Cross-Domain Generalization of Multimodal LLMs for Global Photovoltaic Assessment

Muhao Guo
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, United States
mguo26@asu.edu

Yang Weng
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, United States
Yang.Weng@asu.edu

***Abstract*—The rapid expansion of distributed photovoltaic (PV) systems poses challenges for power grid management, as many installations remain undocumented. While satellite imagery provides global coverage, traditional computer vision (CV) models such as CNNs and U-Nets require extensive labeled data and fail to generalize across regions. This study investigates the cross-domain generalization of a multimodal large language model (MLLM) for global PV assessment. By leveraging structured prompts and fine-tuning, the model integrates detection, localization, and quantification within a unified schema. Cross-regional evaluation using the ΔF1 metric demonstrates that the proposed model achieves the smallest performance degradation across unseen regions, outperforming conventional CV and transformer baselines. These results highlight the robustness of MLLMs under domain shift and their potential for scalable, transferable, and interpretable global PV mapping.**

***Index Terms*—LLM, Photovoltaic, Imagery, Solar Panels**

I. Introduction

The rapid deployment of distributed photovoltaic (PV) systems has transformed the landscape of modern power distribution networks [1]. Accurate visibility of small-scale and residential PV installations is essential for power grid operators to perform reliable load forecasting [2]–[4], hosting capacity analysis [5], [6], and distributed energy resource (DER) management [7]. However, many of these installations remain undocumented in official records. Existing databases, such as those maintained by the U.S. Energy Information Administration and state interconnection registries, are typically updated infrequently and lack sufficient spatial granularity. This information gap limits situational awareness in the planning and operation of active distribution networks and microgrids.

Satellite imagery provides a globally consistent and continuously updated source of information for PV assessment. Yet traditional computer vision (CV) approaches, including convolutional neural networks (CNNs) [8] and U-Net-based segmentation models [9], encounter several critical challenges. They demand large volumes of labeled data [10], struggle to distinguish visually similar surfaces such as dark roofs and parking lots, and exhibit poor generalization when transferred to regions with different rooftop geometries or illumination conditions. Furthermore, these models are designed for narrow, single-task objectives and lack interpretability, making their predictions difficult to validate and integrate into decision-making processes for power systems.

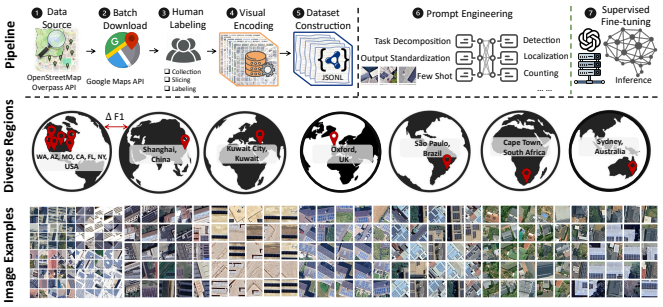

Fig. 1: Framework of the proposed MLLM for global PV assessment. The pipeline includes data engineering (Steps 1 - 6), prompt engineering, and supervised fine-tuning stages. Rooftop imagery from seven global regions is used to evaluate cross-domain generalization, with ΔF1 representing performance changes between the fine-tuning and unseen regions.

Recent advances in LLMs have opened a new direction for multimodal learning [11]. By jointly reasoning over text and imagery, LLMs are capable of performing high-level semantic interpretation, contextual reasoning, and adaptive decision-making with minimal supervision. Such capabilities make them particularly promising for scalable solar infrastructure mapping. In earlier work, the PV Assessment with LLMs (PVAL) [12] framework demonstrated that fine-tuned MLLMs can detect, localize, and quantify rooftop PV installations with competitive accuracy using structured prompts and auto-labeled datasets. While this line of research established the feasibility of LLM-based solar detection, questions remain regarding their robustness and adaptability across geographically and climatically diverse environments.

