2026 NDIA MICHIGAN CHAPTER
GROUND VEHICLE SYSTEMS ENGINEERING
AND TECHNOLOGY SYMPOSIUM
APPLIED ARTIFICIAL INTELLIGENCE TECHNICAL SESSION
AUGUST 11-13, 2026 - NOVI, MICHIGAN

CHANGE DETECTION FILTERING WITH VISUAL LANGUAGE MODELS

Eric Martinson[1], Igri Fishta[1]

[1]Lawrence Technological University, Southfield, MI

ABSTRACT

Intelligence, surveillance and reconnaissance (ISR) often require review of significant quantities of video. While machine vision is used to flag objects for human review, too many flags are generated. Integrating newer methods like Open-Vocabulary Object Detection (OVOD) that support zero shot detection, but with significantly lower accuracy only make the problem worse. This work addresses the utility of OVOD in ISR missions by focusing only what has changed between successive runs through an environment. A Vision Language Model (VLM) compares current observations against a registered "cleared" baseline to focus only on what has changed. Testing across three distinct environments, and using either monocular camera phones or RGB-D equipped vehicles, demonstrates that integrating change detection can automatically remove as much as 80% of unchanged objects without impacting recall.

Citation: E. Martinson, I. Fishta, "Change Detection Filtering with Visual Language Models," In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 11-13, 2026.

1. INTRODUCTION

The volume of video data collected by military personnel and autonomous systems has far outpaced the capacity of human analysts. While data reduction tools and deep learning models (e.g., YOLO) provide automated filtering, they remain rigid, requiring extensive hand-labeled datasets and prior knowledge of specific threats. This lack of agility is a critical failure point in dynamic reconnaissance environments where analysts must adapt to emerging threats or dual-use materials.

Open-Vocabulary Object Detection (OVOD) models, such as CLIP and SAM 3, offer a solution by allowing arbitrary text queries. However, their reliability varies significantly across different object classes. This work proposes enhancing OVOD performance by leveraging temporal and spatial context. In scenarios involving repeated patrols, we argue that analysis should prioritize environmental changes rather than re-evaluating static features. By focusing exclusively on what has changed since a previous "clear" state, we demonstrate that as much as 90% of false-positive flags can be discarded without impacting recall – depending on the complexity of the environment.

Our approach utilizes image registration to map multiple passes into a unified coordinate space. Potential detections are then verified by a Vision Language Model (VLM), which compares multi-angle frames to confirm the physical presence of a new object. We validate this pipeline using both robotic RGB-D and handheld mobile data, comparing VLMs of varying sizes to establish performance benchmarks for resource-constrained military applications.

2. RELATED WORK

Traditional ISR pipelines rely heavily on supervised object detection frameworks like YOLO or PointNet [1]. While highly effective for known targets, these systems are hindered by the "labeling bottleneck," where adding new object classes requires significant human effort and expensive datasets [2]. In dynamic military environments, where threats may involve arbitrary dual-use materials or novel IED components, this rigidity is a liability. To address this, recent research has pivoted toward Open-Vocabulary Object Detection (OVOD). Foundations such as CLIP [3], ALIGN [3], and more recently, SAM v3 [4] allow operators to use natural language “concept prompts” or queries to search for targets without prior training, significantly increasing mission agility. The integration of OVOD with LLMs and VLMs has further expanded the scope of autonomous scene understanding. Recent work has demonstrated that multi-modal fusion can improve detection accuracy by using LLMs to refine search queries, incorporate environmental context, and apply physical constraints like object size [5]. Furthermore, models like Kosmos-2 [6] and LISA++ [7] have moved beyond simple classification to provide language-based segmentation and abstract reasoning. However, while these "foundation models" are versatile, their performance in high-stakes ISR tasks can be inconsistent across different natural language queries and/or object classes; consequently, seemingly simple detection tasks result in single digit F-scores when analyzing streaming image sources [5].

