Evaluating SAM 3 and YOLO26 Photovoltaic Panel Instance Segmentation Performance in Large-Scale Solar Power Plants

Ruben Dario Ramos Ciprian[1]; Rafael Omar Batista Jorge[2]; Darvy Betances Medrano[3]; Cesar Luis Rijo Cedano[4]
[1,4] Universidad Central del Este (UCE), Dominican Republic, rubenramos@uce.edu.do; cr2023-3478@uce.edu.do
[2,3] Pontificia Universidad Católica Madre y Maestra (PUCMM), Dominican Republic, r.batista@ce.pucmm.edu.do; dabm0001@ce.pucmm.edu.do

***Abstract*— The increasing deployment of photovoltaic systems has created a demand for accurate automated methods capable of identifying and segmenting photovoltaic panels in aerial imagery. Convolutional neural network architectures, particularly the You Only Look Once (YOLO) family, have been widely used for this purpose because of their inference speed and adaptability. More recently, foundation models have emerged as a new class of large pretrained models and are being widely explored in image segmentation, remote sensing, and other computer vision applications. However, few comparative studies have evaluated recent YOLO architectures and foundation segmentation models under the same experimental conditions. This work presents a benchmark comparison of the Segment Anything Model 3 (SAM3), YOLO26s, and YOLO26l for photovoltaic panel instance segmentation. The models were evaluated on 16 aerial images containing 49 annotated photovoltaic panel instances using standard segmentation metrics. SAM 3 achieved the best overall performance, with an AP50 of 96.45%, a mAP50–95 of 85.00%, and a recall of 91.84%. YOLO26s obtained the highest precision at 79.31% but showed lower recall and mAP50–95 values of 46.94% and 42.99%, respectively, while YOLO26l produced the weakest results. Overall, SAM 3 provided more accurate mask delineation and substantially greater detection coverage, highlighting the potential of foundation models for photovoltaic inspection.**

***Keywords*— Aerial Imagery, Image Segmentation, Photovoltaic Panels, SAM, YOLO.**

I. INTRODUCTION

The continuous expansion of photovoltaic energy systems has increased the need for automated methods capable of inspecting large solar installations accurately and efficiently. Utility scale photovoltaic plants may contain thousands of modules distributed across extensive areas, making manual inspection time consuming and difficult to scale. Unmanned aerial vehicles and high resolution imaging systems provide an effective mechanism for collecting visual information from these installations. However, the resulting images must be processed using computer vision methods capable of locating each photovoltaic panel and accurately separating it from adjacent modules, soil, vegetation, rooftops, and other background elements. Previous work demonstrated that YOLO models can effectively detect photovoltaic panels in single frame and orthomosaic images [1]. Nevertheless, detection through bounding boxes may not provide the detailed panel boundaries required for applications such as module counting, area estimation, fault localization, and spatial correlation with thermal anomalies. Consequently, instance and semantic segmentation have become relevant tasks for photovoltaic plant monitoring. PVNet, for example, was designed to extract detailed photovoltaic regions and boundaries from high resolution remote sensing imagery [2].

Convolutional neural networks have traditionally been the main architecture used for photovoltaic panel detection and segmentation. These models learn task specific visual features from annotated examples and can provide high accuracy when the training and testing data have similar characteristics. The You Only Look Once family is frequently selected for this type of application because it combines detection accuracy with relatively fast inference. In a previous evaluation, YOLOv8 and YOLOv11 models were trained using a mixed airborne dataset containing single frame and orthomosaic images [1]. The results showed that compact YOLO configurations could achieve high photovoltaic panel detection precision while reducing training time. More recently, YOLO26 has extended this CNN based model family through an end to end design that supports object detection and instance segmentation across multiple model scales. Its architecture includes mechanisms intended to improve small object assignment, simplify the prediction head, and eliminate non-maximum suppression during inference [3]. These characteristics make YOLO26 a relevant candidate for photovoltaic applications in which numerous panels must be processed within high resolution aerial images.

