Synthesizing Images with Aligned Masks Using Text-to-Image Based Generative AI for Robust PV Segmentation

Hongjun Tan [a, b, c], Zhiling Guo [a, c *], Jiaze Li [d], Yuntian Chen [b, c], Qi Chen [e, b], Junwei Liu [a, c], Haoran Zhang [f], Jinyue Yan [a, c *]

[a] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[b] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang 315200, China
[c] International Centre of Urban Energy Nexus, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[d] Beijing-Dublin International College, Beijing University of Technology, Beijing, China
[e] School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China
[f] China School of Urban Planning and Design, Peking University, 2199 Lishui Rd, Nanshan District, Shenzhen, China

Email address:

hongjun.tan@connect.polyu.hk (H.T.), zhiling.guo@polyu.edu.hk (Z.G.), lijiaze@emails.bjut.edu.cn (J.L.), ychen@eitech.edu.cn (Y.C.), chenqi@cug.edu.cn (Q.C.), junweei.liu@polyu.edu.hk (J.L.), h.zhang@pku.edu.cn (H.Z.), j-jerry.yan@polyu.edu.hk (J.Y.)

**Corresponding authors: zhiling.guo@polyu.edu.hk; j-jerry.yan@polyu.edu.hk (J.Y.)*

Synthesizing Images with Aligned Masks Using Text-to-Image Based Generative AI for Robust PV Segmentation

Hongjun Tan [a, b, c], Zhiling Guo [a, c *], Jiaze Li [d], Yuntian Chen [b, c], Qi Chen [e, b], Junwei Liu [a, c], Haoran Zhang [f], Jinyue Yan [a, c *]

[a] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[b] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang 315200, China
[c] International Centre of Urban Energy Nexus, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[d] Beijing-Dublin International College, Beijing University of Technology, Beijing, China
[e] School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China
[f] China School of Urban Planning and Design, Peking University, 2199 Lishui Rd, Nanshan District, Shenzhen, China

Email address:

hongjun.tan@connect.polyu.hk (H.T.), zhiling.guo@polyu.edu.hk (Z.G.), lijiaze@emails.bjut.edu.cn (J.L.), ychen@eitech.edu.cn (Y.C.), chenqi@cug.edu.cn (Q.C.), junweei.liu@polyu.edu.hk (J.L.), h.zhang@pku.edu.cn (H.Z.), j-jerry.yan@polyu.edu.hk (J.Y.)

**Corresponding authors: zhiling.guo@polyu.edu.hk; j-jerry.yan@polyu.edu.hk (J.Y.)*

Abstract

The robust detection of photovoltaic (PV) panels from multi-region and multi-background has traditionally relied on aligned remote sensing imagery with manually labeled annotations, especially in supervised learning. However, it faces challenges in training dataset collection, including large volumes, accessibility concerns, and quality inconsistencies. To efficiently produce paired PV images and masks, this study introduces the SynthPV, a Text-to-Image-based Generative AI (GenAI) approach that incorporates diverse backgrounds. By training on less than 1% of real-world data, it can generate highly aligned PV masks simultaneously. The experimental results obtained from three distinct scenarios using the Heilbronn, Jiaxing, and BDAPPV datasets demonstrate that synthetic images and mask pairs could significantly enhance PV segmentation performance. Notably, the Intersection over Union (IoU) values increase by over 13.09% on average and by 20.55% for the Jiaxing dataset in particular. Cross-validation using synthetic data from Heilbronn with real datasets from Jiaxing and BDAPPV shows a further IoU improvement compared to solely real data. These findings underscore the effectiveness and robustness of the proposed GenAI method, offering a feature-adaptive and integrated approach to visual data augmentation that significantly enhances PV segmentation accuracy, paving the way for more efficient and scalable applications in solar energy system analysis and deployment.

Keywords: GenAI; Mask Generation; Text-to-Image; SynthPV; PV Segmentation; Generalizability

Highlights

- Utilize the Text-to-Image based GenAI method to enhance the PV dataset.
- Propose a SynthPV network for generating aligned PV images with masks.
- Demonstrate the robustness of synthetic datasets in PV segmentation.
- Explore the potential of SynthPV for adaptability and generalization.