This paper specifically addresses the domain generalization and robustness analysis of MLLMs for global solar infrastructure mapping. We analyze how performance degrades with increasing domain shift across diverse geographic regions (See Fig. 1). Through experiments conducted on cities across six continents, we quantify accuracy, calibration, and cross-domain transferability. The results demonstrate that MLLMs maintain high reliability even under significant visual and climatic variations, highlighting their potential as a founda-

tion for globally scalable, interpretable, and data-efficient PV monitoring.

II. Problem Formulation

Let a satellite image be denoted as I. The objective is to construct a mapping function $f(I) \rightarrow \{D, L, Q\}$, where D indicates the binary detection of PV presence, L specifies spatial localization within the image, and Q estimates the panel count or surface area. The function $f(\cdot)$ should generalize across diverse geographical domains and architectural contexts without exhaustive retraining.

In this work, we formulate PV detection as a multimodal inference problem leveraging LLMs conditioned on both imagery and structured textual prompts. The proposed framework employs fine-tuning and few-shot learning re-anchoring to adapt a base LLM for PV infrastructure mapping. To quantitatively assess the model's robustness to domain shift, we introduce the $\Delta F1$ metric, which measures generalization performance across geographical regions.

$$\Delta F1 = F1_{\textit{target region}} - F1_{\textit{training region}}, \tag{1}$$

where $F1_{\textit{target region}}$ and $F1_{\textit{training region}}$ denote F1-Scores on unseen and source domains, respectively. A positive ΔF1 indicates improved transferability, while a negative value reflects performance degradation under domain shift. This metric provides a standardized means to evaluate how well the model generalizes beyond its fine-tuning region.

III. Methodology

This section outlines the proposed framework for cross-domain generalization of MLLMs in global PV mapping. Building on the PVAL foundation, the framework is designed to evaluate model robustness under diverse spatial, climatic, and architectural conditions. It consists of three main components: data engineering, MLLMs configuration, and domain-shift evaluation.

A. Data Engineering

Data engineering ensures the model is trained and evaluated on high-quality, structured, and reproducible datasets. The pipeline includes data collection, image slicing, and annotation. Geographic coordinates of PV installations were obtained from OpenStreetMap (OSM) using the Overpass API[1]. Each region's query, endpoint, and export timestamp were recorded to preserve the exact OSM snapshot. These coordinates anchored the retrieval of satellite imagery from the Google Maps Static API[2], with fixed parameters (`maptype=satellite`, `zoom=20`, and `size=400x400`). To capture a wide range of environmental and architectural conditions, the dataset covers six U.S. regions (Seattle, WA; Orlando, FL; Osage Beach, MO; Harlem, NY; Tempe, AZ; and Santa Ana, CA) and six international cities (Sydney, Australia; Cape Town, South Africa; Kuwait City, Middle East; Oxford, UK; São Paulo, Brazil; and Shanghai, China), supporting cross-regional and cross-continental generalization.

[1]https://overpass-api.de/

[2]https://developers.google.com/maps/documentation/maps-static/overview

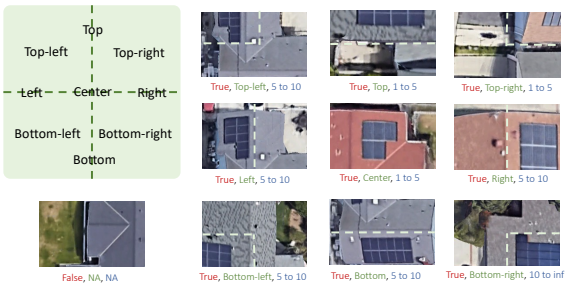

Fig. 2: Solar panel labeling schema and representative examples. The left panel shows nine spatial regions (top, bottom, left, right, center, and four diagonals), plus the "NA" case indicating no solar panels. The right panel displays annotated rooftop tiles with labels specifying (i) presence (True/False), (ii) location, and (iii) quantity range.

