Enhancing Photovoltaic Panel Segmentation in Remote Sensing Imagery: A Comparative Study of Attention-Integrated UNet Models

Mohammed Q. Alkhatib [a], Mina Al-Saad [a], Nour Aburaed [a], Mohammed S. Zitouni [a], Saeed Almansoori [b], Hussain Al-Ahmad [a]

[a] College of Engineering and IT, University of Dubai, Dubai, UAE - (mqalkhatib, minaalsaad, nour.aburaed)@ieee.org, (mzitouni, halahmad)@ud.ac.ae

[b] Mohammed Bin Rashid Space Centre, Dubai, UAE - saeed.almansoori@mbrsc.ae

Keywords: Semantic Segmentation, UNet, Remote Sensing, Attention Module, Photovoltaic (PV) Panel

Abstract

This study explores the enhancement of UNet-based semantic segmentation for photovoltaic (PV) panels in remote sensing images by integrating attention mechanisms. Given the critical role of solar energy in achieving global sustainability, accurate PV panel detection is essential for effective energy management. Using the high-resolution PV01 dataset, which includes UAV-captured rooftop PV samples, we evaluate the impact of four distinct attention modules: Convolutional Block Attention Module (CBAM), Squeeze-and-Excitation Networks (SE-Net), Efficient Channel Attention (ECA-Net), and Coordinate Attention (CA) on segmentation performance. Comparative analysis demonstrates that UNet models with SE and CA modules substantially outperform the baseline, achieving the highest scores in Average Accuracy (AA), Average Precision (AP), Average Recall (AR), mean Intersection over Union (mIoU), and Average F1-score (AF). In particular, UNet + SE achieved an AA of 0.9809, AP of 0.9756, AR of 0.9629, AF of 0.9692, and mIoU of 0.9403, highlighting the efficacy of attention mechanisms in refining feature representations and advancing PV panel segmentation, thereby contributing to large-scale solar energy monitoring and deployment.

1. Introduction

The global demand for renewable energy has significantly increased due to growing concerns about climate change and the depletion of conventional fossil fuels. The widespread reliance on fossil fuels has led to higher greenhouse gas emissions, higher global temperatures, and a range of environmental issues. To tackle these challenges, many countries have enacted policies aimed at achieving carbon neutrality, with a focus on limiting global warming to within 2°C. Within this framework, solar energy has emerged as a sustainable and environmentally friendly power source, playing an essential role in mitigating greenhouse gas emissions and reducing dependence on fossil fuels (Zhao et al., 2024). Consequently, it is crucial to establish solar energy as the primary electricity generation source for future cities. According to UN report in 2022, transitioning to renewable energy technologies is viewed as a vital strategy for developing a clean and secure energy system that aligns with climate neutrality objectives (Pena Pereira et al., 2024).

Recent findings show that desert-climate countries in the Gulf Cooperation Council (GCC), particularly the UAE and Saudi Arabia, are among the foremost investors in solar energy in the region. With an average of 10 hours of sunlight per day and a Global Horizontal Irradiance (GHI) of 2.12 MWh/m^2/year, these countries have significant potential for large-scale solar power production.

Accurate data on the locations, types, quantities, specifications, and power capacities of solar panels is essential for effective policy-making, energy planning, and grid management. However, challenges in gathering this information often arise due to privacy concerns, reluctance from installers to share details, and incomplete datasets that frequently lack precise location information (Lekavičius and Gružauskas, 2024). Recent advancements in remote sensing technology, which offer imagery across various temporal and spatial scales, have become vital for efficiently mapping and detecting solar panels. The integration of remote sensing with Artificial Intelligence (AI) techniques helps address issues related to incomplete data and the labor-intensive nature of manually mapping photovoltaic (PV) facilities, making the mapping process both faster and more accurate (Dui et al., 2023). Nonetheless, detecting solar panels from remote sensing imagery remains challenging due to the diversity in panel shapes, sizes, colors, and the various angles at which they may be installed on rooftops.

Semantic segmentation, the task of classifying each pixel in an image according to its semantic category, is increasingly applied to identify and map PV panels in satellite and aerial imagery. This process is critical for renewable energy planning, enabling precise tracking of solar installations and potential energy outputs. Deep learning methods, particularly UNet architectures, have been adapted for segmenting PV panels, with innovations such as CrossNets that use cross-learning to enhance segmentation accuracy in residential solar panel datasets (Zhuang et al., 2020). CNNs further support PV detection in low-quality satellite images, providing valuable data for local power assessment (Golovko et al., 2017). UNet models also scale effectively for urban applications, where rooftop solar potential is quantified for large-scale energy estimates (Huang et al., 2019a). Enhanced segmentation methods, such as those with Constraint Refinement Modules (CRMs) incorporating color and shape priors, have improved IoU scores for city-wide PV panel detection (Tan et al., 2023). Additionally, SegFormer models extend applications to PV defect detection, excelling in fault classification accuracy within solar cells (Mahboob et al., 2024).