1. Introduction

1.1. Background

The urgency of addressing global climate challenges—such as rising carbon emissions, population growth, and the degradation of terrestrial and marine ecosystems—has underscored the critical need for renewable energy solutions and international collaboration. Renewable energy development has shown remarkable progress globally, with significant potential for capacity expansion in the coming years. This positions renewables to become the dominant energy source, particularly solar photovoltaic (PV) energy. The global PV installation capacity is projected to reach 1954.6 GW by the end of 2024, with its share expected to rise to 12.6% by 2028. This growth could contribute over 40% to the global renewable energy portfolio, highlighting its pivotal role in the energy transition (1, 2).

As solar energy adoption accelerates, accurately detecting and assessing PV panels has become increasingly important. PV installations are now found on building rooftops, water bodies, grasslands, mountains, and roadsides, making it essential to evaluate solar potential across diverse environments. Traditional methods for PV panel detection and energy capacity assessment often suffer from limited precision, data accessibility issues, and acquisition challenges (3). The widespread application of remote sensing tools in photovoltaic (PV) module identification has increasingly revealed their inherent data limitations (4, 5). Although Deep Learning (DL) techniques demonstrate superior accuracy in processing remote sensing imagery, they encounter significant obstacles in practical implementation. A major constraint lies in their data dependency, as these

advanced models demand extensive collections of precisely labeled remote sensing data, which are challenging to acquire and prepare (6).

The process of collecting such data is not only time-consuming and costly but also hindered by issues of data diversity. Satellite PV images often feature complex backgrounds and low resolution, further complicating data collection and processing (7). Additionally, technological constraints and a lack of diverse PV samples limit the effectiveness of DL methods in detecting urban PV modules (8, 9). The sensitivity of these models to heterogeneous environments also makes it challenging to transfer detection algorithms across different urban areas (10). Moreover, the resource-intensive nature of image annotation restricts the scalability and generalizability of these models. To address these limitations, there is a growing need for more adaptable and efficient data collection methods that enhance algorithm robustness and applicability.

In this context, General Generative AI (GenAI) has emerged as a promising solution to improve PV data acquisition and processing (11). By generating synthetic images and corresponding masks, GenAI enables DL models to train on large-scale synthetic datasets, eliminating the need for manual labeling. These models have demonstrated significant success in generating time-series data (12) and medical imagery (13), offering a pathway to overcome the limitations of remote sensing data. GenAI enhances the accuracy, generalization, and robustness of algorithms while addressing challenges related to data diversity and accessibility. Furthermore, GenAI supports local data processing, ensuring data privacy and accessibility, which is particularly valuable for urban planners and developers. This technology can aid in strategizing PV installations and optimizing renewable energy resource management, ultimately contributing to a more sustainable energy future.

1.2. Related Works

Remote sensing techniques have become indispensable tools for detecting and segmenting photovoltaic (PV) images, playing a pivotal role in advancing PV sector applications. Chen et al. (5) highlighted the extensive use of remote sensing in PV segmentation, several studies have explored the application of advanced computational techniques for photovoltaic (PV) system analysis using remote sensing data. Research efforts (14, 15) have successfully demonstrated deep learning's capability in distinguishing PV installations from diverse backgrounds across multiple satellite imagery sources. Parallel investigations (16) (17) have revealed the effectiveness of combining deep learning architectures with multi-resolution remote sensing data for

extracting information from various PV system configurations, confirming the cross-resolution applicability of pre-trained PV segmentation models. Additionally, the scientific community has leveraged machine learning approaches, particularly in analyzing aerial imagery for PV panel segmentation, as evidenced by study (18). Notable examples include DeepSolar (19), Deep Solar PV Refiner (20) , TransPV (21), and PVNet (22).

To improve the detection and segmentation of photovoltaic (PV) systems at a city scale with high accuracy, researchers have introduced a range of innovative approaches. These include the development of diverse network architectures equipped with various encoders (15). Additionally, advancements such as refined network designs (23), the integration of attention mechanisms to minimize background interference, and the implementation of sophisticated loss functions to tackle class redundancy (24) have significantly boosted the performance of these models. These advancements have improved the detection of multi-level characteristics on urban surfaces, aiding in city-scale PV module identification.