Each high-resolution image was divided into a 4×4 grid, yielding 16 tiles per image. This deterministic tiling strategy expanded the dataset while maintaining spatial resolution and ensuring visibility of small-scale PV arrays. It enabled the model to jointly learn presence detection (D), localization (L), and quantity estimation (Q). Each tile was manually annotated by trained PV specialists for solar panel presence, location, and quantity. Fig. 2 illustrates the standardized labeling schema for the three tasks, along with representative annotated examples. This human-in-the-loop process provides high-quality ground truth for model fine-tuning and evaluation [13].

The dataset integrates both U.S. and global imagery from open sources, ensuring transparency and reproducibility. Each image tile is labeled for *solar panel presence*, with additional structured annotations for *location* and *quantity* (intervals: $(0, 1]$, $(1, 5]$, $(5, 10]$, $(10, +\infty)$, or "NA"). These labels provide a standardized and interpretable representation of rooftop PV characteristics across diverse regions. Both baseline models and the PVAL framework were fine-tuned on 2,000 annotated tiles from Santa Ana, CA. The large-scale evaluation set includes about 100,000 tiles from Tempe and Santa Ana, while 480 tiles per region were used for cross-domain generalization tests across diverse climates and geographies.

B. Multimodal LLM Configuration

Configuring the PVAL system for solar panel detection involves a multi-faceted approach that integrates prompt engineering, output standardization, and supervised fine-tuning. This configuration is critical for steering the foundational GPT-4o model towards the specific, high-precision task of geospatial analysis.

The process begins with prompt engineering, which defines the model's operational contract. As detailed in the system prompt (see Prompt III-B), this is a comprehensive guide. It provides "task decomposition" (analysis, location, quantification) and, crucially, enforces "output standardization" by demanding a strict JSON schema with predefined value ranges (e.g., "top-left", "10 to inf"). This schema-guided approach is

reinforced with "few-shot learning" examples, ensuring the model's responses are consistent and machine-readable.

Prompt

Task Decomposition

```
Identify the presence of solar panels in images of residential rooftops, and
determine their locations and quantity within the images. You will be provided
with images that may contain residential rooftop solar systems. Analyze each
image to detect solar panels.
Steps:
1. **Image Analysis**: Examine the entire image to identify any objects that
appear to be solar panels.
2. **Panel Location**: Determine the coordinates or area within the image where
the solar panels are located.
3. **Panel Quantification**: Calculate or estimate the number of solar panels
based on their appearance and arrangement.
```

Output Standardization

```
The output should be in JSON format, structured as follows, with each field
restricted to specific possible values for consistency and accuracy:
"solar_panels_present": A boolean value indicating if solar panels are
detected.
Possible values: [true, false]
"location": A description or coordinates indicating where the panels are
located within the image.
Possible values: [left, right, bottom, top, top-left, top-right, bottom-right,
bottom-left, center, NA]
"quantity": The number of solar panels detected in the image.
Possible values: [0 to 1, 1 to 5, 5 to 10, 10 to inf, NA]
"likelihood_of_solar_panels_present": A value indicating the probability of
solar panels being present.
Possible values: A decimal range from 0.00 to 1.00
"confidence_of_solar_panels_present": A value indicating the model's confidence
in its prediction.
Possible values: A decimal range from 0.00 to 1.00
```

Few-shot Prompting

```
· Example 1 (Solar):
{ "solar_panels_present": true,
  "location": "top-left",
  "quantity": "0 to 1",
  "likelihood_of_solar_panels_present": 0.98,
  "confidence_of_solar_panels_present": 0.90 }
· Example 2 (No Solar):
{ "solar_panels_present": false,
  "location": "NA",
  "quantity": "NA",
  "likelihood_of_solar_panels_present": 0.21,
  "confidence_of_solar_panels_present": 0.87 }
```

This prompt structure is then used as the basis for supervised fine-tuning, the procedure for which is outlined in Algorithm 1. The training dataset is prepared in JSON Lines (JSONL) format, a format well-suited for the OpenAI API. In this dataset, each entry pairs a satellite image (encoded in *base64* [14]) with its corresponding ground truth JSON output, which adheres to the exact schema defined in the prompt.

