Image Preprocessing and YOLO Architectures for Enhanced Small and Slow-Moving Object Detection

Diana Velychko
Rochester Institute of Technology
Rochester, U.S.A
dv6943@rit.edu

Saurav Singh
Rochester Institute of Technology
Rochester, U.S.A.
ss3337@rit.edu

Panos P. Markopoulos
The University of Texas at San Antonio
San Antonio, TX, USA
panagiotis.markopoulos@utsa.edu

Eli Saber
Rochester Institute of Technology
Rochester, NY, USA
esseee@rit.edu

Jamison Heard
Rochester Institute of Technology
Rochester, NY, USA
jrheee@rit.edu

2024 IEEE Western New York Image and Signal Processing Workshop (WNYISPW) | 979-8-3315-0555-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/WNYISPW63690.2024.10786503

***Abstract*—This paper enhances the detection of small and slow-moving objects in satellite video imagery by integrating classical signal processing techniques, such as Accumulative Multiframe Differencing (AMFD) and Low-Rank Matrix Completion (LRMC), with deep learning models. We conduct experiments on the Video Satellite Objects (VISO) dataset using YOLOv5, YOLOv8, and YOLOv10 models. Notably, AMFD outperformed LRMC and a pre-trained YOLOv5, achieving a precision of 0.540, recall of 0.210, and F1-score of 0.300. Furthermore, a YOLOv10 model trained from scratch on VISO for 250 epochs demonstrated superior performance, with a precision of 0.766, recall of 0.334, and F1-score of 0.465. Low-resolution images (220×286 pixels) achieved the highest precision (0.990) and F1-score (0.427). This study underscores the challenges in satellite imagery object detection, particularly regarding domain adaptation and resolution impacts, and paves the way for more effective object tracking.**

***Index Terms*—Computer vision, YOLO, remote sensing imagery, activity-based intelligence, domain adaptation**

I. Introduction

Activity-Based Intelligence (ABI) leverages large-scale remote sensing data to detect, analyze, and understand patterns of human activity in complex environments. Advancements in very high-resolution (VHR) satellite imagery, with spatial resolutions of 1 meter or less, have significantly improved the ability to detect small objects in satellite videos, a critical first step toward accurate object tracking [1]. Along with spatial improvements, the frame rate for sequential VHR imagery has increased to 10 frames per second, providing the necessary temporal resolution to support the detection and eventual tracking of moving objects in dynamic environments [2].

This research was supported by an academic grant from the National Geospatial-Intelligence Agency (Grant No. HM04761912014, Project Title: Target Detection/Tracking and Activity Recognition from Multimodal Data.) Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NGA, DoD, or US government. This material is based upon work supported by the National Science Foundation under Award Numbers DGE-2125362 and 2332744. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Detection is an integral component of the universal tracking system, serving as the foundation for identifying objects before tracking their movements over time. Accurate detection is especially important in remote sensing, where small objects like vehicles may appear as only a few pixels and exhibit irregular shapes. Moreover, satellite imagery often suffer from variability in resolution, elevation, and sensor quality, adding to the complexity of detection. Current state-of-the-art models, such as YOLOv8, offer fast and lightweight detection capabilities but face limitations in the remote sensing domain, especially in handling dense scenes where resizing images can degrade detection accuracy [3].

Improving detection in satellite imagery is therefore a key step towards overcoming challenges in object tracking. Effective detection ensures that objects are consistently identified across frames, enabling better tracking performance by algorithms such as BotSort[4] and ByteTrack [5] by minimizing issues such as object occlusion, scene clutter, and misidentification caused by resolution changes. Prior approaches to remote sensing detection, including both mathematical methods [2], [6], [7], [8] and deep learning models [3][9], have shown potential but often struggle with the unpredictability of object appearance, occlusions, and environmental factors [2].