Despite these advancements, deep learning-based methods still face significant challenges, particularly their reliance on large volumes of labeled PV data. Paletta et al. (25) discussed issues related to cloud masks in sky images obtained through thresholding methods for solar potential evaluation. Collecting and labeling large-scale datasets with pixel-level annotations remains time-consuming and costly (26) exacerbating challenges in data accessibility and manual labeling efforts for PV potential assessments. To mitigate these limitations, researchers have proposed various solutions. Kasmi et al.(27) introduced a composite dataset combining aerial images, segmentation masks, and installation metadata. Zech et al. explored deep active learning to address limited PV system information, while Chen et al. (28) utilized generative adversarial networks (GANs) (29) to create renewable energy scenarios based on weather conditions. Wang et al. (30) demonstrated GANs' ability to generate high-quality samples for weather classification and PV power forecasting. Dong et al. (31) developed a data-driven GAN-based model with interpretable latent space features for generating renewable scenario patterns. Wen et al. (32) employed solar irradiance maps to train multi-scale GANs for image-based forecasting. Zhang et al. (12) addressed the lack of 3D models by using fisheye images derived from 3D models as categorical shading masks in deep generative networks (DGNs).

Some studies have further enhanced data utility by combining imagery with time-series data. Paletta et al. (33) found that combining numerical data with images could improve the accuracy of solar irradiance forecasting. Similarly, the DGN-based model (12) was proposed to generate stochastic hourly solar irradiance time series for urban building facades. In another study, Jamie et al. (34) developed a synthetic PV irradiance time series data at a minute granularity using average hourly meteorological data. Wen et al. (35) employed generative models to create solar irradiance maps (SIMs) for entire regions, supporting the use of distributed PV systems. Furthermore, Wang et al. (36) introduced a GAN and CNN-based weather classification model that enriched training datasets for diverse weather conditions and improved day-ahead PV power forecasting.

For creating large training datasets from limited data, the Generative AI has emerged as a powerful tool. Song et al. (26), utilized tools like Blender (37), Python, GPT-4 (38), and Stable Diffusion (39) to generate a synthetic remote sensing 3D dataset featuring six global city styles and eight land cover types. The score-based diffusion model (40, 41) was utilized to forecast day-ahead solar irradiance, integrating atmospheric factors to deliver detailed probabilistic weather predictions. Meanwhile, Zhu et al. (42) developed an innovative scenario generation approach by merging conditional diffusion models (43) with few-shot learning, enabling the effective use of non-extreme historical data and overcoming limitations in training datasets for weather classification tasks. Yuan et al. (44) demonstrated the superiority of diffusion models over VisualFormer (45) and CycleGAN (46) in generating remote-sensing fake samples. Nie et al. (47) proposed SkyGPT, a deep generative model for generating future sky images from past sequences to aid probabilistic solar forecasting.

The Stable Diffusion model (48), developed by Stability AI, is a text-to-image generator that shows significant potential in transforming the approach to training data augmentation for image segmentation tasks (11). Its application could notably alleviate the resource-intensive demands typically linked with data labeling processes. Lin et al. (49), discussed various PV data generation scenarios, highlighting the advantages of general generative AI, including faster deployment, reduced computational costs, enhanced adaptability, and reduced reliance on large datasets and labeled masks. This study explores the simultaneous generation of PV images and aligned masks to improve PV module detection in urban environments.

1.3. Objective and Contribution

To achieve robustness and accuracy, PV segmentation models based on deep learning necessitate extensive datasets. However, the process of data collection faces significant challenges, including limited scope, accessibility issues, and diverse backgrounds, which hinder the models' generalization and optimization. Additionally, creating labeled data for validation is a complex, costly, and time-intensive task (50).

This paper introduces SynthPV, a Text-to-Image based General Generative AI framework, aimed at generating comprehensive urban PV data and aligned annotations to enhance PV segmentation accuracy. By increasing the dataset's diversity and quantity, this method reduces the costs associated with data labeling and preprocessing while improving the adaptability and robustness of detection algorithms across different urban contexts. The main contributions of this paper could be concluded as follows:

(1) The SynthPV is introduced to generate PV masks simultaneously with image generation, and its effectiveness is evaluated through transformative learning techniques.