Despite their effectiveness, supervised CNN models usually require a representative annotated dataset and additional training for each new task. Their performance may decrease when the images contain changes in spatial resolution, viewing angle, illumination, panel arrangement, sensor characteristics, or geographic location. Photovoltaic modules present additional challenges because they may appear as small and densely arranged objects with similar shapes, colors, and textures. The PVNet architecture addressed this problem by combining coarse region prediction and boundary refinement [2]. Other studies have evaluated transformer based segmentation architectures using multiresolution satellite and aerial data [4], for instance, applied Mask2Former to photovoltaic module segmentation and compared its performance with conventional models such as UNet and DeepLabv3+. These studies demonstrate that specialized models can produce detailed photovoltaic masks, but their generalization remains closely related to the diversity of the training data and the availability of accurate pixel level annotations.

Foundation models have recently emerged as a different approach to computer vision. Instead of being developed exclusively for one predefined dataset or class, these models are pretrained using large and diverse collections of images and can subsequently be applied to multiple tasks. The Segment Anything Model introduced promptable segmentation through points, boxes, or masks and demonstrated the possibility of transferring segmentation knowledge to previously unseen image distributions [5]. Its successor, SAM 2, expanded this approach to images and videos while improving accuracy, inference speed, and interactive segmentation capabilities [6]. The introduction of these models has increased research interest in prompt based and zero shot segmentation across medical imaging, autonomous systems, industrial inspection, and remote sensing. This growing use is motivated by their ability to reduce task specific training requirements and support the segmentation of objects that were not explicitly defined during the original pretraining process.

The use of foundation models for photovoltaic segmentation has also increased. In [7] SAM, SAM 2, and SAM 2.1 are evaluated for the segmentation of photovoltaic panels in aerial RGB imagery under different lighting and spatial resolution conditions. Their study also used YOLOv9 to generate automatic prompts, illustrating that CNN detectors and foundation segmentation models can be integrated within the same processing pipeline. The results indicated that prompt selection and image resolution have a significant effect on segmentation quality. More recently, PVSAM introduced geometric and multiscale prompt modules to adapt SAM features to photovoltaic structures in multisource remote sensing images [8]. This approach improved boundary sensitivity and cross scenario generalization, but it also required domain specific modifications to the original foundation model. A different direction was explored in [9] through a CLIP based framework that generated photovoltaic segmentation masks without downstream parameter training . Together, these studies show that foundation models are moving from general purpose experimentation toward specialized photovoltaic and remote sensing applications.

SAM 3 represents a further development of this model family by introducing prompt based concept segmentation. It can detect and segment all instances associated with a concept described using text, visual examples, or a combination of both [10]. This capability differs from conventional task specific CNN segmentation, where the target class is learned through supervised optimization using annotated images. The comparison between SAM 3 and YOLO26 therefore represents more than an evaluation of two recent architectures. It examines two different computer vision paradigms: a large foundation model that performs concept guided segmentation and a specialized CNN architecture trained directly for photovoltaic panel instances. Although previous studies have evaluated earlier SAM versions, adapted SAM frameworks, and different YOLO generations, limited information is available regarding the direct performance of SAM 3 and YOLO26 under the same photovoltaic segmentation protocol. In particular, it is necessary to determine whether the broader visual knowledge and prompt driven operation of SAM 3 provide measurable advantages over trained YOLO26 models in terms of detection coverage and mask quality, or whether the compact and specialized CNN approach remains more suitable for this application.

For this reason, this work presents a benchmark comparison of SAM 3, YOLO26s, and YOLO26l for photovoltaic panel instance segmentation in aerial images. All models are assessed using the same ground truth annotations and a unified evaluation procedure. The principal contributions of this work can be summarized as follows:

- A direct comparison between a concept driven foundation model and two CNN based YOLO26 configurations for photovoltaic panel instance segmentation.
- A common evaluation framework based on average precision, precision, recall, F1 score, Intersection over Union, false positives, and missed photovoltaic instances.
- An analysis of the differences between the models in terms of mask delineation, detection coverage, confidence threshold selection, and suitability for photovoltaic inspection.
- An extension of previous photovoltaic panel detection research from YOLO bounding box evaluation toward pixel level instance segmentation using recent computer vision architectures.