To mitigate the reliability issues of open-vocabulary systems, researchers have looked to change detection as a powerful contextual filter when repeated, temporally distinct observations are available. Classically, this begins with image registration to align disparate captures [8] and create matching before and after images from different runs. While relatively straightforward with downward facing aerial or satellite imagery, ground level imagery, particularly indoors or anywhere with limited sightlines, have remained significantly more challenging due to viewpoint variations and sensor noise. Attempts to overcome these limitations have included super-pixel-based difference maps [9], dense 3D reconstructions [10], and Neural Radiance Field models [11]. These approaches then require an additional filtering step to separate meaningful change from unimportant differences like lighting or seasonal variations. Solutions vary from deep learning models trained on hand-labeled data [10] to open vocabulary methods like CLIP [11] or an LLM [12].

When change can be defined by specific object categories (e.g. signs, cars, buildings, etc), then object-level differencing can be used in place of side-by-side imagery. Applications range from narrow tasks like detecting storm damage in satellite imagery [13], to creating full semantic maps of all detected objects in urban and indoor areas [14], comparing newly detected objects with the existing map. More recently, open-vocabulary methods have been integrated to provide more flexibility in the objects being tracked [15], but the high variability of

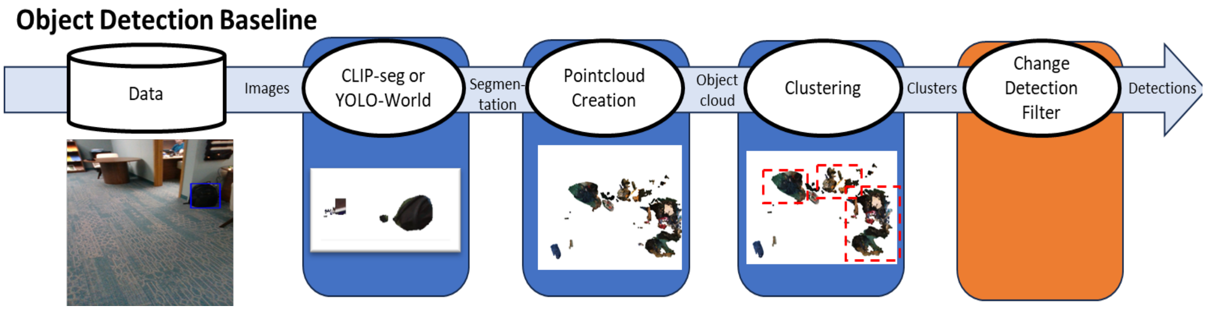

Figure 1. General fusion architecture for improving open vocabulary object detection. Change filtering is the last step after clustering.

detection between runs significantly limits overall efficacy of the approach in practical scenarios.

This research introduces a hybrid approach that anchors open-vocabulary queries within a spatially registered temporal framework. By verifying side-by-side changes through a VLM, we achieve the versatility of arbitrary search while maintaining the structural reliability of classical change detection

3. ALGORITHMS

This section describes algorithms used to support change detection filtering beginning with the object detection pipeline and concluding with the VLM empowered change detection filter. Intermediate steps discuss how to synthesize depth data for use with mobile phone cameras and registering images to the same geospatial coordinate system using COLMAP.

3.1. Object Detection Baseline

The baseline detection pipeline utilizes a multi-stage process to fuse inconsistent 2D open-vocabulary detections into a persistent 3D spatial representation (Figure 1). Initially, RGB-D data and 6-DOF camera poses are captured via robotic localization or image registration. Individual color frames are processed using SAM3 to generate segmentation masks based on arbitrary text queries provided by an analyst at runtime. Because zero-shot detections in streaming video are notoriously noisy and viewpoint-dependent, 2D pixels exceeding a 0.5 likelihood threshold are projected into 3D space using aligned depth data. To maintain real-time performance on edge-compute hardware, a spatial filter is applied to discard redundant frames within 5 cm or 0.05 radians of previous captures, reducing the computational load by approximately 75% without compromising detection density.

The resulting unorganized 3D point cloud is then voxel-downsampled and processed using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to extract discrete object hypotheses. This spatial aggregation effectively filters out transient false positives and background noise that plague individual 2D frames. Final object clusters are filtered based on cluster size and mean likelihood; however, as noted in previous work [5], the efficacy of these thresholds remains highly object-dependent.

3.2. Synthesizing Depth Data

Depth data are required to fuse detections into a single point cloud. If data are captured by a mobile phone or equivalent, then depth data need to be synthesized. This work leverages data capture libraries like Spectacular AI [16] and Reality Scan that align images with onboard IMU. Similar results could be achieved with GPS stamped image streams.