During the fine-tuning job, we use OpenAI's supervised fine-tuning API with image-conditioned prompts and JSON-formatted targets. The training objective is the standard autoregressive language-model cross-entropy:

$$\mathcal{L}_{\text{LM}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\log p_\theta(y_{i,t} \mid y_{i,<t},\, x_i)\,, \qquad (2)$$

where x_i denotes the input (image + text), $y_{i,1:T_i}$ are the target tokens of the JSON output, and p_θ is the model distribution. This setup leverages GPT-4o's multimodal reasoning while aligning it to the PV assessment task under a schema-constrained JSON output. The PVAL framework features model-agnostic modularity. While this study utilizes a specific multimodal backbone, the underlying "Unified Schema" is compatible with any contemporary VLM. This architecture serves as a reasoning engine that can seamlessly integrate future foundation model advancements, ensuring the methodology remains scalable and adaptable.

C. Evaluation Metrics

The model outputs schema-constrained categorical values rather than pixel masks, so IoU and mAP are inapplicable. We report schema-aligned metrics. For the binary solar-panel presence task, with labels $y_i \in \{0,1\}$ and predictions $\hat{y}_i \in \{0,1\}$, let TP, FP, and FN denote true positives, false positives, and false negatives, respectively: $\text{Precision} = TP/(TP+FP)$, $\text{Recall} = TP/(TP+FN)$, $F1 = 2\cdot\text{Precision}\cdot\text{Recall}/(\text{Precision}+\text{Recall})$. For categorical attributes such as location and quantity, performance is measured by exact-match accuracy, $\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}[\hat{y}_i = y_i]$. Cross-domain performance is summarized by the difference in F1 between the source and target domains: $\Delta F1 = F1_{target} - F1_{source}$.

Algorithm 1: Fine-Tuning Procedure

Input: Training file data.jsonl, base model GPT-4o, API key, hyperparameters: epochs, batch size, learning rate (lr), temperature (T)
Output: Fine-tuned model θ^*
1) Initialize client: client ← OpenAI(api_key)
2) Upload training file:
file_id ← client.files.create(file="data.jsonl", purpose="fine-tune").id
(3) Create fine-tuning job: job ← client.fine_tuning.jobs.create(
training_file=file_id, model=base_model, n_epochs, batch, lr)
4) Monitor job status: while *status* $\notin \{succeeded, failed\}$ **do**
  job ← client.fine_tuning.jobs.retrieve(job.id);
  status ← job.status;
5) On success: θ^* ← job.fine_tuned_model
6) Use fine-tuned model:
response ←
client.chat.completions.create(model=θ^*, messages=[{user prompt}], T=0)
return θ^*

IV. Numerical Results

A. Experiment Setup

Fine-tuning and inference were performed using the OpenAI GPT-4o multimodal architecture via its API. The training dataset, in JSONL format, included paired satellite imagery and structured text prompts (Section III). Fine-tuning for PV detection, localization, and quantification was carried out for five epochs on OpenAI's infrastructure. Experiments for benchmarking and large-scale inference were conducted on the *Sol* supercomputing cluster at Arizona State University, equipped with NVIDIA A100 GPUs, AMD EPYC 7413 CPUs, and 256 GB of system RAM. Baseline models, including U-Net [9], ResNet-152 [15], Inception-v3 [16], VGG-19 [8], and ViT-Base-16 [17], were implemented in Python 3.9 with consistent preprocessing, data augmentation, and batch configurations for fair comparison with the LLM-based framework. These baseline models were trained for ten epochs with binary cross-entropy loss and the Adam optimizer (learning rate $\eta = 10^{-3}$, batch size = 16), and the MLLMs framework was evaluated on the same dataset splits for consistency.