The rest of this paper is organized as follows. Section II presents the dataset, image annotations, and experimental configuration. Section III describes SAM 3, YOLO26s, YOLO26l, and the unified evaluation procedure. Section IV presents and discusses the quantitative results. Finally, Section V provides the conclusions and recommendations for future work.

II. DATASET

The dataset used in this study extends the image collection introduced in the previous conference paper, where photovoltaic panels were captured at two installations with different technical and environmental characteristics [1]. The original collection included a 1 MW ground mounted photovoltaic plant located in a coastal area and a 300 kW rooftop installation located in an urban environment. These two sites provided variations in panel arrangement, installation type, image scale, surrounding vegetation, roof structures, and background conditions. The first site was captured using a DJI Phantom 4 equipped with a 12 MP camera, while the second site was captured using a DJI Mavic 3E equipped with a 48 MP camera.

For the present work, a third acquisition site was incorporated to increase the visual and structural variability of the dataset. The new site corresponds to a large industrial area composed of several buildings with photovoltaic arrays installed on different sections of their rooftops. In contrast with the two previous locations, this site contains multiple roof materials, large warehouse structures, roads, parked vehicles, open areas, vegetation, and buildings without photovoltaic panels. The photovoltaic modules also appear at different positions and scales throughout the orthomosaic. Some arrays occupy a considerable portion of the roof, while others are distributed in narrower rows or separated into independent groups. These characteristics provide a more complex evaluation environment than a single ground mounted plant or an isolated rooftop installation.



Fig. 1. Orthomosaic of the new industrial rooftop photovoltaic site incorporated into the extended dataset.

Figure 1 illustrates the principal sources of variability introduced by the new site, including different rooftop dimensions, variations in panel density, multiple array orientations, nonuniform backgrounds, road infrastructure, parked vehicles, vegetation, and areas without photovoltaic modules. The presence of several buildings within the same image also creates a more demanding segmentation scenario because the models must distinguish photovoltaic panels from roof textures, skylights, shadows, and other rectangular structures with similar visual characteristics.

The extended image repository contains more than 2,000 aerial images collected from the three photovoltaic installations. The repository includes individual UAV frames, image crops derived from large scenes, and orthomosaic regions representing wider areas of each site. Rather than increasing the dataset using repeated views of a single installation, the new collection was designed to incorporate meaningful differences in geographic context, installation configuration, camera characteristics, ground sampling distance, illumination, roof appearance, and photovoltaic array distribution. This approach is intended to reduce the dependence of the models on visual patterns associated with only one location.

The complete repository was used to support data preparation, annotation, model development, and fine tuning. However, only 117 images were selected for the generalization experiment presented in this work. This reduced subset was not intended to represent the complete size of the available dataset. Instead, it was selected to create a controlled evaluation set containing representative variations from the expanded image collection. The selected images include panels observed at different scales, arrangements, illumination conditions, and background complexities.

III. MODEL TRAINING PROCESS

Three instance segmentation models were considered in the experimental comparison: YOLO26s-seg, YOLO26l-seg, and Segment Anything Model 3 (SAM 3). The YOLO26 models are convolutional neural network architectures fine-tuned for a fixed photovoltaic panel class, whereas SAM 3 is a foundation model applied through prompt-based concept segmentation. These configurations represent two distinct computer vision paradigms. Recent photovoltaic studies have extended supervised CNN pipelines toward YOLO-based semantic segmentation [11], while prompt-guided foundation-model applications have been explored for photovoltaic assessment [12]. Other recent photovoltaic computer-vision investigations also demonstrate the continued expansion of deep learning methods in this domain [13]. Consequently, the benchmark compares supervised task specialization with zero-shot inference using a general-purpose pretrained model. The evaluation workflow was designed to use fixed operating thresholds and the same final ground-truth masks for all three models.

The 117-entry dataset was assembled specifically for the YOLO26s and YOLO26l fine-tuning process and contained 114 training entries and three internal-validation entries. This dataset construction follows the progression of aerial photovoltaic mapping research, where scene scale, image source, and annotation granularity have been shown to influence localization and area estimation [13], [14], [15]. Both YOLO26 models were initialized from pretrained segmentation checkpoints and configured for 200 epochs at an input resolution of 1024 x 1024 pixels. Early stopping was set to 60 epochs without improvement. Recent UAV-based comparisons have similarly emphasized the importance of evaluating different architectures using a common dataset and validation strategy [12].

