Solar Energy 274 (2024) 112539

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Generalized deep learning model for photovoltaic module segmentation from satellite and aerial imagery

Gustavo García [a], Alejandro Aparcedo [b], Gaurav Kumar Nayak [b,c], Tanvir Ahmed [b,c], Mubarak Shah [b,c], Mengjie Li [b,d,e,f,*]

[a] *Ana G. Méndez University, Gurabo, PR, USA*
[b] *Department of Computer Science, University of Central Florida (UCF), Orlando, FL, USA*
[c] *Center for Research in Computer Vision (CRCV), University of Central Florida (UCF), Orlando, FL, USA*
[d] *Florida Solar Energy Center (FSEC), University of Central Florida (UCF), Cocoa, FL, USA*
[e] *Resilient, Intelligent and Sustainable Energy Systems (RISES), University of Central Florida (UCF), Orlando, FL, USA*
[f] *Department of Materials Science and Engineering (MSE), University of Central Florida (UCF), Orlando, FL, USA*

ARTICLE INFO

Keywords:
Solar energy
PV panel detection
Segmentation
CNN
Mask2Former
Image processing

ABSTRACT

As solar photovoltaic (PV) has emerged as a dominant player in the energy market, there has been an exponential surge in solar deployment and investment within this sector. With the rapid growth of solar energy adoption, accurate and efficient detection of PV panels has become crucial for effective solar energy mapping and planning. This paper presents the application of the Mask2Former model for segmenting PV panels from a diverse, multi-resolution dataset of satellite and aerial imagery. Our primary objective is to harness Mask2Former's deep learning capabilities to achieve precise segmentation of PV panels in real-world scenarios. We fine-tune the pre-existing Mask2Former model on a carefully curated multi-resolution dataset and a crowdsourced dataset of satellite and aerial images, showcasing its superiority over other deep learning models like U-Net and DeepLabv3+. Most notably, Mask2Former establishes a new state-of-the-art in semantic segmentation by achieving over 95% IoU scores. Our research contributes significantly to the advancement solar energy mapping and sets a benchmark for future studies in this field.

1. Introduction

In 2023, the global installation capacity of photovoltaic(PV) systems is estimated to reach 1695 GW, marking an increase of up to about 64% from the year 2022 [1–3]. Solar energy plays a pivotal role in our pursuit of a sustainable and cleaner energy future. Accurate record-keeping of solar installation capacity and accurate estimation of energy generation from renewable energy systems, particularly PV systems, are crucial for the smooth incorporation of renewable energy into the electrical grid, assisting in strategic planning, and guiding policymakers. For utility-scale PV installations, continuous monitoring is essential due to the potential conflicts in land use, as well as considerations regarding biodiversity, ecosystem integrity and environmentally sensitive lands. Additionally, the absence of comprehensive information on small-scale rooftop PV installations poses risks to transmission system operators (TSOs), placing extra strain on the electrical grid due to unaccounted power generation. Policymakers must strategically encourage the adoption of renewable energy, aligning with both economic growth and environmental objectives, while being sensitive to the impacts associated with the expansion of renewable energy infrastructures. Traditionall, assessing the extent of solar deployment has been a manual, time-consuming, and labor-intensive task. Furthermore, the existing data often falls short in terms of geospatial accuracy and risks becoming outdated due to the fast growing PV installation. This highlights the critical need for regular and systematic data collection, as well as the development of a more efficient method for data acquisition. The accurate identification of the solar panels in satellite and aerial images offers a valuable opportunity to automate and streamline this process.

Semantic segmentation [4] is a state-of-the-art technique in computer vision that plays a crucial role in understanding and interpreting visual data. It entails the pixel-wise classification of objects within an image, assigning each pixel to a specific category or class, such as identifying roads, buildings, pedestrians, or trees. Traditional segmentation methods often struggle with the challenges posed by multi-resolution imagery, where PV panels exhibit diverse appearances, orientations,

[*] Corresponding author.
E-mail addresses: ggarcia207@email.uagm.edu (G. García), aaparcedo@ucf.edu (A. Aparcedo), gauravkumar.nayak@ucf.edu (G.K. Nayak), tanvir.ahmed@ucf.edu (T. Ahmed), shah@crcv.ucf.edu (M. Shah), mengjie.li@ucf.edu (M. Li).

https://doi.org/10.1016/j.solener.2024.112539
Received 15 January 2024; Received in revised form 4 March 2024; Accepted 11 April 2024
Available online 2 May 2024

and sizes as a result of different environmental conditions and imaging sensors. Prior research on PV detection has extensively explored the use of convolutional neural network (CNN) based models [5–8]. This paper presents a transformer-based model that seeks to advance PV panel segmentation and setting a new benchmark in the field. By leveraging the advanced transformer-based architecture of the Mask2Former model [9], and applying it to a well curated multi-resolution, crowsourced dataset [10], we conduct a comprehensive analysis. This study compares the performance of the transformer-based model with two popular CNN-based models across various contexts: high-resolution aerial images, low-resolution satellite images, rooftop PV installations, and utility-scale PV installations in China and France.

The central challenge tackled in this research is the precise segmentation of PV panels in multi-resolution satellite and aerial imagery sourced from diverse regions and installations. Current approaches typically rely on CNN models for this purpose. However, such models may encounter difficulties in capturing intricate spatial relationships, resulting in suboptimal segmentation outcomes. Furthermore, CNNs may struggle to adapt to variations in PV panel appearance and size across multi-resolution imagery and across different geographocal regions. To overcome these limitations, we propose the utilization of the Mask2Former model, a cutting-edge transformer-based universal segmentation architecture. This approach aims to enhance the accuracy of PV panel segmentation while maintaining robustness across a variety of multi-resolution imagery scenarios.

Our research contributes to the field of PV segmentation in several ways:

- **Introduction of Mask2Former Model**: We introduce the Mask2Former model for PV panel segmentation in multi-resolution imagery, pushing the boundaries of solar energy mapping.
- **Diverse Multi-Resolution Dataset**: We leverage a diverse multi-resolution dataset for PV panel segmentation, sourced from satellite and aerial imagery [6]. This dataset including a total of 24,705 images, enables comprehensive evaluations of our method. Additionally, we challenge our model with a crowdsourced dataset of aerial images that predominantly featuring rooftop photovoltaic (PV) installations [10], achieving new state-of-the-art scores in supervised learning.
- **Comparative Analysis**: Through comparative analysis with established segmentation methods, including U-Net [11] and DeepLabv3+ [12], we demonstrate the superior performance of our model across a range of evaluation metrics.
- **Parameter Exploration and Future Directions**: We conduct a thorough investigation of Mask2Former's performance across various parameters, including the number of queries, loss functions, data augmentation strategies and post-processing recommendations.