B. Comparative Performance Analysis

The initial phase of our evaluation focused on establishing the baseline performance of the PVAL against a suite of traditional CNN-based and transformer-based models. Take Santa Ana as an example, the fine-tuned PVAL demonstrates superior performance across all standard classification metrics. As illustrated in Fig. 3, it achieves the highest Precision (0.961), Recall (0.946), F1-Score (0.950), and Accuracy (0.946), outperforming established architectures like U-Net, ResNet-152, and ViT-Base-16. To further contextualize these

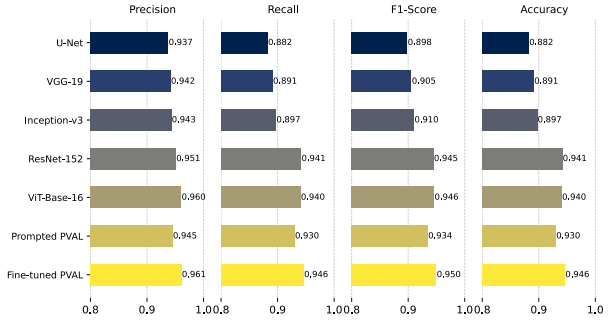

Fig. 3: Comparison of model performance across Precision, Recall, F1-Score, and Accuracy metrics.

results against broader benchmarks, we evaluated the task using foundation models: Florence-2 [18], Llama 3 Vision [19], and CLIP [20], which achieved F1-scores of 0.80, 0.75, and 0.71, respectively. This performance gap underscores that while general-purpose foundation models possess vast world knowledge, the structured prompt engineering and task-specific fine-tuning in PVAL are critical for the sub-meter precision required in global PV assessment.

This superiority, especially in precision, stems from the LLM's underlying architecture. Unlike pure CV models that learn pixel-level patterns from scratch, the MLLMs leverages vast pre-training on world knowledge. It possesses a high-level semantic understanding of constituent concepts (e.g., "roof," "shadow," "skylight," "solar panel"). This allows it to more effectively disambiguate visually similar but semantically different objects, such as mistaking a dark skylight for a solar panel, a common failure case for traditional models. The "Prompted PVAL" baseline's strong showing further underscores this, though fine-tuning is clearly critical for specializing this knowledge.

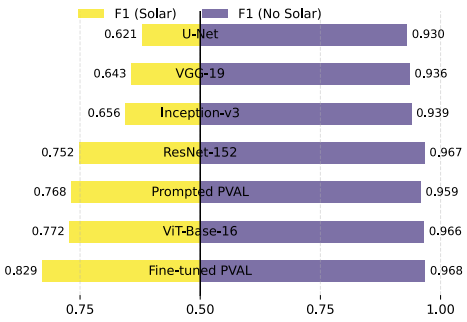

Fig. 4: Comparison of F1-scores for solar and non-solar datasets across different models.

Fig. 4 presents the model's robustness to class imbalance, showing an F1-Score of 0.829 for the "Solar" class, outperforming ViT-Base-16 (0.772). This suggests that traditional CV models overfit to the majority class, while the LLM maintains balanced and robust performance. Fig. 5 shows the PVAL's superior multi-task capabilities, achieving 87.4% accuracy in location detection and 80.6% in quantity estimation. Unlike CV models, which require complex post-processing for

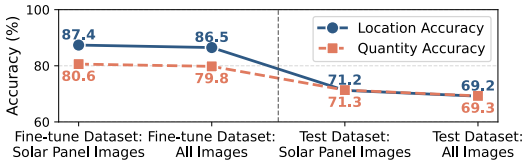

Fig. 5: Accuracy comparison of solar panel location and quantity detection for fine-tuning and test datasets.

these tasks, PVAL generates location and quantity data directly as structured JSON output, demonstrating true reasoning capabilities beyond pixel-wise segmentation.