(2) The work leverages a Text-to-Image-based GenAI framework to generate remote-sensing image pairs, significantly enhancing the diversity of the PV dataset.

(3) By training with a combination of generated and integrated data from three datasets, the approach achieves superior segmentation precision compared to using real data alone.

(4) The generalizability of the generated data is verified through cross-validation with independent datasets.

1.4. Research Framework

Figure 1 illustrates the research framework of this study, addressing challenges like data scarcity, diversity, labeling costs, and accessibility concerns in PV segmentation. To overcome these gaps, the framework incorporates SynthPV, a Text-to-Image-based GenAI model, capable of generating diverse datasets with thousands of synthetic images and their corresponding masks. The Mask Decoder is retrained using real-world data to adapt to urban-specific features, enabling the generation of highly aligned masks integrated with PV images. By focusing on three representative cities, the framework highlights the importance of local characteristics in guiding the generation of 20,000 PV image pairs per city, ensuring synthetic data aligns closely with real-world conditions. This iterative workflow enhances segmentation accuracy and scalability, addressing limitations in traditional PV detection methods.

The structure of the paper is as follows: Section 1 provides the background, reviews related works, and highlights the study's contributions. Section 2 introduces the study areas and data processing procedures, offering a statistical summary of the training and validation datasets. Section 3 details the design of the proposed SynthPV network and the Mask Decoder framework. Section 4 describes the experimental setup, analyzes the results of the proposed approach, discusses the method's ability to distinguish between PV and non-PV-related features, and evaluates its text perception capabilities. Section 5 concludes this work and highlights its potential for broader solar energy analysis and deployment applications.

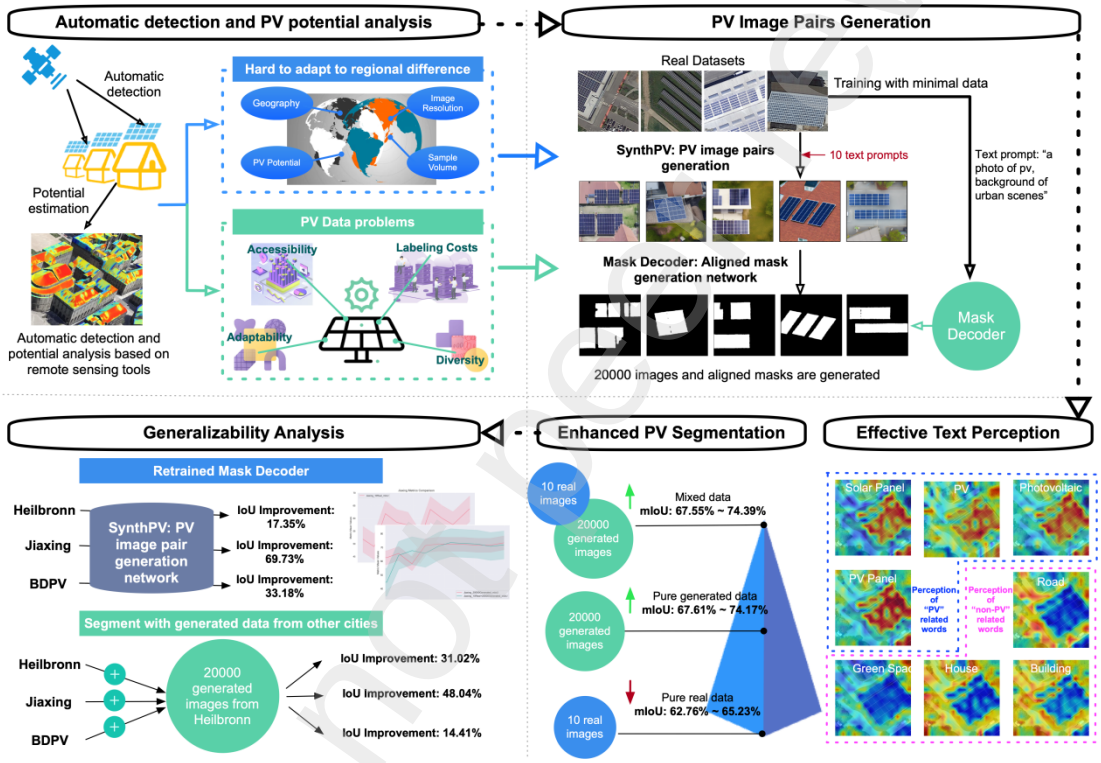


Figure 1. The research framework for enhancing PV segmentation using SynthPV and Mask Decoder networks across multi-regional datasets.

