(x) = \sum_{j=1}^{J_k} w_k^j \mathcal{N}(x; \mathbf{m}_k^j, \mathbf{P}_k^j) \tag{2}$$

where ω_k^j, $\mathbf{m}_k^j$, and $\mathbf{P}_k^j$ are the weight, mean, and covariance for the jth component and J_k is the number of hypothesis in the mixture. Additionally, we add label tracking to the GM-PHD following the work of [17]. We assign a hidden label ℓ_i to each Gaussian mixture component from a label identity set $I_k = \{\ell_1^k, \ell_2^k, \ldots \ell_{J_k}^k\}$, where each $\ell_i \in \mathbf{N}$ is a unique tag assigned to each component. These labels are propagated during the filter's prediction and update step.

The prediction step $D_{k|k-1}(x)$ advances states for each component and accounts for the birth hypotheses. The prediction step has the form:

$$D_{k|k-1}(x) = p_s D_{k|k-1}^s(x) + \lambda(x) \tag{3}$$

where p_s denotes the probability of survival, $D_{k|k-1}^s(x)$ accounts for the predicted surviving hypotheses, and $\lambda(x)$ models the birth components. During the prediction state, the identity set is updated to $I_{k|k-1} = I_k \cup I_k^\lambda$, where $I_k^\lambda = \{\ell_{\lambda,1}^k, \ell_{\lambda,2}^k, \cdots \ell_{\lambda,n_{Bk}}^k\}$ denotes the birth label set and n_{Bk} denotes the number of birth components in frame k. We develop the estimation of $\lambda(x)$ in Section 2.4.

The update step $D_k(x, z)$ associates the filter predicted hypotheses with measurements and has the form of:

$$D_k(x,z) = (1 - p_D)\, D_{k|k-1}(x) + \sum_{\mathbf{z} \in \mathbf{Z}_k} \sum_{j=1}^{J_{k|k-1}} \omega_k^j(z)\, \mathcal{N}(x;\, \mathbf{m}_{k|k}^j(z), \mathbf{P}_{k|k}^j(z))$$

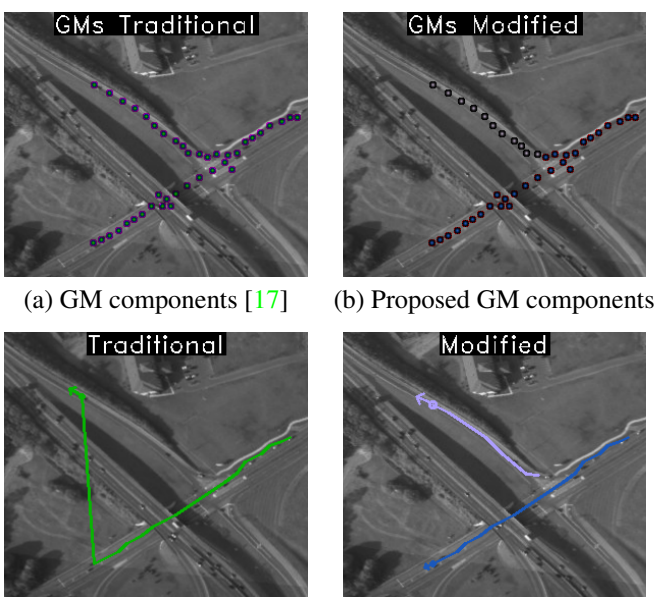
where p_D denotes the probability of detection, $D_{k|k-1}(x)$ the predicted components , and $\omega_k^j(z), \mathbf{m}_{k|k}^j(z), \mathbf{P}_{k|k}^j(z)$ the updated weight, mean and covariance respectively. During the update step, each component propagates its label $\ell_i^k \in I_{k|k-1}$ to $(1 + |Z_k|)$ updated hypotheses. In practice, the number of updated components $J_{k|k-1}$ can be significantly large, but

pruning, merging, and gating methods are employed to reduce the number of hypothesis-measurement associations with low likelihood.

2.3. Enhanced GM-PHD Label Tracking

The hidden label technique [17] propagates object identities in the GM-PHD filter; however, this method struggles to separate tracks with faulty initialization. For example, Fig. 2(a) shows a labeled Gaussian component (denoted green) tracking two nearby vehicles. When the objects separate, the filter propagates the same label to both tracks as possible hypotheses. In fact, during inference time, the filter loses trajectories or recovers discontinuous tracks as shown in Fig. 2(c).