This paper is organized as follows: Section 2 provides an extensive overview of previous research related to PV panel segmentation and deep learning-based approaches. We discuss the existing methods and their limitations, laying the foundation for our proposed approach. Section 3 details the methodology employed in our research. We introduce the three models utilized in our experiments and elaborate on the datasets employed for evaluation. In Section 4, we present the results of our experiments and provide an in-depth analysis of the performance of the our proposed method. We discuss the findings, compare them with existing approaches and highlight the strengths and limitations of our model. Finally, in Section 5, we conclude the paper by summarizing our findings and contributions. We also outline potential avenues for future research in the field of PV panel segmentation, offering insights into further advancements.

2. Related works

- Solar Farm Segmentation.
  In [5], the authors proposed a machine learning framework for solar farm detection and capacity estimation. The study achieved competitive performance with high 96.87% accuracy and 95.5% Jaccard Index scores. However, it primarily focused on the application of U-Net on large-scale PV installation detections through satellite imagery, lacking the versatility needed for diverse scenarios.
- CNN segmentation of multi-resolution dataset.
  The authors of [6] introduced a multi-resolution dataset for PV panel segmentation from satellite and aerial imagery, achieving an average IoU of over 85%. Our research builds upon the dataset and introduces the Mask2Former model, surpassing previous deep learning models and setting new state-of-the-art scores for semantic segmentation.
- Solar park detection from satellite imagery.
  In [13], authors proposed an object-based random forest classification approach to identify solar parks in satellite imagery. Their methodology involved using Sentinel-2 imagery, segmenting the imagery into homogeneous objects, and extracting features for training Random Forest models. The approach achieved an accuracy of 99.97%, demonstrating its suitability for transfer learning and detection of solar parks in new study areas. However, a notable limitation of the work is its reliance on a single period of radar back-scatter properties for the best-performing model, which may restrict its adaptability to varying conditions. In contrast, our research does not depend on radar back-scatter properties and instead focuses on transformer-based PV module segmentation from a diverse set of satellite and aerial imagery.
- HyperionSolarNet Detection from Aerial Images.
  In HyperionSolarNet [14], the authors employed deep learning techniques for automated solar panel detection. Their approach utilized a two-branch model combining an image classifier and semantic segmentation, achieving commendable results with a classification accuracy of 96% and an IoU score of 0.82 for segmentation. Despite its successes, HyperionSolarNet has some limitations, including the use of the two-branch model instead of a single segmentation model. Additionally, it relied on a relatively small dataset consisting of 1963 satellite images. In contrast, out research leverages the single Mask2Former model for accurate segmentation of PV panels. We test our model performance on a diverse, multi-resolution dataset and conduct training and testing with large datasets, encompassing a total of 24,705 images.

3. Methodology

3.1. Datasets

Dataset 1: Large-scale, Distributed and Rooftop PV Installations in Jiangsu, China

The first dataset used in this study comprises aerial and satellite images collected from Jiangsu Province, China, covering an extensive area of 107,200 km^2 [6]. In this region, government policies have actively promoted the use of PV, resulting in widespread installation in areas with minimal land competition. These installations are found in diverse locations, such as sparse shrubs, low-density grasslands, reservoirs, ponds, saline-alkali lands, and rooftops.

Subsets within Dataset 1:

- PV01: This subset contains rooftop PV installations, collected using unmanned aerial vehicle (UAV) images with a ground sampling distance (GSD) of 0.1 m.
- PV03: Aerial imagery with a GSD of 0.3 m, specifically selected for the purpose of identifying distributed ground-mounted PV installations.

Table 1
Overview of the datasets used in our study.

Dataset	Image type	Installation type	Spatial resolution	No. of images	Image size
PV01	Aerial Image	Rooftop	0.1 m	645	256 × 256
PV03	Aerial Image	Distributed Ground	0.3 m	2308	1024 × 1024
PV08	Satellite Image	Large-scale	0.8 m	763	1024 × 1024
Google	Satellite Image	Rooftop	0.1 m	13 303	400 × 400
IGN	Aerial Image	Rooftop	0.2 m	7686	400 × 400

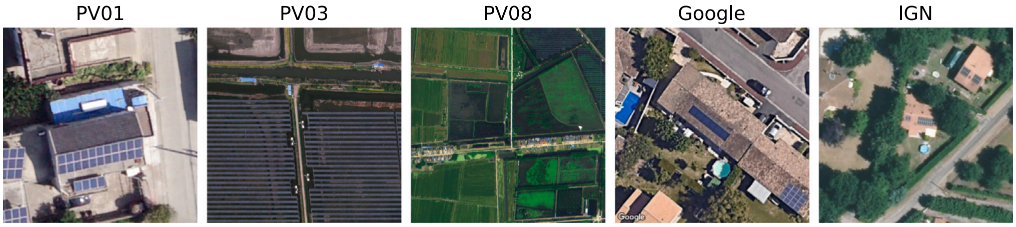


Fig. 1. Preview of the evaluated datasets. From left to right: PV01, PV03, PV08, Google Earth, IGN.

- PV08: PV08 consists of large-scale PV samples extracted from Gaofen-2 and Beijing-2 satellite images. These images feature GSDs of 0.81 m and 0.80 m in panchromatic bands.

Dataset 2: Rooftop PV Installations in France

The second dataset used in this work comprises aerial and satellite images sourced from two image providers: Google Earth Engine (GEE) and French national institute of geographical and forestry information (IGN). This dataset primarily features images of small-scale photovoltaic (PV) installations, particularly rooftop PV systems in France. [10].

Subsets within Dataset 2:

- Google: The dataset includes 13,303 images from Google Earth Engin.
- IGN: A total of 7,685 images are obtained from the French National Institute of Geographical and Forestry Information (IGN).

These images are also accompanied with extensive metadata, such as geolocation information, and energy production data. The database primarily covers information collected from individual system owners, predominantly in France and Western Europe.

The details about the different data sources for the study areas are summarized in Table 1. Examples of each dataset are illustrated in Fig. 1.

3.2. Proposed framework

This work focuses on the application of transfer learning and finetuning with three distinct models, each featuring its corresponding backbones, to segment PV panels in satellite and aerialimages. Fig. 2 describes the proposed framework, including Input, Data Preprocessing, Training and Validation, Testing and Postprocessing stages. We employed MMSegmentation [15], an open-source semantic segmentation toolkit based on PyTorch. Table 2 provides detailed information on the experiment configurations. Table 3 lists the hyperparameters used in the experiments.

3.2.1. Input

In this supervised learning experiment, the models are trained using the original RGB image and the PV panel label images as input.