TABLE I. EXPERIMENTAL DATA PARTITIONS

Parameter	YOLO26s-seg	YOLO26l-seg
Initial checkpoint	yolo26s-seg.pt	yolo26l-seg.pt
Pretrained	Yes	Yes
Training / validation	114 / 3 entries	114 / 3 images
Input size	1024 x 1024	1024 x 1024
Configured epochs	200	200
Early stopping	patience = 60	patience = 60

The selected SAM 3 architecture used zero-shot inference with a consistent text prompt for solar-panel identification. The model image size was 1008 pixels, and the inference wrapper resized the longest image side to 1008 pixels. The visual encoder is a 32-layer vision transformer, and the detector uses 200 object queries. Post-processing adjustments were applied to improve mask quality without fine-tuning the model.

TABLE II. SAM 3 ZERO-SHOT INFERENCE CONFIGURATION

Parameter	SAM 3 configuration
Architecture / model type	Sam3VideoModel / sam3_video
Experimental use	Independent still-image inference; tracking not used
Fine-tuning	None
Parameter	**SAM 3 configuration**
Operating mode	Prompt-based zero-shot inference
Prompt	Text: solar panel; identical for every image
Vision input configuration	Model image size 1008; wrapper maximum side 1008
Vision backbone	ViT, 32 layers, hidden size 1024, patch size 14
Text encoder	CLIP text model, 24 layers, hidden size 1024
Detector structure	6-layer encoder, 6-layer decoder, 200 object queries
Configured precision	FP32; runtime autocast not confirmed
Mask / AP pipeline settings	Mask threshold 0.40; AP population confidence 0.001
Postprocessing information	Hole-filling area 16 pixels; other wrapper operations described generally
Detection limits	External max detections 300; 200 detector queries per forward pass
Implementation	/app/ai/sam3 in Heliogrid; Transformers 5.0.0.dev0
Checkpoint identifier	Not encoded in supplied config (_name_or_path empty)

Model-specific confidence thresholds were selected before final testing using a separate 16-image threshold-validation set containing 64 annotated instances. For each model, a confidence sweep was performed, and the operating threshold was frozen at the maximum of the smoothed instance F1confidence curve at a mask IoU of 0.50.

The final comparison used 16 aerial images containing 49 annotated photovoltaic panel instances. Average Precision was calculated with the python COCO segmentation evaluator. AP50 and AP75 correspond to mask IoU thresholds of 0.50 and 0.75, while mAP50-95 is the mean over thresholds from 0.50 to 0.95 in increments of 0.05 using the standard 101-point COCO interpolation. The COCO evaluator was configured for a maximum of 100 detections per image. Prediction masks were evaluated at the original ground-truth image dimensions. Comparable UAV benchmarking studies have also used precision, recall, F1, IoU, and mAP to quantify architecture tradeoffs under shared test conditions [12].

TABLE III. THRESHOLD SELECTION AND UNIFIED EVALUATION

Evaluation item	Procedure
Task	Instance segmentation
Threshold-validation set	16 images, 64 instances
Final evaluation set	16 images, 49 instances
Threshold rule	Maximum F1 at mask IoU = 0.50
Frozen confidence	SAM 3: 0.6500; s: 0.150150; l: 0.060060
AP implementation	pycocotools COCO segmentation
AP50 / AP75	Mask AP at IoU 0.50 / 0.75
mAP50-95	Mean AP, IoU 0.50:0.05:0.95
COCO interpolation	Standard 101-point
COCO max detections	100 per image
Operating-point counts	Pooled TP, FP, and FN at IoU 0.50
Matched IoU / Dice	Mean over matched true-positive pairs only

Precision, recall, and F1 were calculated from pooled instance counts at the frozen operating thresholds. False positives and false negatives per image were normalized by the number of evaluation images, and the omission rate was calculated as one minus recall. Matched IoU and matched Dice summarize spatial agreement only for true-positive pairs and must therefore be interpreted together with recall and falsenegative counts. The evaluator was used consistently for all three models; lower-level implementation details such as tie handling and duplicate assignment follow the default behaviour of the pycocotools evaluator.