We improve trajectory recovery based on the previoctaus inferred trajectory states. If two updated hypotheses with high likelihood contain the same label, we keep the hypothesis with the shortest distance to the previous labeled target $(\hat{x}_{k-1}, \ell_i)$. We modify our label space to $I_k^+ = I_k \cup I_k^\alpha$ and assign a new label $\ell_\alpha \in I_k^\alpha$ to the rest of the hypotheses, where I_k^α is a latent unique label space. This fix is depicted in Fig.2(b) where a new label is assigned to the Gaussian component moving in a different direction. Fig. 2(d) shows the inferred tracks after solving this issue.



(a) GM components [17] (b) Proposed GM components

(c) Inferred trajectories [17] (d) Proposed inferred trajectories

Fig. 2. Improved GM-PHD Label Propagation.

2.4. Adaptive Birth Estimation

We build on the works presented in [6, 17] to approximate the GM-PHD birth distribution $\lambda(x)$. We classify measurements into birth and surviving measurements $Z_k = Z_k^b \cup Z_k^s$ by matching incoming measurements with predicted GM components; we classify birth measurements as:

$$Z_k^b = \{z \in Z_k : d(\mathbf{F}\, x_i, z) > T_B\} \;\forall\; x_i \in D_{k-1}(x) \quad (4)$$

where $\mathbf{F}$ is the state transition matrix, T_B is a pre-defined threshold, and $d(x_i, z)$ measures the spatial distance between the measurement $z \in Z_k$ and GM component $x_i \in D_{k-1}$. Additionally, we adjust the adaptive-birth procedure developed for the GLMB filter in [18] to improve the GM-PHD hypothesis initialization. We use the position and velocity from inferred objects $\hat{x}_i^k = [\hat{p}_i, \hat{v}_i]$ to build a velocity map $h_v^k(\hat{p}_i)$ as described in [18]. The velocity map contributes to initializing future hypotheses' initial kinematic state and was shown to significantly improve high-speed target tracking [18]; nevertheless, it requires several frames and inferred objects to adapt.

Consequentially, we use the information given by the CNN to improve the initialization when the velocity history map has not been adapted yet. Hence, we define the birth distribution as:

$$\lambda(x, Z_k^b) = \frac{1}{|Z_k^b|} \sum_{z_j \in Z_k^b} \mathcal{N}(x; \mathbf{x}_b^k(z_j), P_b) \quad (5)$$

where $z_j = [\bar{p}_j, \bar{v}_j]$ denotes the measurement obtained from the CNN, P_b is the birth covariance, and $\mathbf{x}_b^k$ is the measurement-driven initial state given by:

$$\mathbf{x}_b^k(z_j) = \begin{cases} [\bar{p}_j, \bar{v}_j]^T & \text{if} \quad ||h_v^k(\bar{p}_j)|| = 0 \\ [\bar{p}_j, h_v^k(\bar{p}_j)]^T & \text{if} \quad ||h_v^k(\bar{p}_j)|| \neq 0 \end{cases} \quad (6)$$

3. EXPERIMENTS

3.1. Experimental Setup

We test our method in the WPAFB benchmark. This datasat contains videos of the Wright-Patterson Air Force Base, in OHIO, Unites States of America. We use the videos imaged with Unmanned Aerial Vehicles (UAVs) at a resolution of 1 m/pixel at a rate of 2 FPS. We perform image registration in the same manner as [6] in order to stabilize the images, and we focus our method only on moving objects. We filter the ground truth annotations to keep only objects that moved more than 5 pixels during the last 15 frames.

We train the network using Areas of Interest (AoIs) 40, 41, and 42 as mentioned in [19], and we test on AoI 2, and AoI 34. We train the network using the same loss and parameters as dictated by CenterTrack [7].

3.2. Metrics

During each frame, we perform a one to one matching between the set of ground truth annotations $P_{gt}^k = \{p_1^k, \ldots, p_N^k\}$ and the set of estimated target locations $\hat{X}_k = \{\hat{x}_1^k, \cdots, \hat{x}_{n_k}\}$.