Attention mechanisms have significantly improved semantic segmentation by enhancing spatial and contextual understanding. Early works like PAN combined spatial pyramids with attention for refined pixel labeling on datasets like PASCAL VOC and Cityscapes (Li et al., 2018), while CCNet introduced criss-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-63-2025

cross attention to efficiently capture global pixel dependencies (Huang et al., 2019b). Later, EAPNet utilized multi-scale context and channel attention for effective segmentation at varying scales (Yang et al., 2021), and SegNeXt presented a convolutional alternative to self-attention, achieving high mIoU on ADE20K and Pascal VOC (Guo et al., 2022). Specialized models like CAS-Net used coordinate attention for enhanced small-object segmentation in remote sensing (Yang et al., 2023), and attention-enhanced UNet variants proved valuable in medical image segmentation by retaining spatial details and context (Fang, 2022, Huang et al., 2024). BiSeNet V3 employed edge-focused attention for real-time segmentation on Cityscapes (Tsai and Tseng, 2023), while HRNet used attention to maintain high-resolution features (Kim et al., 2023). These studies collectively highlight attention's crucial role in advancing segmentation accuracy and efficiency across diverse applications.

In this paper, we evaluate the effectiveness of various attention modules; Convolutional Block Attention Module (CBAM) (Woo et al., 2018), Squeeze-and-Excitation Networks (SE-Net) (Hu et al., 2018), Efficient channel attention (ECA-Net) (Wang et al., 2020), and Coordinate Attention (CA) (Hou et al., 2021) on segmentation performance when integrated with the UNet model. Each attention mechanism brings unique strengths and optimizations for enhancing feature representation. CBAM combines channel and spatial attention to refine feature maps with minimal computational cost, while SE-Net uses channel recalibration to significantly boost representational power, demonstrating considerable improvements in classification accuracy. ECA-Net emphasizes efficiency, achieving competitive accuracy gains with minimal parameters by avoiding dimensionality reduction, making it lightweight and computationally efficient. Coordinate Attention, designed for mobile networks, captures long-range dependencies with embedded positional information, making it particularly effective for spatially complex tasks. Our comparisons focus on identifying which module best enhances UNet's performance for segmenting semantic features, specifically analyzing their trade-offs between accuracy and computational efficiency.

The remainder of the paper is structured as follows: Section 2 presents the literature review of semantic segmentation based on DL approaches in the field of remote sensing, Section 3 explains the methodology , attention mechanism , and dataset used in this research work, Section 4 illustrates and discusses the results, and finally, Section 5 summarizes and concludes the paper.

2. Literature Review

Several studies in the literature have focused on segmenting solar panels from remote sensing images, showcasing substantial progress in this field. Motivated by the need for accurate mapping and monitoring of PV installations for efficient energy management and urban planning. For example, Bouaziz et al. (Bouaziz et al., 2024) developed a UNet-based model to enhance PV panel detection in high-resolution satellite and aerial imagery from Google Earth Pro, focusing on a case study in Sfax, Tunisia. The model incorporates data enhancement techniques, including zoom-in, zoom-out, and blur adjustments, to improve adaptability across different zoom levels. The experiments show robust segmentation results with an IoU score of 86%, even with limited labeled data.Their findings suggest that solar panels are predominantly installed on high-income residences, emphasizing the potential for targeted renewable energy initiatives to support lower-income households. Similarly, authors in (Zhao et al., 2024) introduced a variant of the UNet model, called PV-UNet, which is optimized for detecting solar panels from multisource remote sensing data. By integrating attention and feature fusion modules, PV-UNet effectively addresses both spatial and spectral inconsistencies in images from different sensors. The model achieved an F1-score of 98.04% and IoU of 96.15%, showcasing its effectiveness in adapting to diverse remote sensing data sources, making it a valuable tool for large-scale solar panel mapping. Furthermore, the datasets available for solar panel segmentation are scarce. Moreover, annotating these datasets is often labor-intensive, time-consuming and prone to errors. To overcome these challenges, Lekavičius and Gružauskas (Lekavičius and Gružauskas, 2024) utilized the pix2pix generative adversarial network (GAN) to generate supplementary remote sensing (RS) data. This method enhances the original training data, which varies in ground sampling distances (GSDs), while preserving its integrity. Their experiments combined the DeepLabV3 model with a ResNet-50 backbone and the pix2pix GAN architecture to evaluate the effectiveness of GAN-based data augmentation in improving RS imagery segmentation accuracy. The proposed semantic segmentation model utilized transfer learning and incorporated 60% GAN-generated RS imagery as additional training data. Their findings indicated that GAN-generated images effectively addressed the limitations of existing datasets, enhancing overall IoU and F1 metrics by 2% and 1.46%, respectively, compared to conventional augmentation techniques.