This paper focuses on improving object detection in satellite videos with the ultimate goal of paving the way for superior tracking performance. We propose a novel detection architecture which attains enhanced detection of small and slow-moving objects. Accumulative Multiframe Differencing (AMFD) and Low-Rank Matrix Completion (LRMC) [2] are used to address key challenges, such as varying resolution and occlusions. Our method attains superior detection accuracy, laying the groundwork for more effective tracking systems in future applications.

II. Proposed Method

Our method integrates classical signal processing techniques with deep learning to detect moving objects in video frames. Specifically, our approach utilizes AMFD [2] and LRMC [10]

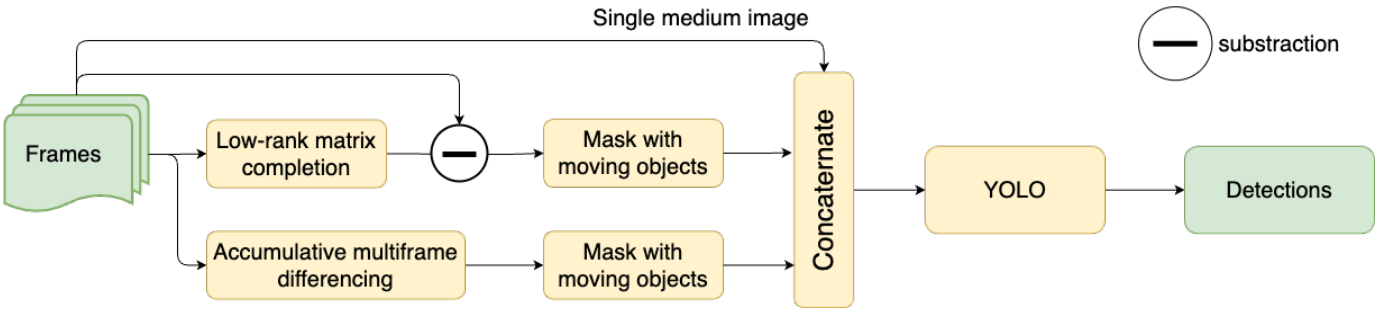


Fig. 1: Proposed object detection framework.

to generate masks with moving objects from the collection of consecutive frames within a one-second time interval as a preprocessing and feature extraction stage. Then extracted feature masks are concatenated with a central image from an original set of frames into a multi-channel input. These features are fed into a modified YOLO network that performs object detection and classification (see Figure 1). Incorporating mathematical methods in the preprocessing stage increases robustness and domain adaptability.

A. Accumulative Multi-Frame Differencing

AMFD is employed to detect moving objects in satellite video sequences. For each frame I_t, a sub-group of three neighboring frames (I_{t-1}, I_t, I_{t+1}) is considered. Three difference masks $(D_{t_1}, D_{t_2}, D_{t_3})$ are computed between these frames and then accumulated and normalized to create an accumulative response image I_D. This image is then binarized using a threshold T, calculated as $(\mu + 4\sigma)$, where μ and σ are the mean and standard deviation of I_D, respectively. Morphological operations are then applied to the binary image. False alarms are removed by retaining only connected areas with sizes between 4 and 100 pixels and aspect ratios between 0.25 and 6.0. This process is repeated for all frames in the sequence, resulting in a set of candidate-moving pixels.

B. Low-Rank Matrix Completion

LRMC is based on fast robust matrix completion (FRMC) and models video frames as low-rank matrices with perturbations to detect moving objects. FRMC employs the in-face extended Frank-Wolfe algorithm to solve this optimization problem. First, the number of observation matrices $N = M/(L * f)$ is calculated, where M is the total number of frames, f is the frame frequency, and L is the number of frames required to model the background (set to 4 in this implementation). For each group of L frames, an observation matrix $\mathbf{V}$ is constructed, where each column represents a video frame. The problem is then formulated as:

$$\min \ \text{rank}(\mathbf{B}) \ \text{s.t.} \ \mathbf{V} = \mathbf{B} + \mathbf{F} \tag{1}$$

where $\mathbf{B}$ is the background matrix, and $\mathbf{F}$ represents the foreground perturbation (moving objects). This optimization problem is relaxed into a nuclear-norm minimization (the convex envelope of the rank function) and solved using the computationally efficient Frank-Wolfe iterative algorithm, which estimates the background and extracts the foreground [11]. The resulting foreground image is binarized and morphological operations are applied to refine the detection results. This process is repeated for all frames in the video sequence, producing preliminary areas with moving targets.