Area	**Method**	GT	MT (%)↑	PT (%)↑	ML (%)↓	FP ↓	FN ↓	IDS↓	MOTA↑	MOTP↓	FPS ↑
02	**Proposed GM-PHD**	552	82.78	15.03	2.17	1501	2074	1624	**72.70**	2.66	3.44
	Labelled-GMPHD [5]	552	63.58	31.70	4.71	1012	3805	1902	64.80	2.64	4.35
	GLMB [15]	552	54.89	**38.20**	6.88	**751**	4384	**1548**	64.90	2.67	0.43
	SORT [13]	552	72.82	25.00	2.17	1072	2698	5069	53.60	3.39	7.16
	Greedy Match [7]	552	**90.03**	8.69	**1.26**	1203	**1459**	12087	22.60	**2.53**	**7.25**
34	**Proposed GM-PHD**	225	51.55	39.11	9.33	681	1010	**306**	**63.90**	2.57	2.13
	Labelled-GMPHD [5]	225	48.00	44.44	7.55	661	1083	625	57.20	2.61	2.18
	GLMB [15]	225	35.11	**54.66**	10.22	**550**	1503	805	48.46	2.69	0.23
	SORT [13]	225	48.88	28.00	23.11	659	993	1324	46.23	3.28	7.16
	Greedy Match [7]	225	**57.45**	37.28	**5.26**	938	**745**	3510	7.60	**2.44**	**3.56**

Table 1. Tracking scores. The arrow's direction represents better scores.

We call an estimated target $\hat{x}_i^k$ True Positive (TP) if it is located within $c = 10$ pixels from an unmatched ground truth annotation p_j^k, otherwise, we count the estimated target as a False Positive (FP). We count all the unmatched ground truth annotations as False Negative (FN) if they are unmatched with a target. Finally, we call a target Identity Switch (IDS) if it is matched with more than one ground truth trajectory.

We use the ClearMOT [20] and the MOTChallenge [21] metrics to evaluate the object detection and tracking. The ClearMOT metrics account for Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP). Additionally, the MOTChallenge metrics account for Mostly Tracked (MT) , Partially Tracked (PT), and Mostly Lost (ML) trajectories.

3.3. Results

Table 1 shows results for object trackers using CenterTrack as object detector. The trackers SORT and Greedy Matching obtain high FPS and high MT scores due to their simplified but pragmatic approach; however, both filters fail to propagate track identities in time. This is reflected in the high quantity of IDS for both methods, and hence, their low MOTA scores.

On the other hand, the RFS-based filters such as the GM-PHD and the GLMB perform well in the MOTA and MOTP metrics but their performance drops significantly in the percentage of trajectories recovered: they have lower MT scores than SORT and Greedy Matching as they present a large number of FNs. Additionally, the GLMB filter performs well in the MOTA score and shows low IDSs due to its complex probabilistic approach, but it has large computational demands as shown by its FPS.

Our enhanced GM-PHD obtains competitive scores across all metrics and gets the highest MOTA score while preserving an reasonable FPS. The improved birth estimation reduces false negatives as a correct initialization is essential to track high speed moving targets. Additionally, the improved label tracking reduces the number of identity switches by up to 40% when compared to the traditional labeled GM-PHD filter. Fig. 3 shows a sample output of our object detection and GM-PHD filter method.

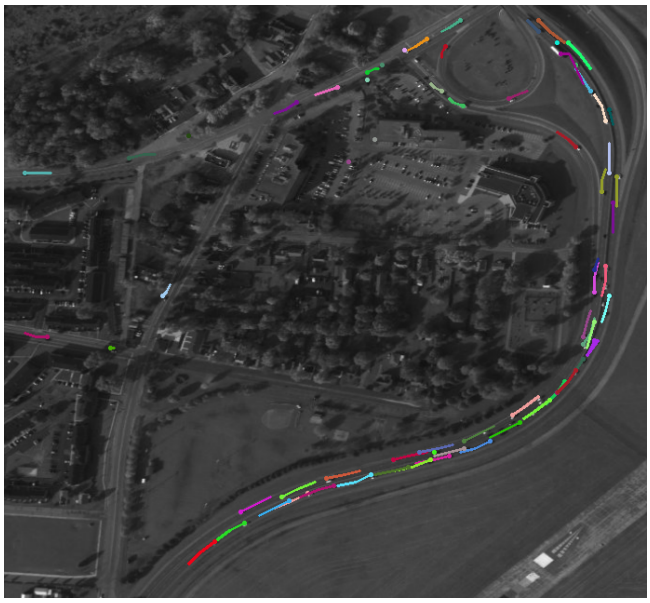


Fig. 3. Sample enhanced GM-PHD result.