3.2.2. Preprocessing

To begin, the PV area labels undergo conversion into a semantic segmentation map. This conversion involves assigning a custom color palette, where white is foreground (PV panel) and black is background. This step effectively associates each pixel in the image with either the PV class (white) or the background class (black). Subsequently, the images and their labels are partitioned randomly into the training, validation, and testing sets. Following the preprocessing approach as outlined by Jiang et al. [6], samples from each subset (e.g., PV01, PV03, PV08), are divided into an 80% training set (of which 20% were used for validation) and a separate 20% testing set. For the Google Earth and IGN datasets, which had no sub-categories, a similar division is applied: 80% of the data serves as the training set (of which 20% as validation) and 20% allocated to the testing set.

To improve training outcomes, a widely used strategy is the adoption of weight transfer from models pretrained on other datasets [16]. This practice significantly enhances the performance of both CNNs and transformers compared to randomly initializing the model's parameters. In our work, we employ transfer learning by initializing the model weights with pre-trained backbones trained on Cityscapes dataset [17] at the outset of training.

Furthermore, augmentations play a crucial role in preprocessing and enhancing images with their corresponding annotations for semantic segmentation tasks. The process starts with the retrieval of images and annotations from respective datasets. Subsequently, the images undergo random resizing, selected from a predefined set of scales, to generate various input resolutions. For consistency, a random crop of size (512, 512) is extracted from the enlarged image, with constrains on the maximum ratio of item categories within the crop. In addition, as part of the data augmentation, a random horizontal flip is applied with a probability of 0.5. Photometric distortion techniques are also employed to further diversify the data. Finally, the inputs are compressed into a format compatible with that segmentation neural networks.

3.2.3. Training and validation

DeepLabv3+ [12] is a cutting-edge semantic segmentation model that extends the DeepLab architecture family. It effectively captures multi-scale contextual information through a modified atrous (dilated) convolution approach. The model adopts a fully convolutional network with an encoder–decoder architecture. In this work, the encoder incorporates the ResNetV1c backbone network [18], followed by multiple convolutional layers to capture intricate spatial context. The decoder uses bi-linear up-sampling to achieve precise segmentation. DeepLabv3+ includes a spatial pyramid pooling module to enhance multi-scale feature representation, yielding accurate and fine-grained segmentation results.

In contrast, U-Net [11] is a well-established architecture for semantic segmentation tasks. Although renowned for its success in biomedical image segmentation [19], U-Net finds widespread use in various domains. The U-Net architecture is symmetric, with a contracting (encoder) and an expansive (decoder) path. The encoder employs

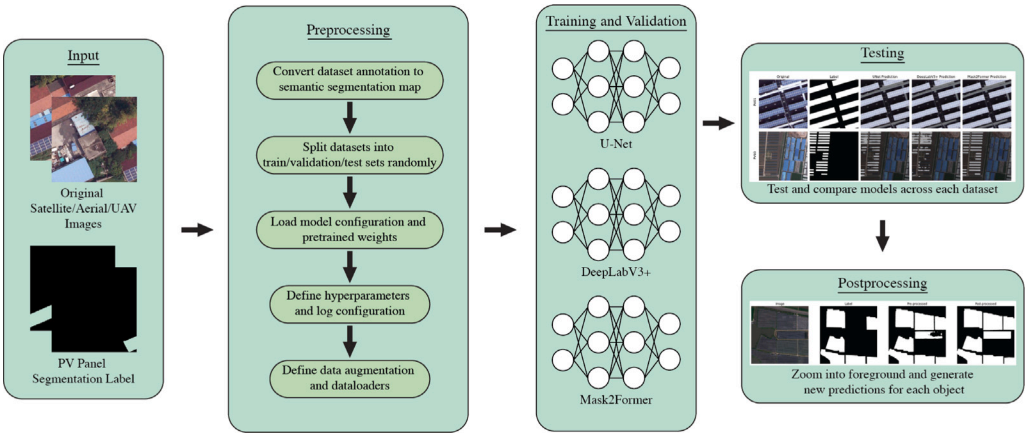

Input
Original Satellite/Aerial/UAV Images
PV Panel Segmentation Label
Preprocessing
Convert dataset annotation to semantic segmentation map
Split datasets into train/validation/test sets randomly
Load model configuration and pretrained weights
Define hyperparameters and log configuration
Define data augmentation and dataloaders
Training and Validation
U-Net
DeepLabV3+
Mask2Former
Testing
Test and compare models across each dataset
Postprocessing
Zoom into foreground and generate new predictions for each object

Fig. 2. Our proposed framework for the segmentation of photovoltaic panels. The input (images and corresponding annotations) is passed to the preprocessing stage, followed by training on the selected three different deep learning models. The appropriate hyperparameters are selected using the validation split. The performance is finally evaluated on the test set. Lastly, predictions are refined by a postprocessing step.

Table 2
Details of hardware and environment configuration, including different items and their corresponding configuration used in the study.

Category	Item	Configuration
Hardware	Cloud Platform	UCF Newton HPC Cluster
	GPU	NVIDIA Tesla V100 × 2
	CPU	Intel(R) Xeon(R) Gold 6226R CPU @2.90 GHz ×10
Environment	PyTorch	1.13.1
	Python	3.10.11
	Cuda	11.7
	MMCV	2.0.0
	MMSegmentation	1.0.0

Table 3
Hyperparameter configuration.

Model	Batch size	Optimizer	Learning rate	Weight decay	Loss function
U-Net	4	SGD	0.01	0.0005	Cross-entropy
DeepLabv3+	2	SGD	0.01	0.0005	Cross-entropy
Mask2Former	2	AdamW	0.0001	0.05	Binary cross-entropy and dice

convolutional and pooling layers to down-sample the input images, progressively capturing context and spatial information. The decoder incorporates transposed convolutions (also known as upsampling or deconvolution) to up-sample the feature maps and restoring the original resolution. For the decoder backbone, we opt for UNet-S5-D16.

Lastly, we introduce the Masked-attention Mask Transformer (Mask2Former) [9], representing the third model for comparison. Mask2Former is a state-of-the-art universal image segmentation architecture, offering the most recent advancements compared to other methods. It excels in various segmentation tasks while maintaining ease of training for each task. The foundation of Mask2Former builds upon a straightforward meta-architecture, consisting of a backbone feature extractor, a pixel decoder, and a Transformer decoder (Fig. 3). Notably, the model incorporates masked attention into the Transformer decoder, restricting attention to localized features centered around predicted segments, which can represent objects or regions based on specific semantic for grouping.

3.2.4. *Testing*

The performance of the trained models is rigorously assessed using five key metrics: Accuracy, Precision, Recall, F1-Score, and Intersection-over-Union (IoU) (Fig. 4). These metrics provide comprehensive insights into model performance, each serving a distinct purpose.

