SSSwin: Sequential Spectral Swin Transformer for Solar Panel Mapping in Satellite Imagery

Zhiyuan Yang and Ryan Rad
Khoury College of Computer Science, Northeastern University
Vancouver, Canada
{yang.zhiyu, r.rad}@northeastern.edu

***Abstract*—Global sustainability is increasingly reliant on solar energy. However, effectively monitoring solar farms and accurately assessing our progress in transitioning towards this renewable energy source remains a challenge, especially on a global scale. This study introduces SSSwin, a novel vision transformer model specifically developed to improve the mapping of solar panel farms using satellite imagery. The cornerstone of SSSwin is its Sequential Spectral Embedding module, uniquely designed to address the three-dimensional aspects of multispectral satellite images, enabling intricate capture of spatial-spectral data. To validate its effectiveness, we incorporated our new module into the design of two state-of-the-art models: UPerNet-SwinB and Mask2Former-SwinB. The experimental results demonstrate that this integration enhanced both of their performance without compromising efficiency. https://github.com/yzyly1992/SSSwin**

***Index Terms*—solar energy, vision transformer, segmentation, multispectral, sequential**

I. Introduction

Solar panel farms are integral to the global shift towards renewable energy, playing a vital role in sustainable development [1]. Effective monitoring of these installations is crucial for multiple reasons. Firstly, it enables the evaluation of solar farms' energy generation and performance over time, providing essential data for optimizing their operation and maintenance [2]. Additionally, monitoring aids in assessing the environmental impact and land use changes associated with solar energy production [3]. On a broader scale, remote sensing analysis of solar farms offers valuable insights into global renewable energy trends and development patterns. Given the anticipated increase in solar installations worldwide, the development of efficient and precise monitoring solutions maximizes the benefits derived from this key source of clean energy [4].

Multispectral satellite imagery presents unique challenges in remote sensing, particularly in applications like land cover categorization and solar panel farm monitoring [5]. While deep convolutional neural networks (CNNs) have been groundbreaking in semantic segmentation of such images, they exhibit inherent limitations. The main challenge lies in their local convolutional kernels, which may not efficiently capture the complex spatial long-range and global cross-band dependencies inherent in multispectral data [6]. This limitation becomes more pronounced in the context of monitoring and analyzing solar panel farms, where precise and comprehensive data interpretation is crucial.

In response to these challenges, Vision Transformers have emerged as a promising alternative, offering a novel approach to image processing. Among these, the Swin Transformer stands out, integrating self-attention mechanisms and relative position biases. This design enables it to leverage both local and long-range dependencies effectively [7]. However, while promising, standard Swin Transformers are not fully optimized for the unique three-dimensional structure of multispectral image cubes. This gap highlights the need for a more tailored approach that can effectively handle the intricacies of multispectral satellite imagery, especially in applications like solar panel farm monitoring.

In this work, we introduce the SSSwin Transformer for multispectral image segmentation. The SSSwin introduces a Sequential Spectral Embedding module to better handle the spatial-spectral properties of multispectral data. It facilitates the capturing of detailed information across bands while maintaining the advantages of vision transformers. By combining sequential modeling with multispectral feature extraction, our SSSwin method aims to achieve superior performance. By improving the existing Swin Transformer to better handle the multispectral data structure, our approach seeks to advance the state-of-the-art. This is especially important for applications involving solar panel farm monitoring from multispectral satellite imagery.

The contributions of this work can be summarized as follows:

- Modifications have been made to Swin Transformer architecture to better suit the characteristics of multispectral data, like those from Sentinel-2 satellites. These adjustments aim to improve feature capture in multispectral image segmentation.
- The study addresses dataset limitations by developing a detailed multispectral image dataset, specifically focusing on solar panel farms. This serves as a valuable resource for research and practical applications in solar energy.
- The work demonstrates significant improvements in multispectral image segmentation for solar panel farms through extensive testing. The results show enhanced accuracy and efficiency, suggesting the potential of these methods for more effective and scalable monitoring solutions.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00237

II. Related Work

A. Deep Convolutional Neural Network

When it comes to deep convolutional neural networks, UNet [8] stands as a benchmark for biomedical image segmentation, employing an encoder-decoder structure with skip connections. Complementing this, DeepLabV3 [9] introduced atrous convolution to explicitly address the multi-scale context in image segmentation tasks. In this context, UPerNet [10] distinguishes itself by incorporating a pyramid attention module, elevating feature representation, and global context capture for even more precise biomedical image segmentation.