All experiments were executed on an NVIDIA RTX A2000 with approximately 11.62 GiB of available memory. The YOLO environment used Python 3.12.3, PyTorch 2.11.0 with CUDA 12.8, and Ultralytics 8.4.76. The SAM 3 environment used Python 3.12, PyTorch 2.9.1 with CUDA 12.6, NumPy 1.26.4, and a development build of Transformers 5.0.0. For SAM 3, the threshold-validation execution required approximately 467.7 ms per image, corresponding to 2.14 frames per second, with 6.78 GB peak GPU memory. The YOLO26s model inference process took 21.2 ms per image for inference.

IV. RESULTS

The evaluation results generated were consolidated using the process presented in the methodology. The final test set contained 16 aerial images and 49 annotated photovoltaic panel instances. Each model was evaluated using the confidence threshold previously selected on the external validation set, without modifying the operating point after observing the test results.

As shown in Table IV, SAM 3 achieved the highest overall segmentation performance. Its AP50 reached 0.9645 and remained at 0.9506 under the stricter AP75 requirement. The resulting mAP50-95 of 0.8500 was almost twice the value obtained by YOLO26s and more than four times the value obtained by YOLO26l. The same ranking was observed for recall, F1 score, and matched IoU.

TABLE IV. QUANTITATIVE PERFORMANCE OF EVALUATED MODELS

Metric	SAM 3	YOLO26s	YOLO26l
Threshold	0.6500	0.1502	0.0601
AP50	0.9645	0.6079	0.5330
AP75	0.9506	0.5141	0.1071
mAP50-95	0.8500	0.4299	0.1989
Precision	0.7895	0.7931	0.7333
Recall	0.9184	0.4694	0.2245
F1 score	0.8491	0.5897	0.3438
Matched IoU	0.9291	0.8565	0.6376

Precision was nearly equivalent for SAM 3 and YOLO26s, with values of 0.7895 and 0.7931, respectively. However, SAM 3 obtained a recall of 0.9184, compared with 0.4694 for YOLO26s and 0.2245 for YOLO26l. This difference increased the F1 score of SAM 3 to 0.8491, while the two YOLO models reached 0.5897 and 0.3438.

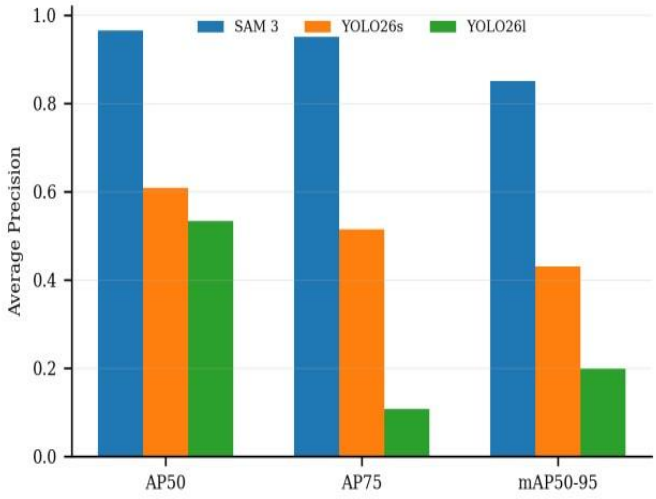


Fig. 2. Average Precision comparison under different mask IoU requirements.

Figure 2 shows that SAM 3 maintained its performance as the mask overlap requirement increased. The reduction from AP50 to AP75 was only 0.0139, corresponding to a retention of 98.6%. YOLO26s retained 84.6% of its AP50 value, whereas YOLO26l retained only 20.1%. The selected YOLO models detected some panel regions at moderate overlap but frequently failed to reproduce their contours under stricter localization requirements.

The matched IoU values support the same interpretation. The true-positive masks produced by SAM 3 obtained a mean IoU of 0.9291, compared with 0.8565 for YOLO26s and 0.6376 for YOLO26l. Although this metric considers only successfully matched instances, SAM 3 simultaneously achieved the highest overlap quality and the greatest detection coverage.