- Accuracy: This metric measures the overall correctness of the model's predictions, indicating the proportion of correctly classified pixels or objects. A higher accuracy score signifies more accurate segmentation results.
- Precision: Precision assesses the model's ability to make accurate positive predictions. It quantifies the ratio of true positive predictions to the total positive predictions made by the model.
- Recall: Recall, also known as sensitivity or true positive rate, evaluates the model's capacity to identify all relevant instances. It calculates the ratio of true positive predictions to the total actual positive instances in the dataset.
- F1-Score: The F1-Score is a balanced metric that considers both precision and recall. It offers a harmonized measure of the model's accuracy in capturing positive instances while minimizing false positives and false negatives.
- Intersection-over-Union (IoU): Notable, IoU holds particular importance in evaluating the performance of semantic segmentation models due to its ability to address the class imbalance issue. IoU quantifies the similarity between the predicted area and the corresponding ground-truth region for an object.

Higher values of these metrics indicate superior model performance, providing a comprehensive understanding of the segmentation quality achieved by the models.

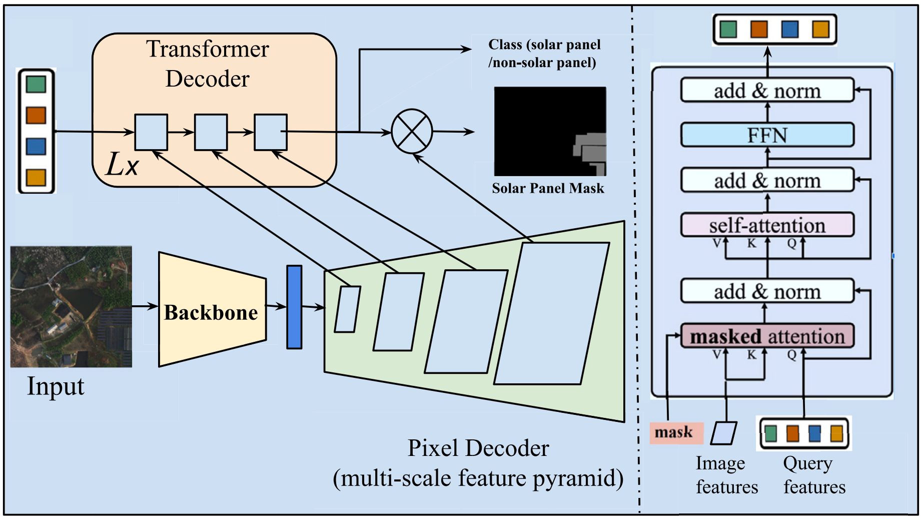
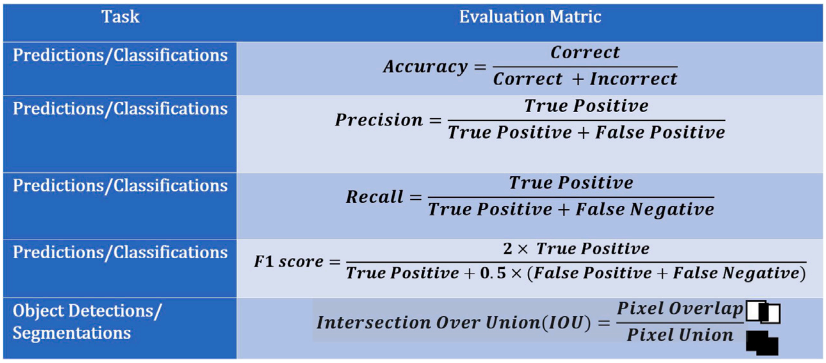


Fig. 3. Illustration of Mask2Former architecture containing backbone network (encoder) and two decoders (transformer and pixel decoders)(based on [9]). The transformer decoder (shown on the right) uses *masked attention* to enforce attention on object regions. We use Mask2Former for segmentation to obtain masks of photovoltaic panels.



Task	Evaluation Matric
Predictions/Classifications	$Accuracy = \frac{Correct}{Correct + Incorrect}$
Predictions/Classifications	$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$
Predictions/Classifications	$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$
Predictions/Classifications	$F1\ score = \frac{2 \times True\ Positive}{True\ Positive + 0.5 \times (False\ Positive + False\ Negative)}$
Object Detections/ Segmentations	$Intersection\ Over\ Union(IOU) = \frac{Pixel\ Overlap}{Pixel\ Union}$

Fig. 4. Performance evaluation of trained models using different metrics.

4. Experimental results and discussion

4.1. *Mask2Former augmentations*

To conduct a preliminary assessment of Mask2Former's performance, we trained it for a total of 10 epochs, utilizing two different settings: one without any data augmentation and the other with a combination of modified large scale jittering (LSJ) data augmentation and padding techniques. Padding is a common image preprocessing step in deep learning. It involves adding extra pixels around the border of an image. The augmentations employed in the latter setting include random resize, random cropping, random flipping, and photometric distortion. The resulting metrics for the multi-resolution datasets are summarized in Table 4. Notably, PV08 (satellite image, large-scale PV, China) exhibited superior IoU results when trained without any augmentations. Similarly, the PV03 (aerial image, distributed PV, China) dataset displayed better accuracy and IoU metrics without the use of data augmentations. Intriguingly, the PV01 (aerial image, rooftop PV, China) dataset demonstrated enhanced accuracy and IoU metrics when trained with the modified LSJ data augmentation combined with Pad. These findings highlight the Mask2Former model's dataset-dependent performance and emphasize the effectiveness of specific augmentation techniques in improving UAV imagery segmentation accuracy and IoU. Such disparities highlight the significance of dataset-specific exploration and optimization when using the Mask2Former model in a variety of satellite, and aerial imagery applications. Fig. 5 compares PV segmentation results of Mask2Former for this experiment. As we can see in the first row, we find that integrating data augmentation into certain datasets, such as PV01, improves feature extraction of rooftop PV modules.

Table 4
Performance across different datasets with and without augmentations.