3. Proposed Methodology

3.1 Network Architecture

The overall framework is depicted in Figure 1. The model is a UNet-based architecture for image segmentation, with attention blocks integrated rtyu to enhance feature focus. It begins by processing a 256×256 input RGB image through convolutional layers and downsampling with max pooling to capture features at different resolutions. Attention blocks are placed after each downsampling layer and in the upsampling path, helping the model retain crucial spatial information by focusing on relevant features during both encoding and decoding. details of the model and the attention block are described in the subsections below.

3.2 UNet Baseline Model

UNet architecture introduced by Olaf Ronneberger (Ronneberger et al., 2015), has become fundamental in the field of image segmentation. Known for its effectiveness in accurately segmenting structures in biomedical images, UNet has also found extensive applications across various computer vision tasks including remote sensing segmentation. The UNet architecture is distinguished by its U-shaped structure, which consists of a contracting path followed by an expansive path. The contracting path captures contextual information through convolutional and pooling layers, while the expansive path enables precise localization using transposed convolutions. Skip connections bridge the contracting and expansive paths, facilitating the transfer of high-resolution features to the final output.This allows the model to restore spatial information lost during downsampling, leading to more accurate segmentation (Liu et al., 2020).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-63-2025 | © Author(s) 2025. CC BY 4.0 License.

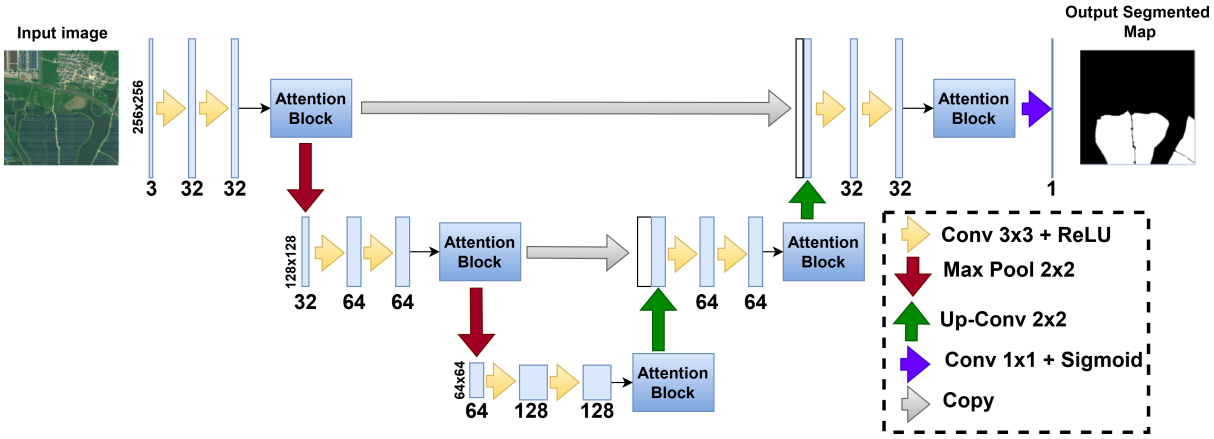
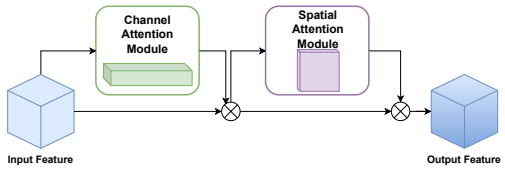
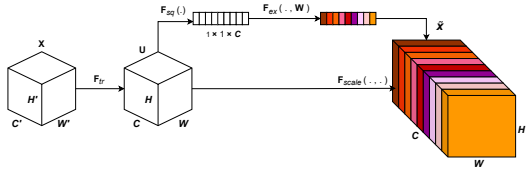