III. Experimental Design

This study uses the Video Satellite Objects (VISO) dataset [2], which consists of 47 videos (15455 total images) and 1,646,038 labeled instances across 4 classes: cars, airplanes, trains, and ships. The dataset offers various scenes and resolutions. The training-test split is 80:20.

The dataset includes a range of image resolutions, which were divided into high-, medium-, and low-resolution images with a width/height in range $[1024, 1348]$, $[451, 512]$, and $[220, 286]$, respectively, to evaluate the impact of resolution on the model performance. Experiments were conducted with different mathematical algorithms and YOLO architectures:

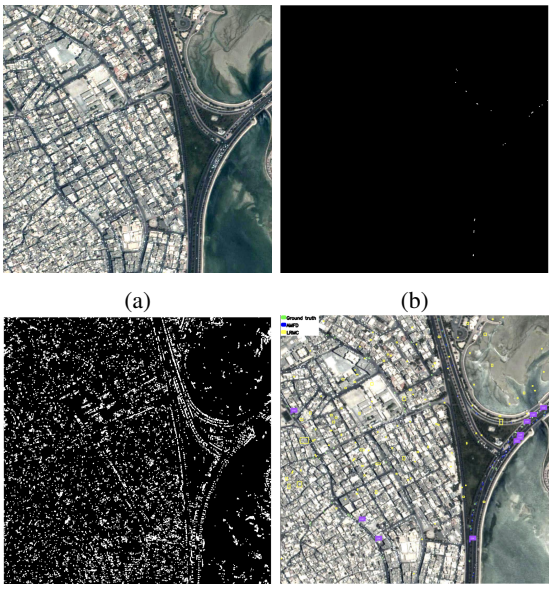
- YOLOv5 pre-trained on DOTA.
- YOLOv5x pre-trained and fine-tuned for VISO dataset.
- YOLOv8x pre-trained and fine-tuned for VISO dataset.
- YOLOv10 pre-trained and fine-tuned for VISO dataset.
- YOLOv10 from scratch trained on VISO dataset.

IV. Preliminary Results

AMFD shows a smaller yet more precise number of detections, while LRMC has multiple false detections (see Figure 2). In addition, Figure 2(c) shows that due to jitter many buildings are optimized as moving objects by LRMC.

Table I shows the evaluation results for preprocessing. AMFD achieves higher precision, recall, and F1-score than LRMC and YOLOv5 which was trained on an out of distribution dataset (DOTA [12]). Hence, the deep-learning model YOLOv5 that was trained on out-of-distribution remote sensing imagery generalizes worse than one of the mathematical models, leading to the expectation that incorporating features extracted by AMFD can permit a better domain adaptation.

We ran a few experiments on YOLO backbone architectures YOLOv5 and YOLOv10, their performance is provided in Table II. The best F1-score of 0.465 is achieved on YOLOv10 which was fully trained on VISO for 250 epochs. At the same time, precision is the highest in YOLOv8 which was fine-tuned



(a) (b) (c) (d)

Fig. 2: Preliminary results for all elements utilized in the preprocessing stage. (a) Original image; (b) Mask of moving objects created with AMFD method; (c) Mask of moving objects created with LRMC method; (d) Photo with ground truth, AMFD, LRMC, and proposed method detections (green, blue, yellow, and purple-colored bounding boxes, respectfully).

Metric	AMFD	LRMC	YOLOv5 (trained on DOTA)
Precision	0.540	0.006	0.330
Recall	0.210	0.252	0.039
F1-score	0.300	0.012	0.078

TABLE I: Results on preprocessing methods.

for 50 epochs. Interestingly, the trade-off between precision and recall is observed while increasing the number of epochs for fine-tuning YOLOv10.