Dataset	Augmentation	Accuracy	IoU
PV01	Yes	97.71	95.84
	Yes (+ Pad)	**97.84**	**95.97**
	No	96.99	95.10
PV03	Yes	97.09	94.36
	Yes (+ Pad)	97.65	94.98
	No	**97.79**	**95.44**
PV08	Yes	**96.98**	91.86
	Yes (+ Pad)	96.83	92.17
	No	96.36	**92.57**

4.2. *Mask2Former object queries*

We conducted additional experiments on the Mask2Former model to explore the impact of varying numbers of object queries during masked attention process. Object queries are important in the modified cross-attention mechanism, enabling the model to selectively attend to specific features within the generated feature maps of backbone network. These attended feature maps are subsequently used to generate object-level predictions, which include both class labels and spatial masks for each individual objects. In our experiments, we experimented

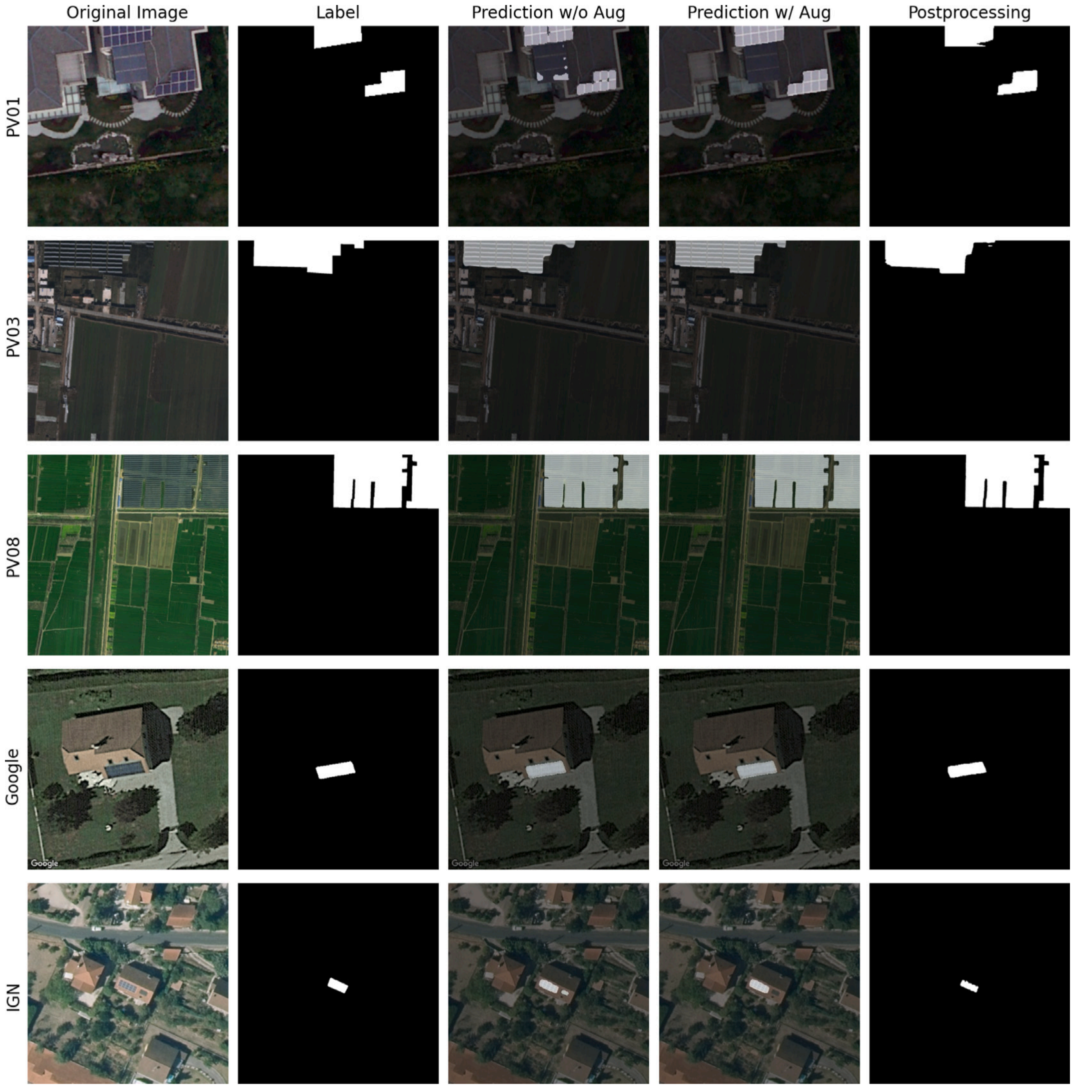


Fig. 5. Segmentation results of our Mask2Former model. We compare the results of our model when trained just on the original dataset versus data augmented dataset. The last column is the postprocessed prediction of the Mask2Former when trained with data augmented dataset.

different numbers of object queries, specifically 2, 100, and 200 (Table 5). Interestingly, the PV08 (satellite image, large-scale PV, China) dataset performed the best with 100 queries, PV03 (aerial image, distributed PV, China) with 200 queries, and PV01 (aerial image, rooftop PV, China) with only two queries. Notably, in the majority of cases, increasing the number of queries did not result in significant improvements in segmentation results. These findings suggest that the PV segmentation systems within masked attention mechanisms exhibits reduced sensitivity to the number of object queries. Moreover, it shows the dataset-specific nature of their impact on segmentation performance.

4.3. Mask2Former loss functions

In this subsection, we conduct experiments to investigate the impact of using focal loss as an alternative to cross-entropy (CE) loss during training. Focal loss introduces a modulating factor that assigns varying weights to different training examples based on their level of difficulty. Originally designed for object detection, focal loss has shown promise when applied to segmentation tasks, particularly in mitigating the challenges posed by class imbalance. However, our findings, as presented in Table 6, show that in the cases of PV03 (aerial image) and PV08 (satellite image), cross-entropy loss outperforms focal loss. The extent of the enhancement, although marginal, underscores the robustness and efficacy of cross-entropy loss in these scenarios.

4.4. Augmentation modification

We employed the “Random Choice Resize” method as part of our modified data augmentation technique. This approach involves resizing both the input images and their corresponding bounding boxes and masks from a range of multiple scales. Specifically, the input image is resized to a scale randomly selected from a range of 128 to 1024. The bounding boxes and masks are resized using the same scale factor applied to the image. Using this modified data augmentation strategy, the models were trained for 50 epochs. The experimental results (Table 7) demonstrated the effectiveness of this modified data augmentation technique in improving the segmentation performance of all three models. To contextualize these results, we conduct a comparative analysis with state-of-the-art scores [6,7], as showcased in Table 8. Our model outperformed existing methods across various performance metrics on PV01 (aerial image, rooftop PV, China), PV03 (aerial image, distributed PV, China), PV08 (satellite image, large-scale PV, China), and Google (satellite image, rooftop PV, France). Notably, even on the challenging IGN (aerial image, rooftop PV, France) dataset, our approach exhibits robust performance, evidenced by an average success

Table 5
Performance of our method when the number of queries is varied in Mask2Former. We observe competitive performance even with just two queries.