B. Vision Transformer

The emergence of Vision Transformer (ViT) [11] marked a paradigm shift by applying transformers to natural images through patch-based tokenization. Swin Transformer [7] later refined this approach with a shifted window mechanism, achieving linear computational and memory complexity. Building upon Swin Transformer, Swin-UNet [12] integrated it into a UNet-like framework, showcasing the versatility of transformer architectures in segmentation. Concurrently, SegFormer [13] proposed a lightweight vision transformer focusing on self-attention, while Mask2Former [14] leveraged masked attention mechanism to enhance segmentation mask predictions.

C. Hybrid Vision Transformer

In the domain of hybrid vision transformers, models like MAE [15] and BEiT [16] explored self-supervised reconstruction for pretraining, followed by fine-tuning for downstream tasks such as segmentation. These approaches contribute to the growing flexibility and adaptability of transformer-based architectures.

D. Vision Transformer for Multispectral Imagery

Expanding the scope to multispectral imagery, 3D Swin Transformer [17] extends the Swin Transformer for multispectral imagery by processing volumetric data with shifted windows. SpectralSWIN [18] is the first transformer-based model for hyperspectral image classification, leveraging the Swin Spectral Module (SSM) for concurrent spatial and spectral feature capture, demonstrating superior performance on two Hyperspectral imaging (HSI) datasets. SpectralFormer [19], a novel transformer backbone, addresses limitations in spectral sequence attribute capture, outperforming classic transformers and state-of-the-art (SOTA) networks in hyperspectral image classification across three datasets.

Current Vision Transformer models are not optimized for multispectral data, limiting their use in fields like remote sensing. They mainly focus on classification over segmentation and their complex optimization methods reduce practicality. Additionally, their application in solar panel farm mapping, crucial for energy management, remains unexplored. Addressing these gaps could significantly broaden their utility.

III. Methodology

A. Global Dataset

To develop an effective methodology for our research, we first generated a comprehensive dataset by combining various sources of information [20]. This dataset consisted of three main components: the Historical Global Ground Truth from 2017-2018 [21], the Manual Annotated US Ground Truth from 2022-2023 using Google Earth Engine, and the Sentinel-2 Multispectral Satellite Imagery spanning the years 2017-2023 [22]. By merging these datasets, we created a comprehensive initial training set with $10,230$ multispectral images with dimensions of $256 \times 256 \times 13$. The dataset has been published at https://github.com/yzyly1992/GloSoFarID

After having the initial training set, we employed multiple state-of-the-art models to train on the dataset with a batch size of 16, and for 50 epochs. We achieved very good accuracy in segmenting the solar panel farms from multispectral images with these models. The best model can reach an IoU of 96.47%, and an F-score of 98.2%.

From these models, we selected the top three performers based on their predictive accuracy. These selected models were then utilized to predict and combine labels for the global dataset spanning the years 2021-2023. This step allowed us to leverage the strength of the best-performing models to provide accurate predictions for the desired time frame.

To ensure the quality and reliability of our dataset, we implemented a process to remove noise and weak predicted samples. By applying rigorous filtering techniques, we were able to eliminate any erroneous or unreliable data points. The resulting dataset was then considered to be the final dataset, containing the most up-to-date and accurate information available.

B. SSSwin: Sequential Spectral Swin Transformer

The Swin Transformer was initially built with 2D Patch Partition, Linear Embedding, and Swin Transformer Blocks. However, its design and testing primarily focused on standard RGB images. When dealing with images containing more than 3 spectral bands, its architecture may not ensure optimal performance. Our proposed SSSwin Transformer takes a novel approach by customizing the original Swin Transformer Architecture, specifically addressing the limitations of the Patch Partition Method and Embedding Strategy when handling multispectral data. These tailored modules aim to boost the overall performance of the Swin Transformer.

1) 3D Patch Partition: The purpose of the patch partition is to enable the effective processing of images by dividing them into smaller patches. The Swin Transformer utilizes this patch partition by breaking down images into patches along both height and width dimensions, facilitating the extraction of features from localized regions. However, a limitation arises as each patch encompasses information from all spectral bands. During the embedding process, all spectral bands within a patch are treated as a single unified representation. This approach becomes challenging when dealing with multispectral