C. Cross-Domain Performance Analysis

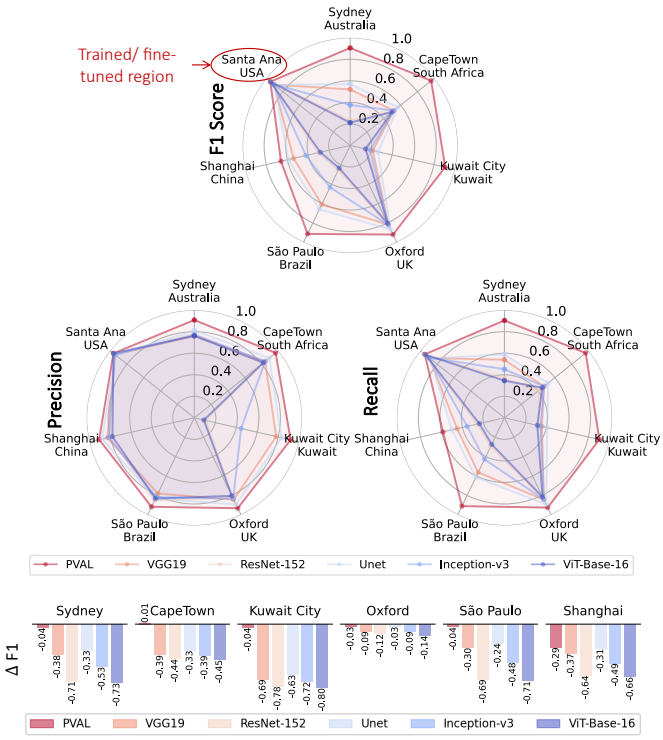

Fig. 6: Top: Radar plots comparing model performance across seven regions for Precision, Recall, and F1-Score. The Santa Ana is the fine-tuning region. Bottom: ΔF1 of each model across six unseen regions, where positive values denote improved transferability and negative values indicate reduced generalization.

A core objective of this research is to assess the domain adaptation and generalization capabilities of the MLLMs for global-scale PV mapping. To this end, we conducted a comprehensive cross-domain evaluation, applying the PVAL exclusively on data from Santa Ana to six unseen regions.

The radar plots in Fig. 6 (top) highlight the model's stability across domains. The PVAL's performance (red line) maintains the broadest and most consistent region for F1-Score, Precision, and Recall, outperforming conventional vision models. In contrast, models such as U-Net and ResNet-152 rely on low-level visual features (e.g., "dark rectangles on light-colored

roofs") that are strongly domain-dependent. When exposed to distinct rooftop materials: lighting conditions, or architectural designs, such as those in Kuwait City or São Paulo (see Fig. 1), these models exhibit sharp degradation due to their limited semantic understanding.

This performance decline is quantified in Fig. 6 (bottom) using the ΔF1 metric. While all models experience performance degradation under domain shift, the PVAL framework's resilience, evidenced by the minimal $\Delta F1$ drop in regions like Sydney (-0.04) and Oxford (-0.03), is rooted in its transition from local texture matching to semantic reasoning. Conventional CV models like U-Net and ResNet-152 rely on brittle, low-level visual cues such as 'dark rectangles on light-colored roofs,' which are highly sensitive to regional variations in roofing materials and illumination. In contrast, the MLLM leverages its extensive pre-training on world knowledge to disambiguate visually similar but semantically distinct objects. For instance, in complex urban environments like Shanghai or Kuwait City, PVAL successfully distinguishes PV panels from dark skylights or HVAC units, features that frequently trigger false positives in traditional models. Furthermore, PVAL's ability to natively output structured JSON for location (L) and quantity (Q) demonstrates a form of 'spatial reasoning' that eliminates the need for the brittle post-processing logic required by pixel-mask architectures. This suggests that the model is not merely matching pixels but is performing high-level semantic interpretation of the solar infrastructure across diverse architectural contexts.

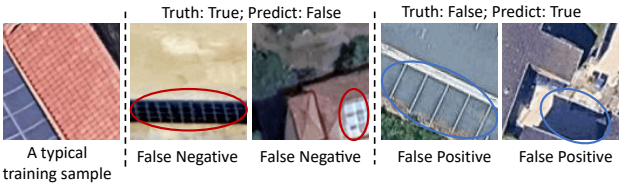

Fig. 7: Example prediction errors of the ViT model. Middle: false negatives (missed PV panels); Right: false positives (non-solar structures misclassified as solar panels).