Dataset	Queries	Accuracy	Precision	Recall	F1-Score	IoU
PV01	2	**98.26**	**97.99**	**97.90**	**97.95**	**95.99**
	100	98.19	97.96	97.76	97.86	95.83
	200	98.20	97.98	97.77	97.87	95.85
PV03	2	**98.06**	94.89	96.76	95.80	92.10
	100	97.83	97.17	97.64	97.40	94.95
	200	97.91	**97.25**	**97.73**	**97.49**	**95.12**
PV08	2	98.05	94.79	**96.85**	95.79	92.08
	100	**98.10**	**95.07**	96.72	**95.87**	**92.24**
	200	98.03	94.82	96.71	95.73	91.98

Table 6
Ablation on loss functions (cross-entropy (CE) vs focal loss (Focal)). We observe better performance with CE loss.

Dataset	Loss	Accuracy	Precision	Recall	F1-Score	IoU
PV03	CE	**98.14**	**97.40**	**98.16**	**97.77**	**95.65**
	Focal	97.14	96.89	96.17	96.52	93.32
PV08	CE	**98.12**	**95.01**	**96.94**	**95.94**	**92.36**
	Focal	98.09	94.94	96.88	95.88	92.25

Table 7
Demonstration of the efficacy of our modified augmentations across different datasets.

Dataset	Model	Accuracy	Precision	Recall	F1-Score	IoU
PV01	U-Net	98.00	97.81	97.46	97.63	95.39
	DeepLabv3+	97.63	97.66	96.72	97.17	94.52
	Mask2Former	**98.38**	**98.17**	**98.00**	**98.09**	**96.25**
PV03	U-Net	96.58	95.78	95.99	95.89	92.15
	DeepLabv3+	88.37	89.09	82.29	84.76	74.27
	Mask2Former	**98.46**	**97.96**	**98.34**	**98.15**	**96.37**
PV08	U-Net	97.70	95.51	94.24	94.86	90.48
	DeepLabv3+	96.83	93.64	92.15	92.88	87.16
	Mask2Former	**98.32**	**96.17**	**96.43**	**96.30**	**93.00**
Google Earth	U-Net	99.74	96.55	96.15	96.35	93.17
	DeepLabv3+	99.72	96.87	95.36	96.10	92.75
	Mask2Former	**99.80**	**97.25**	**97.12**	**97.19**	**94.66**
IGN	U-Net	99.75	90.46	90.02	90.24	83.63
	DeepLabv3+	99.77	**93.86**	87.82	90.62	84.17
	Mask2Former	**99.79**	92.07	**91.86**	**91.96**	**86.13**

Table 8
Comparison with state-of-the-art methods on various performance metrics across different datasets. Our proposed framework, Mask2Former (M2F), consistently outperformed the existing methods such as Detail-oriented network (DON), DeeplabV3+ (DV3+), RefineNet (RN), and others found on the table.

Dataset	Model	Source	Accuracy	Precision	Recall	F1	IoU
PV01	U-Net	[6]	96.10	83.10	90.00	86.40	78.70
	RN	[6]	98.10	90.90	89.70	90.30	85.90
	DV3+	[6]	98.30	92.80	89.40	91.10	86.80
	DNLNet	[7]	96.36	86.08	96.36	90.93	83.37
	UPerNet	[7]	93.09	90.86	93.09	91.96	85.12
	U-Net	[7]	95.84	91.06	95.84	93.39	87.60
	DMNet	[7]	88.16	89.20	88.16	88.68	79.66
	PSPNet	[7]	88.43	87.25	88.43	87.83	78.31
	DV3+	[7]	96.82	83.77	96.82	89.82	81.53
	DON	[7]	94.77	89.51	94.77	92.06	85.30
	M2F	Ours	**98.38**	**98.17**	**98.00**	**98.09**	**96.25**
PV03	U-Net	[6]	97.30	89.70	93.50	91.60	85.80
	RN	[6]	97.60	95.70	93.70	94.70	87.80
	DV3+	[6]	98.30	95.90	93.10	94.50	90.80
	M2F	Ours	**98.46**	**97.96**	**98.34**	**98.15**	**96.37**
PV08	U-Net	[6]	98.00	87.10	86.40	86.80	77.60
	RN	[6]	97.90	84.80	88.40	86.60	77.30
	DV3+	[6]	**98.40**	87.70	85.70	86.70	79.00
	M2F	Ours	98.32	**96.17**	**96.43**	**96.30**	**93.00**
Google	DON	[7]	89.15	92.09	89.15	90.59	82.81
	M2F	Ours	**97.28**	**97.01**	**97.28**	**97.14**	**94.58**
IGN	M2F	Ours	91.97	92.08	91.97	92.02	86.22