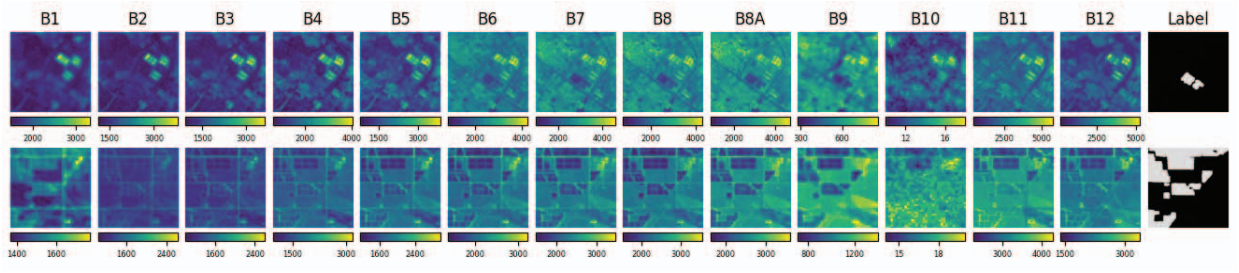

Fig. 1. Examples from our proposed multispectral dataset [20] Each sample is 256 by 256 pixels with a 10m resolution and contains 13 bands of satellite information, labeled as B1 to B12. The image on the right represents the mask for the ground truth, where the gray color indicates the solar panel area and the black color represents the non-solar panel area.

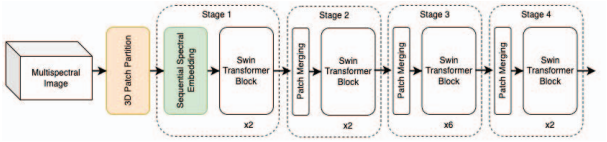

Fig. 2. SSSwin Transformer Architecture

bands, as it fails to adequately capture the intricate relationships among the various spectral bands.

To address this challenge more effectively, we introduced a 3D Patch Partition Module. Unlike the conventional approach of solely splitting images along the height and width dimensions, the 3D Patch Partition Module extends its capability by incorporating spectral dimension into the partitioning process. By doing so, each patch unit not only retains a clear spatial relationship but also captures distinct spectral information.

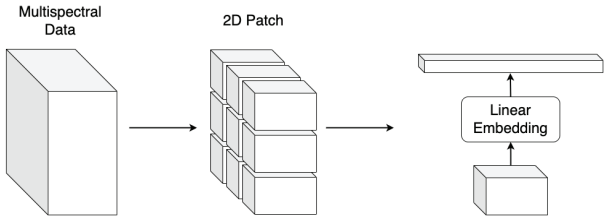

Fig. 3. Original Swin Patch Embedding

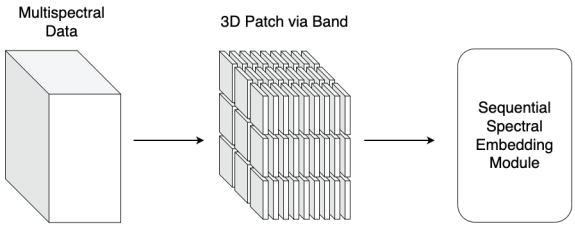

Fig. 4. SSSwin with 3D Patch and Sequential Spectral Embedding

$$\text{Multispectral Image } (X) \in R^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$$

$$\text{3D Patched Image } (X_p) \in R^{\mathrm{N}\times(\mathrm{P}\times\mathrm{P}\times 1)}$$

$$\text{Number of Patches } N = \mathrm{H}\times\mathrm{W}\times\mathrm{C}/(\mathrm{P}\times\mathrm{P}\times 1).$$

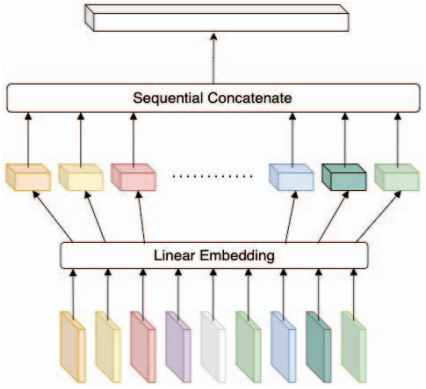

Fig. 5. Sequential Spectral Embedding Module

2) Sequential Spectral Embedding Module: The original Swin Transformer utilizes linear embedding to directly incorporate 2D partitioned patches into the target embedding dimension. As discussed in the partition module section, this direct embedding strategy consolidates all spectral features into a single representation, making it challenging to capture intricate spectral relationships.