Fig. 7 presents representative error cases from the ViT baseline. The false negatives demonstrate the model's difficulty in identifying small or partially shaded PV panels, which differ from the clean, high-contrast panels seen during training. The false positives occur when roof features such as skylights or reflective surfaces mimic PV textures, leading to misclassification. These examples highlight the inherent limitations of conventional computer vision models that rely primarily on local texture cues, underscoring the advantage of MLLMs that leverage semantic reasoning to disambiguate visually similar objects across varying domains.

V. Conclusion

This study explored and demonstrated that MLLMs can achieve strong cross-domain generalization for global PV assessment. The PVAL framework, fine-tuned with structured prompts, maintains high accuracy across diverse regions and shows minimal $\Delta F1$ degradation under domain shift. By leveraging semantic reasoning rather than low-level visual cues, the LLM outperforms conventional vision models in robustness and adaptability. These findings highlight the potential of MLLMs as scalable and transferable tools for global PV mapping and monitoring.

References

[1] M. Guo, Q. Cui, and Y. Weng, "Graph mining for classifying and localizing solar panels in distribution grids," in *2023 Panda Forum on Power and Energy (PandaFPE)*. IEEE, 2023, pp. 1743–1747.

[2] H. Li, M. Guo, M. Ilic, Y. Weng, and G. Ruan, "External data-enhanced meta-representation for adaptive probabilistic load forecasting," *arXiv preprint arXiv:2506.23201*, 2025.

[3] H. Li, M. Guo, Y. Weng, M. Ilic, and G. Ruan, "Exarnn: An environment-driven adaptive rnn for learning non-stationary power dynamics," *arXiv preprint arXiv:2505.17488*, 2025.

[4] H. Li, C. Xiao, M. Guo, and Y. Weng, "Latent mixture of symmetries for sample-efficient dynamic learning," *arXiv preprint arXiv:2510.03578*, 2025.

[5] J. Wu, J. Yuan, Y. Weng, and R. Ayyanar, "Spatial-temporal deep learning for hosting capacity analysis in distribution grids," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 354–364, 2022.

[6] J. Yuan, Y. Weng, and C.-W. Tan, "Determining maximum hosting capacity for pv systems in distribution grids," *International Journal of Electrical Power & Energy Systems*, vol. 135, p. 107342, 2022.

[7] M. Chen, S. Ma, Z. Soltani, R. Ayyanar, V. Vittal, and M. Khorsand, "Optimal placement of pv smart inverters with volt-var control in electric distribution systems," *IEEE Systems Journal*, vol. 17, no. 3, pp. 3436–3446, 2023.

[8] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] C. Bouaziz, M. EL Koundi, and G. Ennine, "Integrated solar panel detection and energy production estimation using unet and high-resolution imagery."

[10] S. Luo, Y. Weng, E. Cook, R. Trask, and E. Blasch, "Solar panel identification under limited labels," in *2022 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2022, pp. 1–5.

[11] M. Guo, M. Guo, E. T. Dougherty, and F. Jin, "Msq-biobert: ambiguity resolution to enhance biobert medical question-answering," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 4020–4028.

[12] M. Guo and Y. Weng, "Solar photovoltaic assessment with large language model," *Applied Energy*, vol. 402, p. 126835, 2025.

[13] E. Blasch, É. Bossé, and D. A. Lambert, *High-level information fusion management and systems design*. Artech House, 2012.

[14] S. Josefsson, "The base16, base32, and base64 data encodings," Tech. Rep., 2006.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[17] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, "Florence-2: Advancing a unified representation for a variety of vision tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4818–4829.

[19] J. Chi, U. Karn, H. Zhan, E. Smith, J. Rando, Y. Zhang, K. Plawiak, Z. D. Coudert, K. Upasani, and M. Pasupuleti, "Llama guard 3 vision: Safeguarding human-ai image understanding conversations," *arXiv preprint arXiv:2411.10414*, 2024.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.