2. Data and Materials

2.1. Study Aera

This study adopts three datasets from diverse regions to investigate the transformability and generalizability of synthetic datasets (**Figure 2**). The first dataset, from Heilbronn, Germany, represents European PV systems in temperate zones, characterized by diverse terrain and high-resolution imagery (0.15 m/px) with 512 × 512 tiles. The second dataset, from Jiaxing, China, exemplifies dense urban PV adoption in subtropical climates, offering moderate resolution (1.34 m/px) and 512 ×

512 tiles. The third dataset, BDAPPV (27), primarily covers France and Western Europe, showcasing PV systems in coastal plains and areas with variable weather patterns, with a lower resolution (2.0 m/px) and 400 × 400 tiles. **Table 1** shows the detailed differences between the three cities. During the training process, ten real images of each dataset are selected to compare with the generated image pairs training images. SynthPV is retrained using 50 real images, and 20,000 image pairs are generated with the Text-to-Image network, providing a diverse testing ground for PV segmentation methods across varying geographical and installation conditions. The contrasting characteristics of these regions enhance the study's scope, offering valuable insights into PV installation trends and solar energy distribution.

Table 1. The datasets used in this paper and their features.

Location	Geography	PV Potential	Number of all images (Real data training)	Number of Mask Decoder training images	Number of generated image pairs	Image resolution [m/px]	Image tiles resolution [px]
Heilbronn, Germany	Eastern China	Growing PV installations	27945 (10)	50	20000	0.15	512 * 512
Jiaxing, China	Southern Germany	Pioneer in PV installation,	5484 (10)	50	20000	1.34	512 * 512
BDAPPV	Mainly in France and Western Europe	Southern regions with high solar irradiation	13300 (10)	50	20000	2.0	400 * 400