Image data are used to train a Gaussian Splatting model [17]. We use the Nerfstudio environment and the splatfacto model

Figure 2. Gaussian Splatting model of the home environment showing synthetic color images/image collection locations (top) and synthetic depth image (bottom)

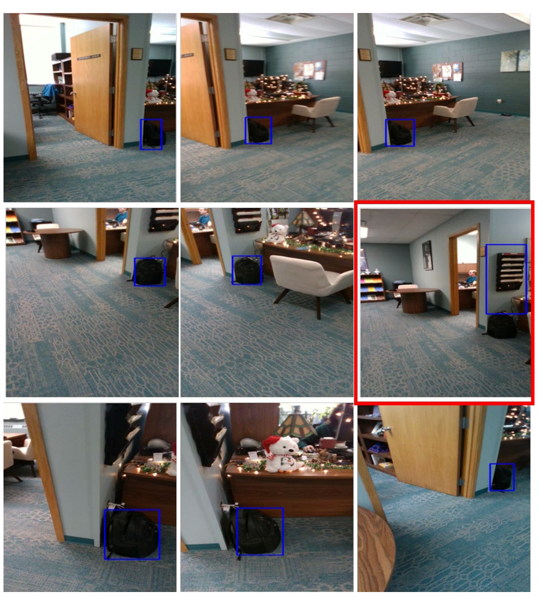

Figure 3. Example images collected and annotated for a single cluster when searching for small items left behind in the office environment. The red box indicates an incorrectly identified annotation

specifically, making sure to register images with respect to geospatial coordinates (Figure 2). The resulting model supplies synthetic depth images for every registered color image.

3.3. Image Registration for Change Detection

Change detection requires images captured at two distinct times, but registered to the same geospatial coordinate system. The approach taken in this paper is to first create a baseline coordinate system from the "cleared" or earlier run. COLMAP [18] works for this purpose, registering images to an arbitrary coordinate space, and then realigning to specific geospatial information (GPS or IMU).

COLMAP also supports adding new images to this same reconstruction. Assuming that the not too much of an image has changed, images taken at a separate time can be added in this way and registered to the same coordinate space. Re-applying known GPS or IMU coordinates for the baseline run helps make sure the scale and coordinate system do not change too much.

This approach for adding a new sequence of images captured at a different time is robust to modest amounts of environmental change. We removed objects, opened windows to change lighting, ransacked bookshelves, scattered papers, etc., and still registered most images correctly.

3.4. Change Detection Filter

Every cluster has a set of associated images viewing the object. A bounding box is drawn around the object in each image reflecting the edges of the cluster point cloud in the current image frame. Figure 3 demonstrates such a set of images for a single object. N random images are selected for further evaluation with the change detection filter.

For each selected image i, the relative distance image between image j in the baseline "cleared" run is calculated using the the sum of guassian distance functions:

$$S\left(dist_{i,j}, \Delta\Theta_{\mathrm{i,j}}\right) = \left(\frac{-dist_{i,j}{}^{2}}{2\sigma_T^2}\right)^{1/2} + \left(\frac{-\Delta\Theta_{\mathrm{i,j}}{}^{2}}{2\sigma_A^2}\right)^{1/2}$$

Where $dist_{i,j}$ is the Euclidian distance between image coordinates, and $\Delta\Theta_{i,j}$ is the delta angle between the vectors along the z-axis of each image. Standard deviations for distance (σ_T) and angle (σ_A) were determined empirically as 1.0-m and 0.25-rad.

The nearest image that can also view the object centroid is selected, and the two images are merged together (Figure 4) and supplied to a VLM along with the query:

The attached image includes two side by side images of the same location captured at different times.A box has been drawn in the right most image to specify the area of interest.Please tell me if there is a new object contained inside the blue box on the right that was not present in the image on the left.If there is a new object, answer True. If the area has not changed, answer False. Return an answer in JSON format as {'is_new_object': <True/False>}

The number of VLM responses of *True* is divided by the number of evaluated images, *N*. To create a single likelihood of this cluster depicted a change to the environment. Clusters with likelihood < threshold are discarded or else reduced in importance. This sample-based response allows the VLM filter to overcome poor image registration

4. EVALUATION DATASETS

Three datasets were collected: 1 with the mobile phone, and 2 with the Stretch-2 robot. Datasets were originally published in [11]. Home space data were collected with permission of the owner without people present:

Figure 4. Example merged images processed by the VLM to search for change in the lecture space. (top) A "small item" has been introduced. (bottom) A keyboard in a hanging poster is flagged as "electronics"

- **Home 1 (Mobile Phone)** tracks changes in a living room space of approximately 5m x 4m. 407 baseline images were collected. 4 change tracks with ~100 images each were collected across two days of recording. Tracked changes include dishes, toys, computers, cables, books, clothing and blankets. Figure 4 depicts the gaussian splat model of the home space. The
- **Presentation Space (Robot)** is a simulated lecture space (~4m x 4m) in a visually cluttered environment. Posters along all walls with pictures generate significant false positive detections with all classifiers (Figure 3). To account for people moving furniture and/or walls in this shared space, baseline data were collected at the beginning of each of 3 days of trials. 10 different change runs were completed, tracking phones, chargers, notepads, bags, soda bottles, outerwear (hat, gloves, jacket) – items commonly left behind in lecture halls. Baseline data have ~350 images per run, while change runs have ~180 images.
- **Office (Robot)** is a working office space of 6.5m x 2.5m characterized by a moderate

Table 1. Counting the clusters resulting from each generic query when searching the environment and the numbers of clusters representing changes

	Home		Presentation Space		Office		Combined	
	Unchanged	Changed	Unchanged	Changed	Unchanged	Changed	Unchanged	Changed
clothing	1	2	N/A	N/A	N/A	N/A	1	2
dishes	0	1	N/A	N/A	N/A	N/A	0	1
electronics	N/A	N/A	42	9	31	7	73	16
general clutter	0	0	40	0	0	0	40	0
small items	45	15	24	21	63	21	132	57
trash	N/A	N/A	4	8	1	14	5	22
						Total	251	98

level of visual clutter and frequent human activity. The space contains desks, shelving, personal items and office equipment that remain largely consistent over time, with variability primarily introduced by personal belongings of individuals using the office. Data were collected over multiple consecutive days to capture natural variations in object placement, while maintaining a largely stable scene structure. Change runs reflect realistic usage of the space, where items such as papers, chargers, soda bottles, gloves, backpacks and lunch boxes were temporarily left behind by occupants entering or working in the office (Figure 3). Because the office was still being used during data collection, some sampling images include people within the scene, requiring the system to tolerate partial occlusions and dynamic elements. Baseline data have ~420 images per run, while changes runs have ~200 images.

5. RESULTS

5.1. Object Detection Baseline

Each of the three datasets was searched with 4 different natural language queries. The home dataset was searched for “clothing”, “dishes”, “general clutter”, and “small items”. The office and presentation space were each searched for “electronics”, “general clutter”, “small items” and “trash”.

As demonstrated in Table 1, some queries generated significant numbers of clusters that did not change between runs. If the environment had been cleared prior, then >70% of objects could be removed automatically using change detection.

Note that these queries were non-item specific, representing searches where the operator was uncertain of what they were looking for. Queries focused on specific items like backpacks, wiring, etc, generates far fewer clusters – as demonstrated by searches for clothing or dishes.

5.2. Change Filtering

Clusters from all runs were evaluated with 5 different VLMs of different sizes available from the Ollama [19] libraries that could run on a single A100 card with 80GB of memory:

- Gemma3 (12B, 27B)
- Qwen (7B, 32B)
- Llama4 (90B) [20]

Models were evaluated on their ability to separate clusters representing a real change from those that remained the same in both runs.

Figure 5 plots the ROC-Curve for the best performing model (Llama4:90B) against all three environments. Curves demonstrate the relative difficulty of each space. The home environment, created without depth data, is the noisiest, but >80% of unchanged objects can be discarded while retaining 80% of the