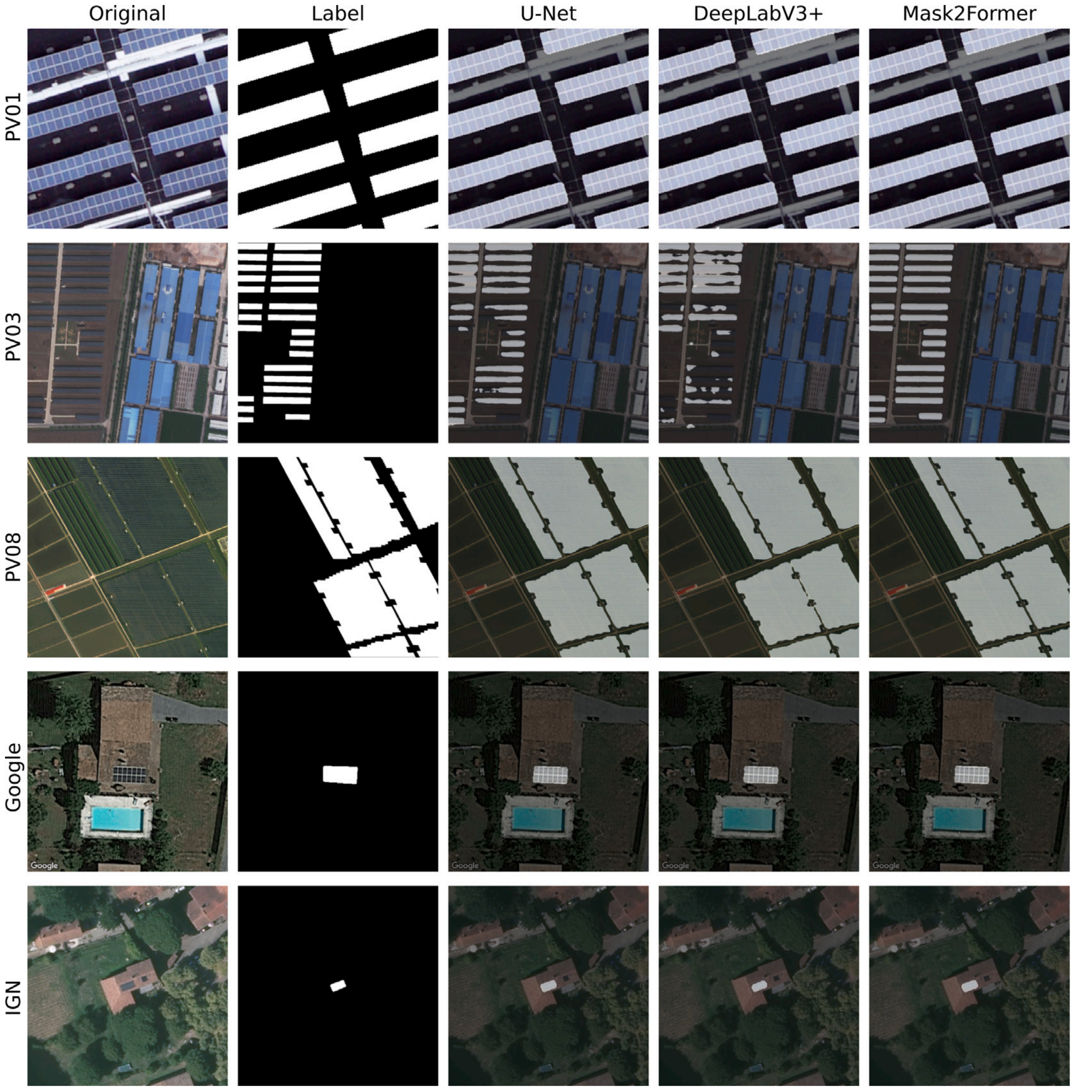


Fig. 6. Inference results of Mask2Former, U-Net, and DeepLabV3+ tested on PV01, PV03, PV08, Google, and IGN. The performance improvements of Mask2Former can be observed mainly on the PV03 dataset, where U-Net and DeepLabV3+ struggle to segment all visible photovoltaic panels. All models where trained on the data augmented version of each dataset.

rate of over 90% across the key metrics accuracy, precision, recall, F1 score, and IoU. Fig. 6 shows five examples where Mask2Former excels over previous methods before our postprocessing stage. On the second row, PV03, we can see Mask2Former is superior at segmenting all the photovoltaic panels in the image.

4.5. Postprocessing

To refine the predictions of the Mask2Former, U-Net, and DeepLabv3+, we employ a post-processing approach during the inference stage. To begin, we first identify the contours of objects within the binary mask image (prediction image) and crop the original image based on the bounding boxes of the objects. These cropped foreground objects are then resized to match the original image size and stored in a list. Subsequently, we leverage this list of newly generated predictions for the foreground objects and replace the corresponding regions in the original prediction image with these new predictions. Fig. 7 shows postprocessing predictions for U-Net, DeepLabV3+, and Mask2Former on images from PV01, PV03, PV08, Google, and IGN. We can see that in general, our postprocessing step does produce qualitatively better segmentation results. Especially in PV03, where the postprocessing prediction is better than the ground truth label. The improvements on PV01, PV 08, Google and IGN are marginal, since the predictions without postprocessing are already of high quality. However, it is important to note that since the IoU score measures against the ground truth label, we observe a decrease in the IoU, when the postprocessing produce predictions better than the ground truth label.

4.6. Discussion

In this section we perform a variety of experiments to optimize PV panel segmentation with different model architectures. During our loss function experiment we find that cross-entropy loss outperforms focal loss on all tested datasets. Additionally, we delve deeper into our data augmentation and postprocessing steps. Data augmentation improves performance across measured metrics by a small margin. Ultimately, Mask2Former demonstrates the best performance across all datasets; see Table 8. We believe that this is mainly due to the novelty and size of the Mask2Former architecture. U-Net, DeepLabV3+, and Mask2former have 29, 62, and 215 million trainable parameters, respectively [15]. That is, Mask2Former has over 7 times more parameters than U-Net and over 3 times more parameters than DeepLabV3+. We conjecture the general segmentation architecture [9] and greater parameter count is the reason Mask2Former outperforms previous methods in PV segmentation task. As for our postprocessing step, it does improve performance