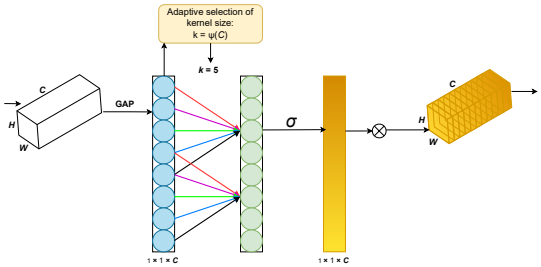
Figure 1. Overall Architecture of the Model



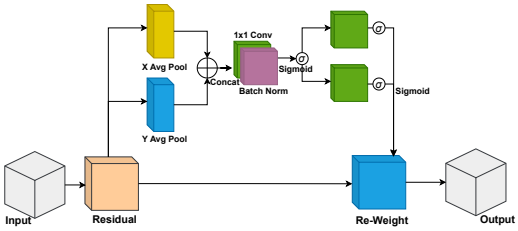
(a) CBAM (Woo et al., 2018).



(b) SE (Hu et al., 2018).



(c) ECA (Wang et al., 2020).



(d) CA (Hou et al., 2021).

Figure 2. Different Attention Modules used in this study, (a) CBAM; (b) SE; (c) ECA and (d) CA.

3.3 Attention Block

As previously discussed, this study uses four distinct attention mechanisms: CBAM, SE-Net, ECA-Net, and CA. Detailed descriptions of each module are provided below.

3.3.1 Convolutional Block Attention Module (CBAM): The CBAM (Woo et al., 2018) is an efficient attention mechanism that enhances convolutional neural networks (CNNs) by refining feature maps through channel and spatial attention sequentially. Given an input feature map $F \in \mathbb{R}^{H \times W \times C}$, CBAM first applies **Channel Attention** to emphasize meaningful channels, followed by **Spatial Attention** to highlight important spatial locations. This sequential approach enables CBAM to focus on “what” (channels) and “where” (spatial locations) to refine, providing adaptive feature refinement. The block diagram of CBAM is shown in Figure 2(a).

The **Channel Attention Module** captures inter-channel dependencies by applying both average-pooling and max-pooling along the spatial dimensions to generate two distinct descriptors, which are passed through a shared multi-layer perceptron (MLP) with a reduction ratio to balance computational efficiency and accuracy. The attention map is computed as:

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \quad (1)$$

where σ denotes the sigmoid function, aggregating pooled features through summation.

Following this, the **Spatial Attention Module** focuses on informative spatial locations. It applies average-pooling and max-pooling along the channel axis to create spatial descriptors, which are concatenated and convolved with a 7×7 kernel to form the spatial attention map:

$$M_s(F) = \sigma(f_{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \quad (2)$$

This attention map captures spatial dependencies effectively, complementing the channel-focused refinement.

3.3.2 Squeeze and Excitation (SE): The Squeeze and Excitation (SE) (Hu et al., 2018) introduces the SE block, an architectural unit designed to enhance the representational power of convolutional neural networks (CNNs) by adaptively recalibrating channel-wise feature responses. This method directly captures channel interdependencies through a two-step process called Squeeze and Excitation. By leveraging global information, the SE block adaptively highlights the most informative features, as shown in Figure 2(b).
In the Squeeze operation, global average pooling aggregates the

spatial dimensions $H \times W$ of a feature map $X \in \mathbb{R}^{H \times W \times C}$:

$$z_i = \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} x_{ijk} \tag{3}$$

where X denote the input feature maps with dimensions $B \times H \times W \times C$, where B is the batch size, H is the height, W is the width, and C is the number of channels. While the excitation process can be expressed as:

$$s_i = \sigma(W_2 \delta(W_1 z_i)) \tag{4}$$

where δ is the ReLU activation function, W_1 and W_2 are learnable weights, and σ is the sigmoid activation function (Hu et al., 2018).

3.3.3 Efficient Channel Attention (ECA-Net): The Efficient Channel Attention (ECA) (Wang et al., 2020) Network introduces an improved channel attention mechanism that simplifies complexity while maintaining effectiveness for convolutional neural networks (CNNs). Unlike previous methods that increase model parameters, ECA achieves efficient attention by avoiding dimensionality reduction and focusing on local cross-channel interactions. This is accomplished through a 1D convolution with an adaptively chosen kernel size based on the channel dimension C, as illustrated in Figure 2(c).