To address this limitation and enhance spectral information handling, we introduced the Sequential Spectral Embedding Module. This module, in conjunction with the 3D Patch Partition Module, embeds each 3D patch into a smaller dimension D_P (derived by dividing the target dimension D_T by the spectral dimension C) and sequentially concatenates them along the spectral dimension. This approach allows us to maintain the same target embedding dimension for the output while preserving the sequential relationships between the spectral bands.

$$D_P = D_T/\mathrm{C}$$

IV. Result and Discussion

A. Experiment Setting

To evaluate the efficacy of the SSSwin Transformer Model, we conducted a comprehensive comparative analysis involving eight SOTA models. Our assessment involved rigorous training and testing on the newly introduced Multispectral Solar Panel

Farms dataset, a purposefully curated collection designed to assess the model's performance in the presence of diverse spectral bands and complex environmental conditions. The chosen SOTA models were selected based on their prominence and relevance in the field, ensuring a robust benchmark for our proposed SSSwin Transformer.

The hardware environment utilized a Cluster, featuring NVIDIA Tesla V100 SXM2 GPUs with 32 GB of memory. Each model underwent training for 50 epochs, employing a batch size of 16. Additionally, we implemented early stopping with a patience of 10, tracking the mean Intersection over Union (IoU) value.

The metrics employed to evaluate the performance of our models encompass IoU and F-score. Given the nature of the task, which involves semantic segmentation, these metrics play a pivotal role in assessing the accuracy and precision of the model's predictions. IoU, measuring the overlap between predicted and ground truth masks, provides insights into the spatial alignment of segmentation results. Concurrently, the F-score, considering both precision and recall, offers a comprehensive assessment of the model's ability to correctly classify pixels belonging to the target classes.

B. Quantitative Results

To validate the efficacy of the SSSwin Transformer, we adopted the Mask2Former [14] and UPerNet [10] models, incorporating the SSSwin Transformer as their dedicated encoder backbone. Subsequently, we conducted a comparative analysis by contrasting the outcomes obtained using these customized models against those generated by Mask2Former and UPerNet utilizing the original Swin Transformer as their encoder. This comparative study serves to elucidate the specific advantages and improvements conferred by the SSSwin Transformer in the context of these segmentation models.

Moreover, we extended our evaluation to include benchmarking against CNN models such as U-Net [8] and DeepLabV3 [9], and SOTA Vision Transformer models such as Swin-UNet [12], SegFormer [13], MAE [15], and BEiT [16]. As TABLE I shows, by encompassing a diverse set of reference models, we aimed to provide a comprehensive assessment of the SSSwin Transformer's performance relative to established methodologies, thereby offering valuable insights into its overall competitiveness and potential advancements in solar panel farm mapping tasks. We also include the number of parameters and floating point operations (FLOPs) to measure the complexity of the model, enabling a nuanced understanding of computational efficiency alongside performance metrics. This comprehensive analysis contributes to the broader understanding of Vision Transformer research, highlighting the strengths and areas for improvement of the SSSwin model in comparison to contemporary benchmarks.

C. Qualitative Results

Fig. 6 presents qualitative results comparing the segmentation masks for solar panel farms generated by the proposed SSSwin Transformer Models, original Swin Transformer Models, CNN Models, and SOTA Vision Transformer Models. The figure consists of ten rows, each representing a sample area of $2,560$m by $2,560$m. Five samples are from North America, three samples are from Europe, and two samples are from Asia. The first column displays the RGB representation of the multispectral data. Columns 2 to 8 show prediction results from different models. In each prediction mask, white represents true positives, black represents true negatives, red represents false positives, and blue represents false negatives. These qualitative results offer a visual assessment of the model's performance in accurately mapping solar panel farm areas, highlighting the effectiveness of the proposed approach in learning from multispectral images for solar farm mapping tasks.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	Params	FLOPs	IoU	F-score
UNet-MobileNetV2 [8], [23]	10.1M	26.16G	42.08	48.18
DeepLabV3-EfficientNet [9], [24]	9.0M	10.17G	66.54	46.28
Swin-UNet [12]	26.6M	9.34G	36.35	45.0
UPerNet-MAE-base [10], [15]	163M	148G	71.24	83.21
UPerNet-BEiT-base [10], [16]	163M	148G	73.29	84.59
SegFormer-B5 [13]	82.0M	16.87G	75.54	85.06
UPerNet-SwinB [7], [10]	120M	81.87G	76.92	86.95
Mask2Former-SwinB [7], [14]	110M	442G	79.32	88.47
UPerNet-SSSwin*	120M	81.77G	**78.49**	**87.95**
Mask2Former-SSSwin*	110M	442G	**80.34**	**89.1**

D. Discussion

The consistent uptick of approximately 1% ($\pm 0.5\%$) in IoU and F-score across both Mask2Former and UPerNet configurations showcases the robust adaptability of SSSwin, irrespective of the underlying segmentation architecture. Notably, the superior performance of the Mask2Former model coupled with SSSwin stands out, achieving an impressive IoU of 80.34% and an F-score of 89.1%. This particular synergy underscores the compatibility between the Mask2Former architecture and the unique features introduced by the SSSwin Transformer. The nuanced understanding required for multispectral segmentation, especially in the context of solar panel farms, appears to be effectively captured by this combined approach.