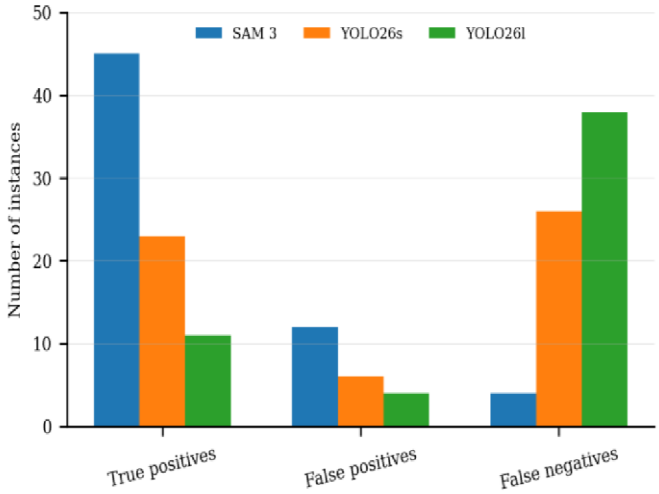


Fig. 3. True-positive, false-positive, and false-negative instance counts.

The instance counts shown in Figure 3 provide a direct interpretation of the metric differences. SAM 3 detected 45 of the 49 annotated panels and missed only four. YOLO26s detected 23 instances and missed 26, while YOLO26l detected 11 and missed 38. SAM 3 produced more false positives than the YOLO models, but its substantially lower number of omissions resulted in the lowest total error count, offering the better performance of the models compared in this work for the segmentation of PV panels. Table V summarizes these findings.

TABLE V. OPERATING-POINT INSTANCE COUNTS

Metric	SAM 3	YOLO26s	YOLO26l
True positives	45	23	11
False positives	12	6	4
False negatives	4	26	38
Predictions	57	29	15
FP/image	0.750	0.375	0.250
FN/image	0.250	1.625	2.375
Omission rate	8.16%	53.06%	77.55%

The omission rate was 8.16% for SAM 3, 53.06% for YOLO26s, and 77.55% for YOLO26l. The combined false-positive and false-negative counts were 16, 32, and 42, respectively. When normalized by the 16 test images, these values correspond to 1.00 error per image for SAM 3, 2.00 for YOLO26s, and 2.63 for YOLO26l.

The prediction-to-ground-truth ratios were 1.163 for SAM 3, 0.592 for YOLO26s, and 0.306 for YOLO26l. SAM 3 slightly overproduced candidate masks but maintained high recall. In contrast, both YOLO models underproduced instances. For photovoltaic inspection, this error profile favors SAM 3 when overlooking an installed panel is more critical than reviewing an additional candidate region.

Relative to YOLO26s, SAM 3 increased mAP50-95 by 0.4201, equivalent to a 97.7% relative improvement. Recall in-

creased by 0.4490 and F1 score by 0.2594, while precision remained effectively equivalent. These results indicate that the principal advantage of the foundation model was its capacity to recover substantially more panel instances while preserving accurate mask boundaries.

YOLO26s consistently outperformed YOLO26l. The small model more than doubled recall and increased AP75 from 0.1071 to 0.5141. It also reduced the number of false negatives from 38 to 26. Therefore, greater model size and computational complexity did not improve the evaluated photovoltaic segmentation checkpoint.

The most relevant result is that SAM 3 achieved the best segmentation without supervised fine-tuning on the photovoltaic dataset. The zero-shot foundation model demonstrated stronger transfer to the evaluated aerial scenes than the available task-specific convolutional neural network models. This finding supports the increasing use of foundation models in remote sensing and industrial inspection. Their largescale pretraining can provide visual representations that remain effective under changes in roof material, panel arrangement, illumination, scale, and surrounding objects.

Nevertheless, the result does not imply that prompt-based models will always outperform supervised CNNs. Performance may change with larger training sets, controlled hyperparameter optimization, model compression, or domain-specific finetuning. The proposed training dataset helps to compare the generalization capabilities of each model.