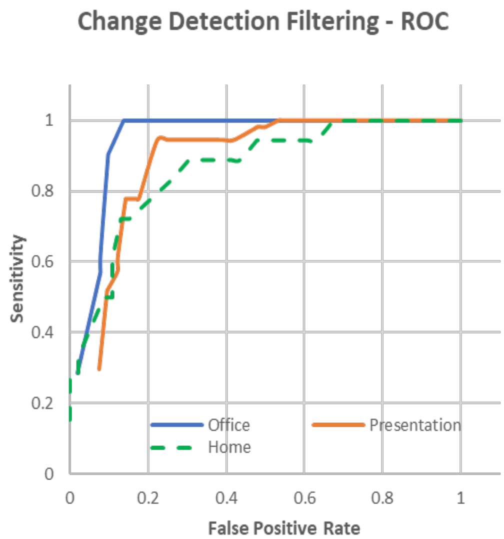

Figure 5. ROC curve describing the effectiveness of the change detection filter in separating out changed clusters in each of the 3 tested environments

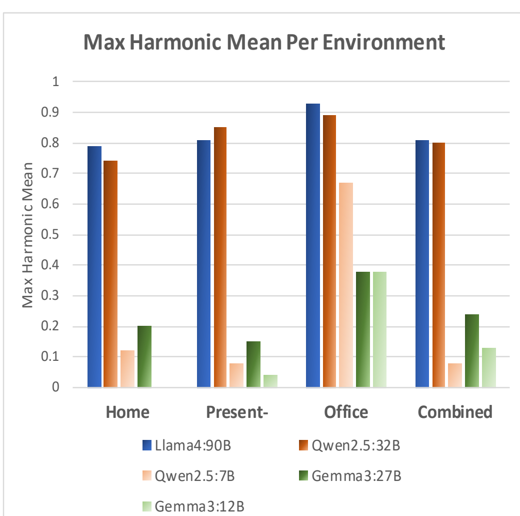

Figure 6. Evaluating the max harmonic mean of specificity and sensitivity per model per environment. Llama and Qwen large models outperform all others by significant margins.

changed objects. The presentation space, with many posters of around the room, is the next most difficult, dropping 80% of unchanged objects while retaining 90% of detected changes. Finally, the office space is the simplest, retaining 100% of detected changes with the same threshold.

Figure 6 examines the variations between different models, focusing specifically on the harmonic mean of specificity and sensitivity:

$$HM = \frac{2 * Spec * Sens}{Spec + Sens}$$

This metric was selected to emphasize the importance of specificity (or 1 – FPR), as the goal is to remove as many unchanged objects as possible while dropping few changed objects.

As expected, the largest models generally did the best. The Llama4 model, which is a 90B parameter model, had the highest harmonic mean score (0.81). The largest qwen model had a similar performance (0.80), with less than half the number of parameters. Beyond that, performance was less predictable. The larger 27B gemma model showed some ability to differentiate (0.24), but was only slightly better than the much smaller Qwen and Gemma models.

6. CONCLUSIONS

This research demonstrates that integrating change detection filtering into Open-Vocabulary Object Detection (OVOD) pipelines can significantly reduce the load on human analysts when clearing new video data. By leveraging temporal context—comparing current observations against a registered "cleared" baseline—the system can automatically discard up to 80% of static, unchanged objects that would otherwise burden an analyst.

Evaluations conducted in this work cross three distinct environments (Home, Presentation Space, and Office), and multiple collection devices (phone and autonomous ground vehicle). These conclusively demonstrate the effectiveness of mid-size Vision Language Models (VLMs) like Llama4:90B and Qwen2.5vl:32B, which reliably distinguish between meaningful environmental changes and persistent scene features across all environments - even in the presence of noise and visual clutter.

6.1. Addressing System Latency

While the proposed pipeline effectively reduces the manual review burden, operational latency remains a significant challenge for real-time deployment. Currently, bottlenecks are concentrated in two areas: initial image registration and high-parameter VLM processing. Spatial registration via COLMAP, while robust to environmental changes, is computationally expensive when adding new image sequences. This can be significantly sped up through parallelization—breaking the problem up into pieces that can be solved separately before combining them together;

The VLM can also be improved through parallelization, but there is also potential to train lighter weight VLM models specifically for the change detection application rather than relying on generalized, larger models. This would enable use on lighter weight, less costly GPUs in the future.

With these proposed optimization strategies, we expect the system to process video significantly faster and be integrate-able ground-level operations in the future.

6.2. Data Representative of ISR

Current datasets, while useful for proof-of-concept, do not fully capture the specific classes of objects or the nuance of queries that analysts utilize. Beyond addressing latency issues described in the previous section, future work will investigate how this change-detection filtering can be better integrated with real ISR missions, evaluating utility in a wider variety of environments and investigating how to ingest external intelligence, such as SIGINT or prior mission reports, to further refine and prioritize detected changes.