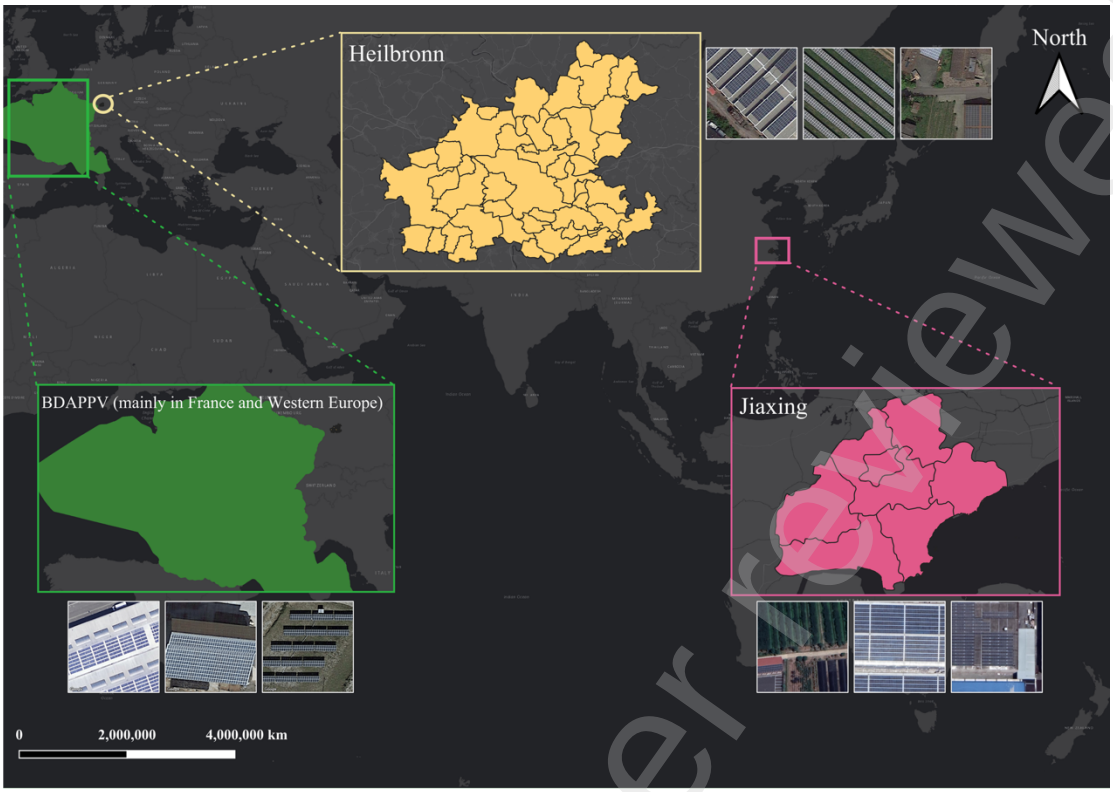


Figure 2. The satellite imagery of the three cities.

2.2. Data Selection

The SynthPV utilizes 50 original PV arrays, representing less than 1% of the original dataset, with diverse background elements to pre-generate training data. PV modules paired with varied environments such as buildings, grasslands, and water bodies are incorporated. **Figure 3** illustrates representative images from the three regions, encompassing varied backgrounds such as urban rooftops, industrial zones, and natural environments. This pre-generated training approach enables the stable diffusion model to better distinguish between the binary objectives, PV, and Background, leading to highly accurate annotations.

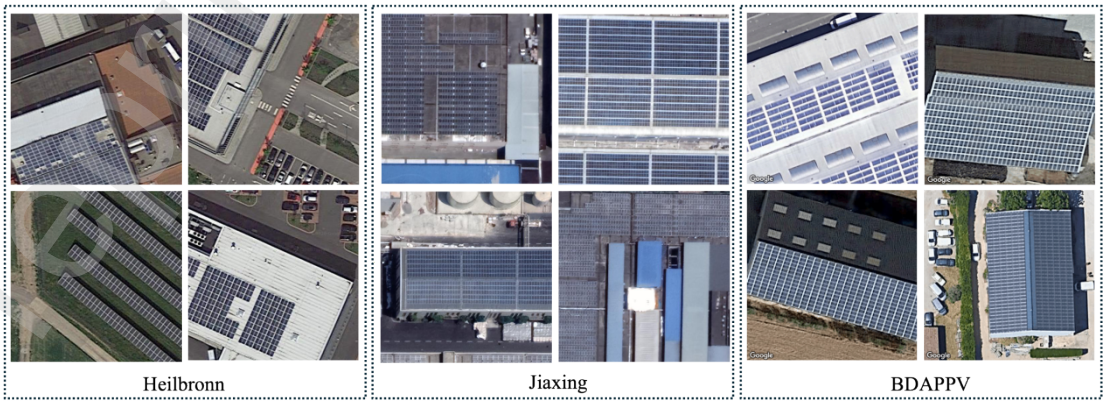


Figure 3. Samples of selected pre-generating training dataset.

The PV datasets consist of binary semantic labels identifying PV installations and non-PV areas (20). To prepare the data for model training, the images were divided into 512 × 512-pixel tiles with a 0.25 overlap ratio from tiff satellite imageries of Heilbronn and Jiaxing. The annotation process applied a consistent cropping approach to ensure uniformity across datasets. BDAPPV dataset contains with open-source images and annotations. In total, this experiment datasets include 27945 tiles for Heilbronn City, 5484 for Jiaxing City, and 13300 for BDAPPV.

3. Method

In this section, the detailed design of the method is explained. As depicted in **Figure 1**, this study introduces the SynthPV network, a novel approach designed to address the challenges of PV data scarcity and regional adaptation in detection methods. The SynthPV network generates paired images and masks to enable more effective downstream tasks. Section 3.1 details the SynthPV network, while the Mask Decoder is discussed in the subsequent section. Rather than focusing solely on image generation, the approach prioritizes the seamless creation of image pairs for downstream tasks. The image component is synthesized using a text-to-image model, while the mask is generated by training a specialized Mask Decoder to align with the corresponding image.