Channel Attention Mechanism: To compute the channel weights, global average pooling (GAP) is applied to aggregate features:

$$y = \text{GAP}(F) \tag{5}$$

where $F \in \mathbb{R}^{C \times H \times W}$ represents the input feature map. ECA then uses 1D convolution without dimensionality reduction to capture cross-channel dependencies:

$$\omega = \sigma(\text{Conv1D}_k(y)) \tag{6}$$

where σ denotes the sigmoid function, and the kernel size k is adaptively determined by:

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + b \right|_{\text{odd}} \tag{7}$$

Here, $\psi(C)$ is a non-linear mapping, ensuring k is an odd integer proportional to the channel dimension C, with γ and b as parameters.

3.3.4 Coordinate Attention (CA): The Coordinate Attention (CA) (Hou et al., 2021) mechanism is an innovative attention module designed to enhance mobile networks by embedding positional information within channel attention. Unlike traditional channel attention that relies on global pooling, CA factorizes attention into two 1D feature encoding processes, effectively aggregating features along two spatial directions to retain precise positional information. This allows CA to capture both channel relationships and long-range dependencies across spatial dimensions, as shown in Figure 2(d).

To generate the attention map, CA first applies 1D global pooling to the feature map $X \in \mathbb{R}^{C \times H \times W}$ along the horizontal and vertical axes separately, producing two direction-aware feature descriptors:

$$z_h(c, h) = \frac{1}{W} \sum_{i=1}^{W} x_c(h, i) \tag{8}$$

Table 1. Number of parameters of each model used in this study.

Model	Parameters
UNET	466,529
UNET + SE	473,545
UNET + ECA	466,546
UNET + CA	475,049
UNET + CBAM	474,035

$$z_w(c, w) = \frac{1}{H} \sum_{j=1}^{H} x_c(j, w) \tag{9}$$

where z_h and z_w represent the aggregated features along the height and width, respectively.

The resulting feature maps are concatenated and passed through a shared 1×1 convolution, non-linearity, and a split operation to produce two separate attention maps. These maps are computed as:

$$f = \delta(F_1([z_h, z_w])) \tag{10}$$

$$g_h = \sigma(F_h(f_h)), \quad g_w = \sigma(F_w(f_w)) \tag{11}$$

where σ represents the sigmoid function, F_1 is the shared 1×1 convolution, and f_h and f_w are the split feature maps.

Finally, the input feature map is recalibrated by element-wise multiplication with both attention maps:

$$y_c(i, j) = x_c(i, j) \cdot g_h(i) \cdot g_w(j) \tag{12}$$

The attention mechanisms explored in this study—namely CBAM, SE, ECA, and CA—serve as specialized modules that enhance the model's ability to focus on salient features, thereby improving feature representation within the U-Net architecture. Each attention block is integrated individually to assess its unique contribution to model performance. In addition, to facilitate a comparison of complexity, Table 1 outlines the parameter count for each model configuration, revealing the relative computational impact of adding these attention mechanisms.

The addition of SE, CA, and CBAM attention mechanisms to U-Net increases model complexity, with CA introducing the most substantial parameter growth, indicating a higher capacity for detailed feature extraction. CBAM and SE also add moderate complexity, balancing parameter increase with refined attention capabilities. In contrast, ECA stands out for its efficiency, minimally raising the parameter count and thus offering a lightweight option for resource-constrained applications. This parameter comparison highlights a trade-off between computational cost and the potential for enhanced feature representation, allowing for strategic selection of an attention mechanism based on specific application requirements.

During model training, a separate instance is created for each attention mechanism, and the model undergoes training with each attention block applied independently. This sequential evaluation allows for direct comparison of their effectiveness under identical training and testing conditions. Following the completion of training for each attention configuration, results are recorded, highlighting the impact of each attention mechanism on the model's performance. This comparative analysis aids in identifying the most suitable attention block for optimizing the model across various tasks.

4. Experiments

4.1 Dataset

The dataset comprises PV samples collected from satellite and aerial imagery, organized into three groups based on spatial resolution: 0.8m, 0.3m, and 0.1m. The 0.8m resolution dataset (PV08) includes rooftop and ground PV samples derived from Gaofen-2 and Beijing-2 satellite imagery. The 0.3m dataset (PV03), captured via aerial photography, contains ground PV samples further classified into five categories based on background land use: shrubland, grassland, cropland, saline-alkali land, and water surfaces. Lastly, the 0.1m dataset (PV01), sourced from UAV orthophotos, includes rooftop PV samples categorized into three groups according to roof type: flat concrete, steel tile, and brick. This research specifically utilizes the PV01 group for the analysis. Further information about the dataset is provided in (Jiang et al., 2021). The dataset consists of 645 images, each accompanied by a corresponding segmentation mask, with all images and masks of size 256×256. The dataset is divided into three subsets: 70% for training the model, 15% for validating its performance during training, and the remaining 15% for evaluating the model's performance after training.