The accuracy advantage of SAM 3 must be considered together with its computational cost. The available validation record reports approximately 467.7 ms per image, 2.14 frames per second, and 6.78 GB of peak GPU memory. The YOLO26s validation log reported 21.2 ms per image, corresponding to approximately 47 frames per second for the inference component. Since the two measurements used different timer boundaries, they should not be treated as a strict end-to-end speed comparison. However, they illustrate the expected tradeoff between segmentation quality and throughput.

TABLE VI. DATA PARTITIONS AND INDEPENDENCE

Partition	Images	Instances	Independent from training data
Training	114	— a	Not applicable
Internal validation	3	— a	No
Threshold validation	16	64	No weight updates
Final test	16	49	Partial: development cycle

Table VI summarizes the partitions used in this study together with their independence status. Two constraints follow from it. First, the final test set contains 49 annotated instances, so a single instance corresponds to approximately 2.04 percentage points of recall; metric differences below that magnitude should not be interpreted as meaningful. In particular, the precision difference between SAM 3 and YOLO26s reported in Table IV corresponds to less than one instance. Second, the test images were involved during the broader development cycle of YOLO26s, which introduces an optimistic bias in favour of the supervised models.

It should be noted that this bias operates against the conclusion reported in this work: even under conditions favourable to YOLO26s, SAM 3 recovered 45 of the 49 annotated instances compared with 23, a difference of 44.90 percentage points in recall that cannot be attributed to the evaluation bias. A fully independent blind evaluation, using a fourth acquisition site not involved in any stage of dataset preparation, model fine-tuning, or threshold selection, together with grouped cross-validation by source photograph, is specified as the immediate next step in Section V.

Under the evaluated conditions, the operational ranking is SAM 3, YOLO26s, and YOLO26l. SAM 3 is the preferred option when minimizing missed panels and preserving detailed contours are the main objectives. YOLO26s provides the stronger alternative for low-latency or resource-constrained deployment. The evaluated YOLO26l checkpoint is not recommended because it combines low recall with weak mask localization.

Future work should also measure all models using identical preprocessing, inference, and mask-postprocessing timer boundaries.

V. CONCLUSION

This work presented a comparative evaluation of SAM 3, YOLO26s, and YOLO26l for photovoltaic panel instance segmentation in aerial imagery. The study examined two different computer vision paradigms: prompt-based foundation models and supervised convolutional neural network architectures. All models were evaluated using the same ground truth annotations, frozen confidence thresholds, and a common COCO segmentation procedure. This methodology enabled a consistent comparison of detection coverage, mask accuracy, false detections, missed instances, and overall segmentation performance.

SAM 3 achieved the strongest overall results, with an AP50 of 96.45%, an AP75 of 95.06%, a mAP50–95 of 85.00%, and a recall of 91.84%. It detected 45 of the 49 annotated photovoltaic instances and obtained a matched IoU of 92.91%, demonstrating both high detection coverage and accurate mask delineation. YOLO26s achieved a slightly higher precision of 79.31%, compared with 78.95% for SAM 3, but its recall decreased to 46.94%. YOLO26l produced the weakest performance, with a recall of 22.45%, a mAP50–95 of 19.89%, and 38 missed photovoltaic instances.

The results demonstrate that the main advantage of SAM 3 was its ability to identify substantially more photovoltaic panels while maintaining precision comparable to the best YOLO26 configuration. This performance was obtained without supervised fine tuning on the photovoltaic dataset, using the same textual prompt for all test images. YOLO26s remains a relevant alternative for applications that prioritize inference speed and reduced computational requirements, since the available execution records indicate substantially lower latency than SAM 3. The evaluated YOLO26l checkpoint did not provide an advantage over the smaller model, although this result must be interpreted considering that both YOLO models were trained under different dataset and batch conditions.

Future work will focus on evaluating the models using a fully independent blind test set containing new photovoltaic installations, acquisition platforms, geographic locations, and environmental conditions. Controlled experiments should also retrain YOLO26s and YOLO26l using identical data partitions, augmentation procedures, optimization parameters, and batch conditions to isolate the influence of model scale. Additional research should investigate supervised or parameter efficient adaptation of SAM 3, alternative textual and visual prompts, multispectral and thermal imagery, and model compression for edge deployment. A unified end to end timing protocol should also be implemented to compare preprocessing, inference, mask reconstruction, memory consumption, and energy requirements under equivalent hardware conditions.