The same experiments were run to test the performance of the models with grayscale images; see Table II. The results show that the performance of the proposed model with masks stays within the precision of the mathematical model, while all other models perform significantly worse.

The last experiment (see Table III evaluated the impact of image resolution (low, medium, and high) on YOLOv10. The Low resolution imagery demonstrates the highest precision (0.990) and F1-score (0.427). In contrast, the high-resolution imagery exhibits the lowest precision (0.482) and recall (0.150), suggesting challenges in identifying relevant instances in complex scenarios.

V. Discussion

Detection of objects in aerial imagery presents unique challenges, including small-scale objects, occlusions, extremely complex scenes, and VHR. LRMC works well in scenarios

Model	Precision	Recall	F1-score	mAP50
RGB input				
YOLOv5	0.751	0.347	0.475	0.487
YOLOv10	0.537	0.307	0.391	0.319
YOLOv10*	0.632	0.13	0.216	0.151
YOLOv10+masks	0.766	0.334	0.465	0.343
Grayscale input				
YOLOv5	0.217	0.058	0.092	0.032
YOLOv10	0.239	0.11	0.151	0.089
YOLOv10*	0.208	0.032	0.055	0.022
YOLOv10+masks	0.549	0.301	0.389	0.226

TABLE II: Results on YOLOv5 pre-trained on ImageNet and fine-tuned on VISO, YOLOv10 trained from scratch on VISO, YOLOv10* trained from scratch on VISO in grayscale, and proposed method YOLOv10 with masks AMFD and LRMC trained from scratch for 250 epochs. Includes results on YOLO models with grayscale input.

Resolution	Precision	Recall	F1-score	mAP50
Low	0.990	0.272	0.427	0.237
Medium	0.850	0.134	0.232	0.069
High	0.482	0.150	0.229	0.038
All	0.766	0.334	0.465	0.220

TABLE III: Performance metrics by resolution.

that contain relatively large moving objects. However, in the case of satellite imagery, small targets typically occupy area of $4-50$ pixels and the change in the movement is relatively minor, so often the targets are modelled as the background yielding a reduction in performance.

LRMC requires calculating the number of observation matrices that are based on the total number of frames captured in the video [10]. Such an approach limits the utilization of the algorithm to be applied to pre-captured videos unless the number of frames required is known a-priori, which is not always satisfied. In addition, background subtraction relies on a few frames being processed through an extensive number of matrix calculations, which increases inference time. In the case of satellite imagery, the jitter from satellite movement requires the LRMC algorithm to run at a predetermined frequency.

Deploying a model in the real world often involves varying degrees of domain shift. In remote sensing, the complexity and variability of aerial imagery datasets can lead to significant differences in feature distributions, challenging model consistency and performance. In the case of YOLOv5 trained on DOTA, we observe that there are issues with the perception of the model (the bounding boxes are too large for the observations). Although the locations of the bounding boxes are mostly correct, their amount and sizes are inaccurate (see Figure 3). These observations demonstrate the heightened need for domain adaptation approaches in satellite imagery.

YOLOv5 trained on out-of-distribution satellite data (trained on DOTA and tested on the VISO dataset), detects a few

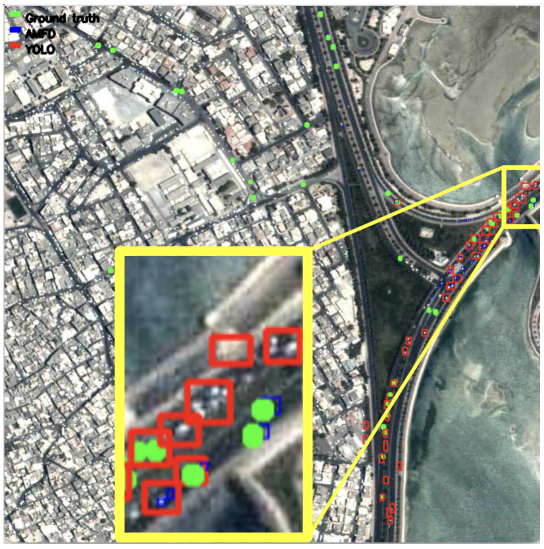


Fig. 3: Object detection results (human labels in green; YOLOv5 trained on DOTA in red; AMFD in blue). Yellow bounding box demonstrates domain shift issue of the model trained on DOTA and deployed on VISO.