The comparison with other SOTA segmentation models, including U-Net, DeepLabV3, Swin-UNet, SegFormer, MAE, and BEiT, would be essential to position the SSSwin Transformer within the broader landscape of segmentation techniques. Future work may delve into exploring the computational efficiency and scalability of SSSwin across larger datasets and diverse environmental conditions.

V. CONCLUSION

This study introduces the SSSwin Transformer, a significant advancement in multispectral image segmentation with a focus on solar panel farm monitoring. Our integration of the SSSwin Transformer into two best-performing SOTA models, Mask2Former and UPerNet architectures, has led to notable

Fig. 6. Qualitative comparison of our predicted solar panel maps (by SSSwin) and the ground truth across different continents. Here, we represent True Positive as **white**, True Negative as **black**, False Positive as **red**, and False Negative as **blue**.

improvements in segmentation accuracy. This highlights the transformative impact of our approach in the field. The SSSwin Transformer has proven its versatility as an effective encoder backbone across different segmentation models. The combination of SSSwin and Mask2Former, in particular, has yielded exceptional results, achieving an IoU of 80.34% and an F-score of 89.1%. In future work, we plan to expand the application of the SSSwin Transformer to other domains. We aim to explore the potential of solar map identification in a semi-supervised manner. As the need for global adoption of solar energy grows, our findings are expected to enhance the sustainability and overall efficacy of this crucial renewable resource.

REFERENCES

[1] V. Fthenakis, "Sustainability of photovoltaics: The case for thin-film solar cells," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 9, pp. 2746–2750, 2009.

[2] K.-C. Liao and J.-H. Lu, "Using uav to detect solar module fault conditions of a solar power farm with ir and visual image analysis," *Applied Sciences*, vol. 11, no. 4, p. 1835, 2021.

[3] D. Turney and V. Fthenakis, "Environmental impacts from the installation and operation of large-scale solar power plants," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 6, pp. 3261–3270, 2011.

[4] P. Choudhary and R. K. Srivastava, "Sustainability perspectives-a review for solar photovoltaic trends and growth opportunities," *Journal of Cleaner Production*, vol. 227, pp. 589–612, 2019.

[5] J. Rogan and D. Chen, "Remote sensing technology for mapping and monitoring land-cover and land-use change," *Progress in planning*, vol. 61, no. 4, pp. 301–325, 2004.

[6] E. Saralioglu and O. Gungor, "Semantic segmentation of land cover from high resolution multispectral satellite images by spectral-spatial convolutional neural network," *Geocarto International*, vol. 37, no. 2, pp. 657–677, 2022.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[8] O. Ronneberger, P. Fischer, and 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.

[9] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587

[10] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," *CoRR*, vol. abs/1807.10221, 2018. [Online]. Available: http://arxiv.org/abs/1807.10221

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[12] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Cham: Springer Nature Switzerland, 2023, pp. 205–218.

[13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *CoRR*, vol. abs/2105.15203, 2021. [Online]. Available: https://arxiv.org/abs/2105.15203

[14] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," *CoRR*, vol. abs/2112.01527, 2021. [Online]. Available: https://arxiv.org/abs/2112.01527

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, "Masked autoencoders are scalable vision learners," *CoRR*, vol. abs/2111.06377, 2021. [Online]. Available: https://arxiv.org/abs/2111.06377

[16] H. Bao, L. Dong, and F. Wei, "Beit: BERT pre-training of image transformers," *CoRR*, vol. abs/2106.08254, 2021. [Online]. Available: https://arxiv.org/abs/2106.08254

[17] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-d-swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[18] S. Ayas and E. Tunc-Gormus, "Spectralswin: a spectral-swin transformer network for hyperspectral image classification," *International Journal of Remote Sensing*, vol. 43, no. 11, pp. 4025–4044, 2022. [Online]. Available: https://doi.org/10.1080/01431161.2022.2105668

[19] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, dOI: 10.1109/TGRS.2021.3130716.

[20] Z. Yang and R. Rad, "Glosofarid: Global multispectral dataset for solar farm identification in satellite imagery," 2024.

[21] L. Kruitwagen, K. Story, J. Friedrich, L. Byers, S. Skillman, and C. Hepburn, "A global inventory of photovoltaic solar energy generating units," *Nature*, vol. 598, no. 7882, pp. 604–610, 2021.

[22] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort *et al.*, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[24] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.