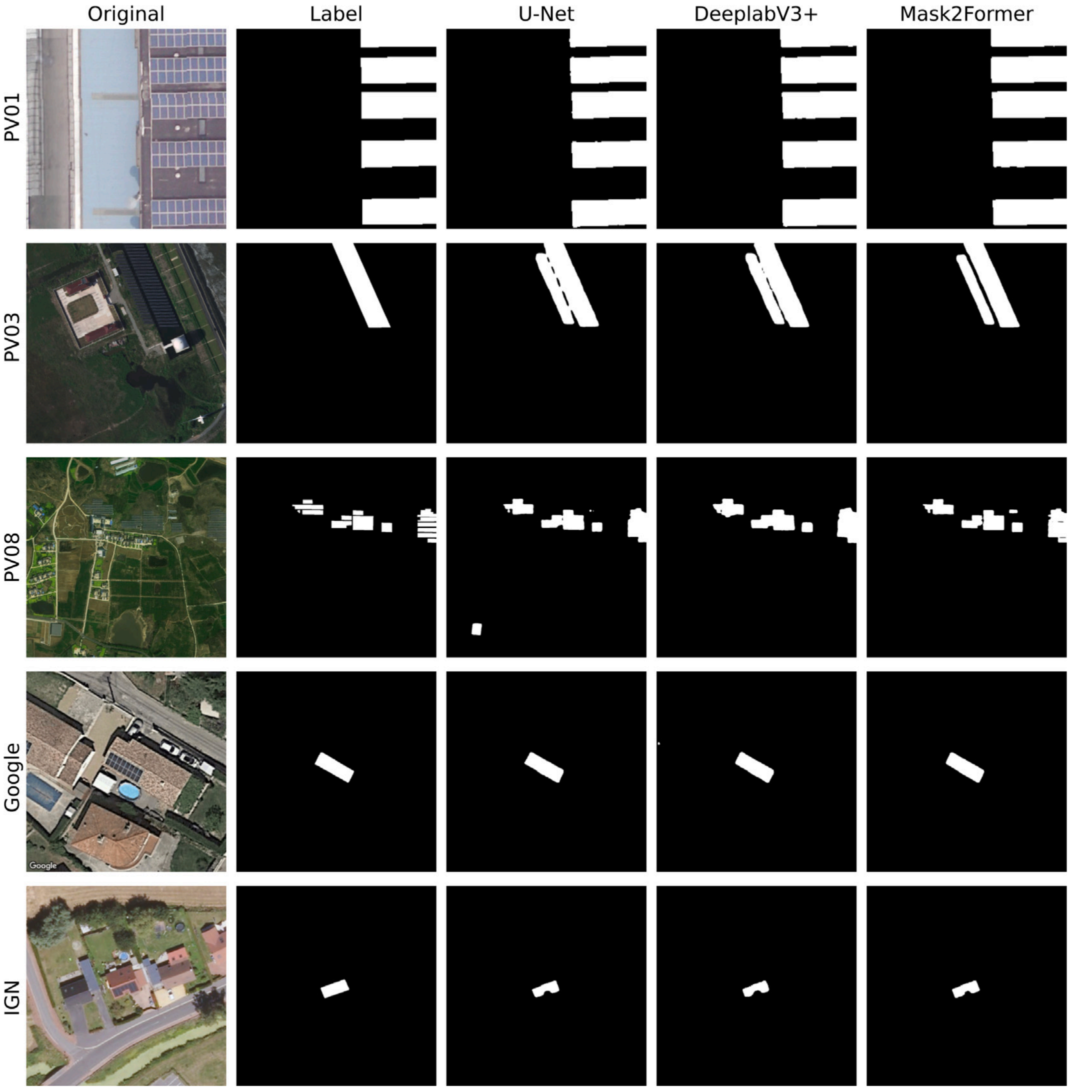


Fig. 7. Qualitative comparison of postprocessing prediction of U-Net, DeepLabV3+, and Mask2Former across PV01, PV03, PV08, Google, and IGN. We observe that in general our postprocessing step result in marginal improvements of predictions in the PV03 and IGN datasets. Especially for the PV03 dataset, the postprocessing prediction is even better than the ground truth label. Postprocessed predictions were generated using predictions from models trained on data augmented dataset.

across tested metrics, expecially on distributed PV installation in aerial images captured in PV03, see Fig. 7. Although the improvements achieved through postprocessing on other datasets are marginal, we hope our work will inspire the development of better, more specialized method for segmenting PV panels.

5. Conclusion

In this study, we explored the capabilities of the novel transformer-based deep learning model, Mask2Former, for PV segmentation in aerial and satellite imagery. Our comparative analysis with popular convolutional neural network (CNN)-based models, such as U-Net and DeepLabv3+, demonstrated that Mask2Former consistently surpassed these models in segmenting various types of PV installations, including large-scale utility, distributed ground-mounted, and small-scale rooftop PV systems, across diverse locations in China and France. Furthermore, we delved into the effects of data augmentation, the number of object queries, loss function types, and post-processing techniques on the model's performance. Our research revealed that data augmentation significantly aids in the detection of rooftop PV in aerial images, though its impact is somewhat limited in distributed and large-scale utility PV installations. In examining the influence of the number of object queries, we discovered that high-resolution aerial images of rooftop PV installations perform optimally with merely two queries, while distributed and large-scale PV installations in satellite imagery do get better segmentation results with over 100 queries. Additionally, our experiments comparing focal loss with cross-entropy loss indicated that in most instances, cross-entropy loss proved more effective. Notably, the implementation of the modified data augmentation technique, 'Random Choice Resize', was a pivotal factor, enhancing the performance of all three models and achieving the highest scores across all metrics. Our post-processing technique demonstrated significant success in the context of distributed PV in aerial imagery; however, its impact on other datasets was relatively marginal. These outcomes underscore the complexity of segmentation tasks and the critical need for tailored exploration and optimization of these models, particularly when applied to satellite, and aerial imagery of different type of PV installation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Gen AI disclaimer

During the preparation of this work the author(s) used ChatGPT to assist in the initial drafting and editing of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgments

This material is based upon work supported by the National Science Foundation, United States under Award No. 2050731. The revision of the manuscript is supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technology Office (SETO) Renewables Advancing Community Energy Resilience (RACER) Award Number DE-EE0010418.

References

[1] Snapshot 2023, 2023, https://iea-pvps.org/snapshot-reports/snapshot-2023/.

[2] Renewables 2023: Executive summary, 2023, https://www.iea.org/reports/renewables-2023/executive-summary#.

[3] Global PV market outlook, 4Q 2023, 2023, https://about.bnef.com/blog/global-pv-market-outlook-4q-2023/.

[4] X. Liu, Z. Deng, Y. Yang, Recent progress in semantic image segmentation, Artif. Intell. Rev. 52 (2019) 1089–1106.

[5] R. Ravishankar, E. AlMahmoud, A. Habib, O.L. de Weck, Capacity estimation of solar farms using deep learning on high-resolution satellite imagery, Remote Sens. 15 (1) (2022) 210.

[6] H. Jiang, L. Yao, N. Lu, J. Qin, T. Liu, Y. Liu, C. Zhou, Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery, Earth Syst. Sci. Data 13 (11) (2021) 5389–5401.

[7] R. Zhu, D. Guo, M.S. Wong, Z. Qian, M. Chen, B. Yang, B. Chen, H. Zhang, L. You, J. Heo, et al., Deep solar PV refiner: A detail-oriented deep learning network for refined segmentation of photovoltaic areas from satellite imagery, Int. J. Appl. Earth Obs. Geoinf. 116 (2023) 103134.

[8] M. Kleebauer, C. Marz, C. Reudenbach, M. Braun, Multi-resolution segmentation of solar photovoltaic systems using deep learning, Remote Sens. (ISSN: 2072-4292) 15 (24) (2023) http://dx.doi.org/10.3390/rs15245687.

[9] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1290–1299.

[10] G. Kasmi, Y.-M. Saint-Drenan, D. Trebosc, R. Jolivet, J. Leloux, B. Sarr, L. Dubus, A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata, Sci. Data 10 (1) (2023) 59.

[11] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.

[13] V. Plakman, J. Rosier, J. van Vliet, Solar park detection from publicly available satellite imagery, GIScience. Remote Sens. 59 (1) (2022) 462–481.

[14] P. Parhar, R. Sawasaki, A. Todeschini, H. Vahabi, N. Nusaputra, F. Vergara, HyperionSolarNet: solar panel detection from aerial images, 2022, arXiv preprint arXiv:2201.02107.

[15] MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark, https://github.com/open-mmlab/mmsegmentation.

[16] L. Zhuang, Z. Zhang, L. Wang, The automatic segmentation of residential solar panels based on satellite images: A cross learning driven U-net method, Appl. Soft Comput. 92 (2020) 106283.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[18] C. Zhang, Z. Huang, D. Ye, G. Cai, F. Xue, SFR-net: A spatial feature enhance method for road extraction, in: 2022 12th International Conference on Information Technology in Medicine and Education (ITME) V, IEEE, 2022, pp. 662–665.

[19] H. Seo, M. Badiei Khuzani, V. Vasudevan, C. Huang, H. Ren, R. Xiao, X. Jia, L. Xing, Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications, Med. Phys. 47 (5) (2020) e148–e167.