4.2 Experimental Configuration

To guarantee a fair and objective comparison, all networks were trained and evaluated within the same controlled environment, utilizing TensorFlow library in Python. Consistent training parameters were applied across all models to ensure uniformity, thereby reducing the risk of confounding variables or biases in the comparisons. In the optimization process, the well-known Adam algorithm was utilized as the optimization function, along with binary cross-entropy as the selected loss function. The learning rate was empirically set to 10^{-3}, and each network was trained for 100 epochs. Moreover, to reduce training time and mitigate potential overfitting to the training data, we applied the early stopping method with a patience of 10, halting training if validation accuracy showed no improvement over 10 consecutive epochs.

4.3 Results and Discussion

Five architectures analyzed and explored in this research; Baseline UNet, UNet+SE, UNet+ECA, UNet+CA, and UNet+CBAM were subjected to thorough training and testing processes using the carefully curated dataset outlined in Section 4.1.

To assess and evaluate the effectiveness and performance of the trained models, various objective quantitative evaluation metrics were applied. These metrics are Average Accuracy (AA), Average Precision (AP), Average Recall (AR), Average F-score (AF), and mean Intersection over Union (mIoU) as described in (Aburaed et al., 2023). Each metric provides distinct insights into the model's classification capabilities and the precision of its segmentation maps.

Table 2 summarizes performance metrics for various UNet models with different attention mechanisms. The baseline UNet achieved an AA of 0.6882, but its scores for AP, AR, mIoU, and AF were all zero. This indicates that its performance was due entirely to correctly identifying True Negatives, with no effective detection of positive samples. Similarly, UNet+ECA displayed identical results to the baseline model, suggesting that its AA is also due to a reliance on True Negative matches without positively impacting other metrics.

Table 2. Summary of each model's performance metrics in terms of AA, AP, AR, mIoU, and AF.

Model Name	AA	AP	AR	mIoU	AF
UNet	0.6882	0.0	0.0	0.0	0.0
UNet+SE	**0.9809**	**0.9756**	**0.9629**	**0.9403**	**0.9692**
UNet+ECA	0.6882	0.0	0.0	0.0	0.0
UNet+CA	0.9783	0.9728	0.9571	0.9322	0.9649
UNet+CBAM	0.8864	0.8282	0.8020	0.6876	0.8149

In contrast, UNet+SE achieved the highest performance across all metrics, with AA = 0.9809, AP = 0.9756, AR = 0.9629, mIoU = 0.9403, and AF = 0.9692, making it the top-performing model in both positive and negative class identification. UNet+CA also performed well, with AA = 0.9783, AP = 0.9728, AR = 0.9571, mIoU = 0.9322, and AF = 0.9649, showing slightly lower but competitive scores to SE. The UNet+CBAM displayed moderate results with AA = 0.8864 and comparatively lower values for AP, AR, mIoU, and AF, indicating a more balanced but less robust performance.

Overall, UNet+SE and UNet+CA demonstrated significant improvements over the baseline, effectively enhancing UNet's capability for both positive and negative class detection, while the baseline and ECA variants remained limited by focusing primarily on True Negatives, as reflected in their shared AA.

Figure 3 visually confirms the quantitative findings presented in Table 2, further validating the results. UNet+SE consistently produces segmentation maps that closely align with the Ground Truth (GT) across various samples, outperforming other models in accurately capturing object boundaries, even amidst complex shapes and cluttered backgrounds. This is especially evident in the second sample, where UNet+SE effectively delineates structures with minimal errors around object edges, unlike other networks. In samples 1 and 3, SE-UNet continues to demonstrate superior boundary precision, while sample 4 showcases its ability to maintain accuracy even with less clutter. Models like UNet with Coordinate Attention (CA) also perform well but show slightly reduced consistency in boundary definition. CBAM offers moderate performance, capturing some structural details but introducing noise in complex areas. In contrast, the baseline UNet and UNet with ECA exhibit minimal detection, reflected in the near-absence of segmentation results, which is consistent with their low quantitative scores. The figure thus reinforces the robustness of UNet+SE as indicated by the table, highlighting the transformative impact of the SE mechanism on segmentation accuracy and visual clarity.