7. REFERENCES

[1] Ultralytics, "YOLOv8," [Online]. Available: https://docs.ultralytics.com/models/yolov8/. [Accessed 22 Jan 2024].

[2] E. Martinson and F. Alladkani, "Interactive, Privacy-Aware Semantic Mapping for Homes," in Ubiquitous Robots, New York, USA, 2024.

[3] A. Radford et al, "Learning transferable visual models from natural language supervision," in Int. Conference on Machine Learning, Online, 2021.

[4] N. Carion et al, "SAM3: Segment Anything with Concepts," 19 November 2025. [Online]. Available: https://ai.meta.com/research/publications/sam-3-segment-anything-with-concepts/. [Accessed 15 December 2025].

[5] E. Martinson and H. Indurthi, "Augmenting Open Vocabulary Object Detection with Large Language Models for Home Service Robots*," in International Conference on Automation Science and Engineering (CASE), Los Angeles, 2025.

[6] Z. Peng et al, "Kosmos-2: Grounding Multimodal Large Language Models to the World," 2023 June 26. [Online]. Available: https://arxiv.org/abs/2306.14824.

[7] S. Yang et al, "Lisa++: An Improved Baseline for Reasoning Segmentation with a Large Language Model," 28 Dec 2023. [Online]. Available: https://arxiv.org/abs/2312.17240.

[8] D. Lu, P. Mausel, P. Mausel, E. S. Brondízio and E. Moran, "Change Detection Techniques," Int J. of Remote Sensing, vol. 25, pp. 2365-2401, 2004.

[9] K. Sakurada and T. Okatani, "Change Detection from a Street Image Pair using

CNN Features and Superpixel Segmentation," in British Machine Vision Conference, Swansea, 2015.

[10] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo and R. Gherardi, "Street-view change detection with deconvolutional networks," Autonomous Robots, vol. 42, pp. 1301-1322, 2018.

[11] E. Martinson, I. Fishta and D. Butani, "Integrating Foundation Models with Change Detection to Identify Tasking for Service Robots," To be published in Frontiers in AI, 2026.

[12] K. Ranasinghe, S. N. Shukla, O. Poursaeed, M. S. Ryoo and T.-Y. Lin, "Learning to Localize Objects Improves Spatial Reasoning in Visual-LLMs," in Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2024.

[13] T. Liu, L. Yang and D. Lunga, "Change detection using deep learning approach with object-based image analysis," Remote Sensing of Environment, vol. 256, no. 112308, 2021.

[14] L. Göncz and A. L. Majdik, "Object-Based Change Detection Algorithm with a Spatial AI," Sensors, vol. 22, no. 6342, 2022.

[15] Y. Obinata et al, "Semantic Scene Difference Detection in Daily Life Patroling by Mobile Robots using Pre-Trained Large-Scale Vision-Language Model," in Intelligent Robotics and Systems, Detroit, 2023.

[16] Spectacular AI, "Gaussian Splatting & NeRFs," 2025. [Online]. Available: https://spectacularai.github.io/docs/sdk/tools/nerf.html. [Accessed 14 December 2025].

[17] B. Kerbl, G. Kopanas, T. Leimkühler and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," ACM Transactions on Graphics, vol. 42, no. 4, 2023.

[18] J. Schoenberger, "Colmap FAQ," [Online]. Available: https://colmap.github.io/faq.html. [Accessed 20 Jan 2024].

[19] "Ollama Models," Ollama, [Online]. Available: https://ollama.com/search. [Accessed March 2026].

[20] Meta, "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation," 2025. [Online]. Available: https://ai.meta.com/blog/llama-4-multimodal-intelligence/. [Accessed 6 March 2026].

8. CONTACT INFORMATION

Eric Martinson
Associate Professor of Computer Science
Lawrence Technological University
College of Arts and Science
Math and Computer Science Department
21000 W 10-Mile Rd
Southfield, MI
emartinso@ltu.edu

9. ACRONYMS

ISR	Intelligence, Surveillance, and Reconnaissance
OVOD	Open-Vocabulary Object Detection
VLM	Visual Language Model
DBSCAN	Density-Based Spatial Clustering of Applications with Noise