densely situated vehicles as one instance in certain cases, while our model detects closely nested objects as separate objects. Due to the difference in altitude of the satellites, the objects may appear bigger or smaller in photos from different datasets. Hence, a model trained only on data from the satellite with one altitude is prone to being biased towards a certain size of the vehicles. As a result, vehicles situated closely to each other are often predicted as one instance if the model is trained on a different dataset. Moreover, the lack of adaptation to the size of the object can be considered a challenge unique to satellite imagery. Photos taken on the ground naturally include perspective, and the same object can appear in different sizes, allowing the model to more properly generalize. In the case of satellite imagery, all objects appear mostly flat, as the distance in Earth elevation in the vast majority of instances is neglectful compared to the distance from the ground to the satellite.

Satellite-based observations inherently do not have a perspective available to ground-based observations because one satellite captures all objects at a fixed elevation level while ground cameras capture the same object in different sizes relative to the size of the image as it approaches the camera. For that reason, we observe a scale-related domain shift.

The main purpose of improving moving object detection is to enhance tracking accuracy. The misregistration phenomenon manifests as different tracking ids are assigned to the same object that temporarily disappeared from detections but not from the scene. As a result, instead of one continuous track for the object, there are multiple disconnected tracklets belonging to the same object. Misregistration can be mitigated by employing bounding box fusion techniques to smooth over abrupt transitions or displacements by considering the most likely path for an object.

VI. Conclusion

In this paper, we propose a novel architecture that integrates temporal elements into YOLO to enhance the detection of small and slow-moving objects in satellite imagery. The proposed method combines AMFD and LRMC preprocessing, enhancing adaptability, robustness, and detection accuracy in satellite video frames with small, slow-moving objects. Preliminary results on the VISO dataset show that AMFD outperforms both LRMC and a pre-trained YOLOv5 model on the DOTA dataset. YOLO models, fine-tuned on VISO dataset, provide better performance. Furthermore, YOLOv10 trained from scratch achieved the best F1 score, while fine-tuned YOLOv8 had the highest precision. Results highlight the effects of resolution on model performance, demonstrating the deficiency of model predictions on high resolution images. Future work involves incorporating a ConvLSTM layer or preprocessing module to leverage temporal information.

References

[1] Z. Cai, H. Wei, Q. Hu, W. Zhou, X. Zhang, W. Jin, L. Wang, S. Yu, Z. Wang, B. Xu, and Z. Shi, "Learning spectral-spatial representations from vhr images for fine-scale crop type mapping: A case study of rice-crayfish field extraction in south china," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 199, pp. 28–39, 2023.

[2] Q. Yin, Q. Hu, H. Liu, F. Zhang, Y. Wang, Z. Lin, W. An, and Y. Guo, "Detecting and tracking small and dense moving objects in satellite videos: A benchmark," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[3] G. Jocher *et al.*, "Yolov8 by ultralytics. 2023," *URL https://github.com/ultralytics/ultralytics*, 2023.

[4] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[5] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[6] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 430–443.

[7] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 778–792.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[9] M. Sharma, M. Dhanaraj, S. Karnam, D. G. Chachlakis, R. Ptucha, P. P. Markopoulos, and E. Saber, "Yolors: Object detection in multimodal remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1497–1508, 2020.

[10] B. Rezaei and S. Ostadabbas, "Background subtraction via fast robust matrix completion," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1871–1879.

[11] M. Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 427–435. [Online]. Available: https://proceedings.mlr.press/v28/jaggi13.html

[12] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.