ACKNOWLEDGMENT

This work was funded by the MESCYT of Dominican Republic by the projects FONDOCYT number 2023-1-3C3-0747 and 2023-1-3C3-0560.

REFERENCES

[1] R. Batista-Jorge et al., "Evaluating YOLOv8 and YOLOv11 Photovoltaic Panels Detection Performance in Large Scale Solar Power Plants," in 2025 International Conference on Sustainable Technology and Engineering (i-COSTE), IEEE, Dec. 2025, pp. 1–6. doi: 10.1109/iCOSTE68047.2025.11467540.

[2] J. Wang, X. Chen, W. Jiang, L. Hua, J. Liu, and H. Sui, "PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery," Int. J. Appl. Earth Obs. Geoinf., vol. 119, p. 103309, May 2023, doi: 10.1016/j.jag.2023.103309.

[3] G. Jocher, J. Qiu, M. Liu, S. Lyu, F. C. Akyon, and M. E. Kalfaoglu, "Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models," Jun. 2026, [Online]. Available: http://arxiv.org/abs/2606.03748

[4] G. García, A. Aparcedo, G. K. Nayak, T. Ahmed, M. Shah, and M. Li, "Generalized deep learning model for photovoltaic module segmentation from satellite and aerial imagery," Sol. Energy, vol. 274, p. 112539, May 2024, doi: 10.1016/j.solener.2024.112539.

[5] A. Kirillov et al., "Segment Anything," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.02643

[6] N. Ravi et al., "SAM 2: Segment Anything in Images and Videos," Oct. 2024, doi: arXiv:2408.00714.

[7] O. Rafaeli, T. Svoray, R. Blushtein-Livnon, and A. Nahlieli, "Solar Panels Segmentation in Remote Sensing Data Using Segment Anything Models 2 and 2.1," IEEE Geosci. Remote Sens. Lett., vol. 22, pp. 1–5, 2025, doi: 10.1109/LGRS.2024.3522183.

[8] X. Yao, S. Zhang, Z. Liang, J. Li, and C. Liu, "PVSAM: Adapting geometric prompts to segment anything model for photovoltaic detection in remote sensing imagery," Appl. Energy, vol. 404, p. 127137, Feb. 2026, doi: 10.1016/j.apenergy.2025.127137.

[9] H. Li et al., "A CLIP-Based Zero-Shot Photovoltaic Segmentation Framework for Remote Sensing Imagery," Remote Sens., vol. 18, no. 6, p. 865, Mar. 2026, doi: 10.3390/rs18060865.

[10] N. Carion et al., "SAM 3: Segment Anything with Concepts," Mar. 2026, [Online]. Available: http://arxiv.org/abs/2511.16719

[11] J. Zhang, D. Chen, B. Li, Z. Zhao, H. Bi, and P. Xiao, "A Yolo-Based Semantic Segmentation Model for Solar Photovoltaic Panel Identification," Sensors, vol. 26, no. 1, p. 75, Dec. 2025, doi: 10.3390/s26010075.

[12] H. İ. Şenol, A. Y. Yiğit, and M. Ulukavak, "Comparative deep learning detection of rooftop photovoltaic panels from UAV imagery," Sol. Energy, vol. 316, p. 114828, Sep. 2026, doi: 10.1016/j.solener.2026.114828.

[13] J. M. Malof, K. Bradbury, L. M. Collins, and R. G. Newell, "Automatic detection of solar photovoltaic arrays in high resolution aerial imagery," Appl. Energy, vol. 183, pp. 229–240, Dec. 2016, doi: 10.1016/j.apenergy.2016.08.191.

[14] J. Camilo, R. Wang, L. M. Collins, K. Bradbury, and J. M. Malof, "Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery," Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.04018

[15] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal, "DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States," Joule, vol. 2, no. 12, pp. 2605–2617, Dec. 2018, doi: 10.1016/j.joule.2018.11.021.