5. Conclusion

This research examines the enhancement of a UNet-based semantic segmentation model for identifying PV panels in remote sensing images through the integration of various attention mechanisms. Using the high-resolution publicly available PV01 dataset, which includes UAV-captured rooftop PV samples, we examine the impact of four attention modules; CBAM, SE-Net, ECA-Net, and CA on segmentation performance. Experiments reveal that UNet models with SE and CA modules significantly outperform the baseline model, achieving the top quantitative scores. Notably, UNet with SE reached an AA of 0.9809 and mIoU of 0.9403, highlighting the effectiveness of attention mechanisms in improving PV panel segmentation for scalable solar energy monitoring and deployment. In future work, ablation experiments will be conducted to validate the effectiveness

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-63-2025 | © Author(s) 2025. CC BY 4.0 License.

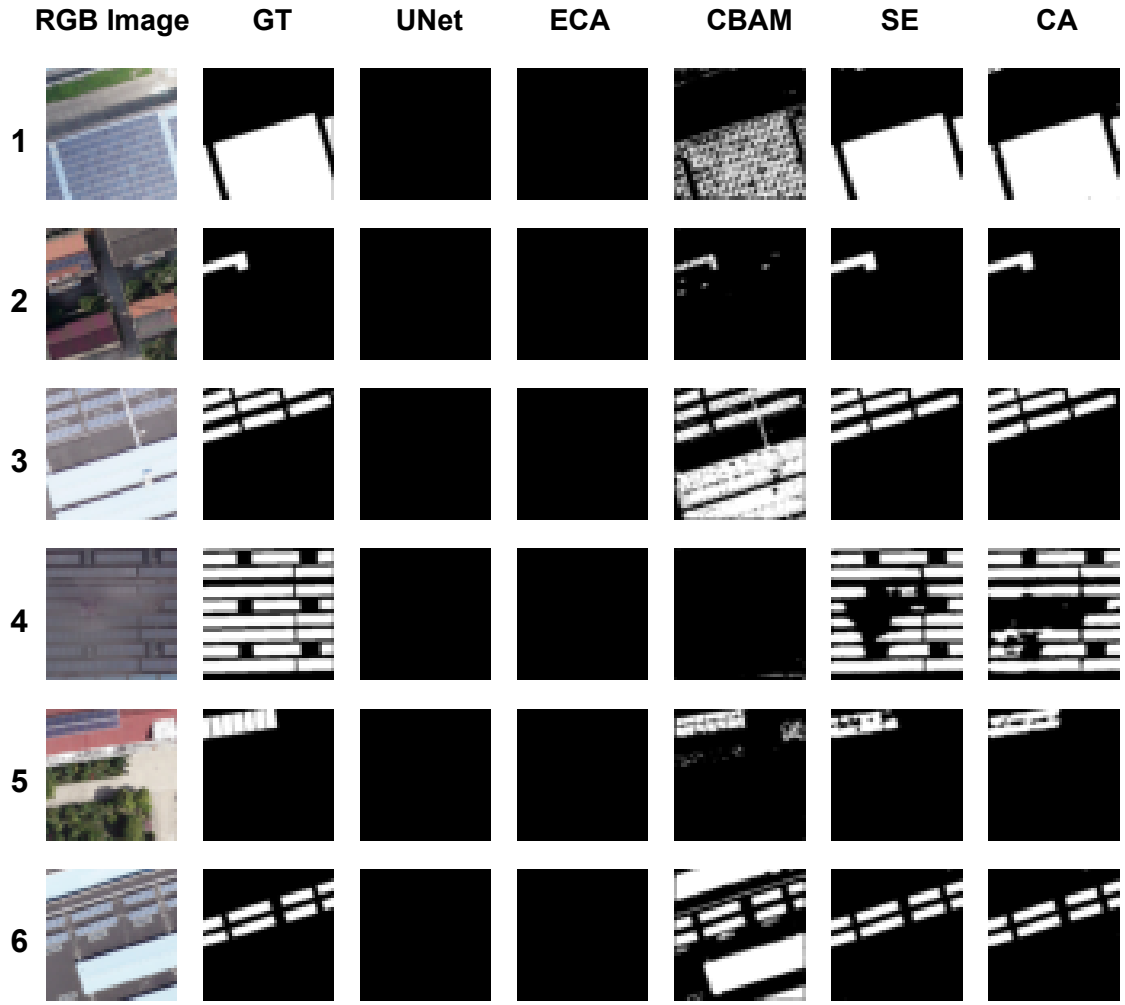


Figure 3. Visual results of segmentation maps produced by different attention blocks. Ground Truth (GT).

and generalization capability of the proposed approach, including testing the network on an additional rooftop PV dataset. The model's performance will also be evaluated using images of varying resolutions and under different environmental conditions, while exploring its potential for real-time applications.