4. CONCLUSIONS

We presented a modified version of the GM-PHD filter to improve object initialization and improve label tracking. We test our method in a real-world scenario using CNN-based detections and achieve real-time results with high FPS performance. Our method uses information from the CNN paired with a history of previous detections to improve both the tracking performance and the label management. We showed that our filter obtains high MOTA scores and mostly tracked trajectories while preserving a larger FPS (3 Hz) than the imaging system (2 Hz).

5. REFERENCES

[1] Jasper S. Wijnands, Haifeng Zhao, Kerry A. Nice, Jason Thompson, Katherine Scully, Jingqiu Guo, and Mark Stevenson, "Identifying safe intersection design through unsupervised feature extraction from satellite imagery," *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 3, pp. 346–361, 2021.

[2] Sanja Bauk, Nexhat Kapidani, Žarko Lukšić, Filipe Rodrigues, and Luís Sousa, "Review of unmanned aerial systems for the use as maritime surveillance assets," in *24th IEEE International Conference on Information Technology (IT)*, 2020, pp. 1–5.

[3] Wei Ao, Yanwei Fu, Xiyue Hou, and Feng Xu, "Needles in a haystack: Tracking city-scale moving vehicles from continuously moving satellite," *IEEE Transactions on Image Processing*, vol. 29, pp. 1944–1957, 2020.

[4] Michael Teutsch and Michael Grinberg, "Robust detection of moving vehicles in wide area motion imagery," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 1434–1442.

[5] Camilo Aguilar, Mathias Ortner, and Josiane Zerubia, "Small moving target MOT tracking with GM-PHD filter and attention-based CNN," in *IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021, pp. 1–6.

[6] Camilo Aguilar, Mathias Ortner, and Josiane Zerubia, "Small object detection and tracking in satellite videos with motion informed-CNN and GM-PHD filter," *Frontiers in Signal Processing*, vol. 2, 2022.

[7] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl, "Tracking objects as points," in *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds. 2020, pp. 474–490, Springer International Publishing.

[8] Hakki Motorcu, Hasan F. Ates, H. Fatih Ugurdag, and Bahadir K. Gunturk, "Hm-net: A regression network for object center detection and tracking on wide area motion imagery," *IEEE Access*, vol. 10, pp. 1346–1359, 2022.

[9] Lars Sommer, Wolfgang Kruger, and Michael Teutsch, "Appearance and motion based persistent multiple object tracking in wide area motion imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 3878–3888.

[10] 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*, vol. 21, no. 4, pp. 1252, 2021.

[11] U.S. Air Force Research Laboratory, "Wright-Patterson air force base (WPAFB) dataset," 2009, data retrieved from the SDMS website, https://www.sdms.afrl.af.mil/index.php.

[12] H. W. Kuhn and Bryn Yaw, "The Hungarian method for the assignment problem," *Naval Res. Logist. Quart*, pp. 83–97, 1955.

[13] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft, "Simple online and realtime tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2016.

[14] Ronald P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion*, Artech House, Inc., USA, 2007.

[15] Ba-Ngu Vo, Ba-Tuong Vo, and Hung Gia Hoang, "An efficient implementation of the generalized labeled multi-Bernoulli filter," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 1975–1987, 2017.

[16] Ba-Ngu Vo and Wing-Kin Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4091 – 4104, 2006.

[17] Kusha Panta, Daniel E. Clark, and Ba-Ngu Vo, "Data association and track management for the Gaussian mixture probability hypothesis density filter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 45, no. 3, 2009.

[18] Camilo Aguilar, Mathias Ortner, and Josiane Zerubia, "Adaptive birth for the GLMB filter for object tracking in satellite videos," in *IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 2022, pp. 1–6.

[19] 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 (CVPR)*, June 2018.

[20] Keni Bernardin and Rainer Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP Journal on Image and Video Processing*, 2008.

[21] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé, "MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 845–881, 2021.