References

Aburaed, N., Al-Saad, M., Alkhatib, M., Zitouni, M., Almansoori, S., Al-Ahmad, H., 2023. Semantic Segmentation of Remote Sensing Imagery Using AN Enhanced Encoder-Decoder Architecture. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 1015–1020.

Bouaziz, M. C., El Koundi, M., Ennine, G., 2024. High-resolution solar panel detection in Sfax, Tunisia: A UNet-Based approach. *Renewable Energy*, 235, 121171.

Dui, Z., Huang, Y., Jin, J., Gu, Q., 2023. Automatic detection of photovoltaic facilities from Sentinel-2 observations by the enhanced U-Net method. *Journal of Applied Remote Sensing*, 17(1), 014516–014516.

Fang, X., 2022. Research on the application of unet with convolutional block attention module to semantic segmentation task. *Proceedings of the 2022 5th International Conference on Sensors, Signal and Image Processing*, 13–16.

Golovko, V., Bezobrazov, S., Kroshchanka, A., Sachenko, A., Komar, M., Karachka, A., 2017. Convolutional neural network based solar photovoltaic panel detection in satellite photos. *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 1, IEEE, 14–19.

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., Hu, S.-M., 2022. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35, 1140–1156.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13713–13722.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, B., Huang, T., Xu, J., Min, J., Hu, C., Zhang, Z., 2024. RCNU-Net: Reparameterized convolutional network with convolutional block attention module for improved polyp image segmentation. *Biomedical Signal Processing and Control*, 93, 106138.

Huang, Z., Mendis, T., Xu, S., 2019a. Urban solar utilization potential mapping via deep learning technology: A case study of Wuhan, China. *Applied Energy*, 250, 283–291.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019b. Ccnet: Criss-cross attention for semantic segmentation. *Proceedings of the IEEE/CVF international conference on computer vision*, 603–612.

Jiang, H., Yao, L., Lu, N., Qin, J., Liu, T., Liu, Y., Zhou, C., 2021. Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery. *Earth System Science Data Discussions*, 2021, 1–17.

Kim, J.-S., Park, S.-W., Kim, J.-Y., Park, J., Huh, J.-H., Jung, S.-H., Sim, C.-B., 2023. E-HRNet: Enhanced semantic segmentation using squeeze and excitation. *Electronics*, 12(17), 3619.

Lekavičius, J., Gružauskas, V., 2024. Data Augmentation with Generative Adversarial Network for Solar Panel Segmentation from Remote Sensing Images. *Energies*, 17(13), 3204.

Li, H., Xiong, P., An, J., Wang, L., 2018. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*.

Liu, Z., Chen, B., Zhang, A., 2020. Building segmentation from satellite imagery using u-net with resnet encoder. *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, IEEE, 1967–1971.

Mahboob, Z., Khan, M. A., Lodhi, E., Nawaz, T., Khan, U. S., 2024. Using SegFormer for Effective Semantic Cell Segmentation for Fault Detection in Photovoltaic Arrays. *IEEE Journal of Photovoltaics*.

Pena Pereira, S., Rafiee, A., Lhermitte, S., 2024. Automated rooftop solar panel detection through Convolutional Neural Networks. *Canadian Journal of Remote Sensing*, 50(1), 2363236.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Tan, H., Guo, Z., Zhang, H., Chen, Q., Lin, Z., Chen, Y., Yan, J., 2023. Enhancing PV panel segmentation in remote sensing images with constraint refinement modules. *Applied Energy*, 350, 121757.

Tsai, T.-H., Tseng, Y.-W., 2023. BiSeNet V3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. *Neurocomputing*, 532, 33–42.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020. Eca-net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. Cbam: Convolutional block attention module. *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Yang, Q., Ku, T., Hu, K., 2021. Efficient attention pyramid network for semantic segmentation. *Ieee Access*, 9, 18867–18875.

Yang, Z., Wu, Q., Zhang, F., Zhang, X., Chen, X., Gao, Y., 2023. A New Semantic Segmentation Method for Remote Sensing Images Integrating Coordinate Attention and SPD-Conv. *Symmetry*, 15(5), 1037.

Zhao, Z., Chen, Y., Li, K., Ji, W., Sun, H., 2024. Extracting photovoltaic panels from heterogeneous remote sensing images with spatial and spectral differences. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Zhuang, L., Zhang, Z., Wang, L., 2020. The automatic segmentation of residential solar panels based on satellite images: A cross learning driven U-Net method. *Applied Soft Computing*, 92, 106283.