3.1. Synthesizing PV Image Pairs with SynthPV Network

The Stable Diffusion model operates as a conditional latent diffusion framework within the realm of Generative AI. It utilizes a pre-trained text encoder to embed textual prompts and employs a Variational Autoencoder (VAE) to enhance memory and computational efficiency. As depicted in **Figure 4**, the SynthPV network demonstrates the generation of images and masks. Within this framework, the VAE employs an encoder $\mathcal{E}$ to compress image inputs into a latent space, while a decoder $\mathcal{D}$ reconstructs images from denoised latent representations, leading to substantial optimization of resource utilization. Furthermore, the model integrates a denoising U-Net ϵ_θ, augmented with transformer blocks and cross-attention layers. This enhanced U-Net accurately predicts noise residuals for image generation and effectively handles various multimodal inputs, including text, segmentation maps, and edges.

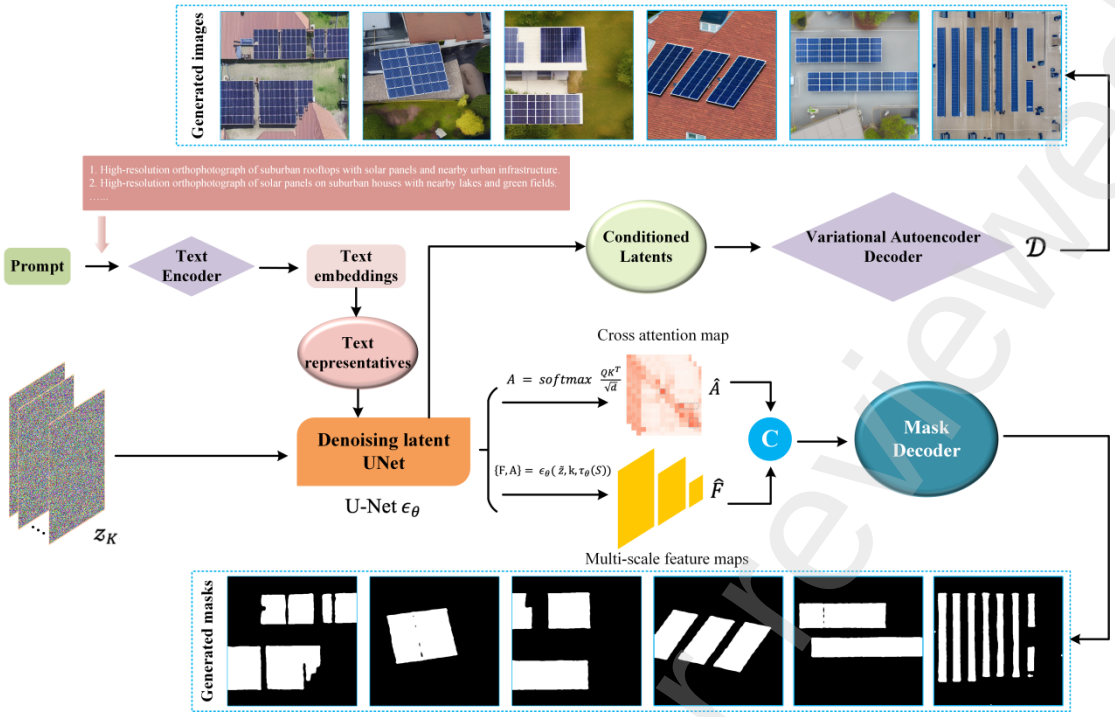


Figure 4. Detailed SynthPV network, a Text-to-Image based Generative AI model, for generating paired PV images and masks, and illustrated with five reference examples.

The inference pipeline is designed for text-guided data generation. To facilitate this process, a conditioner τ_θ is employed to map conditioning prompts y to the intermediate layers of the U-Net. This is achieved by projecting y into an intermediate representation, denoted as $\tau_\theta(y) \in R^{M \times d_\tau}$, which is then integrated into the U-Net through cross-attention layers. During the inference phase, the latent representative $\tilde{z}$ is obtained by progressively subtracting the predicted noise from the given image, mask, and text inputs through the diffusion model. The resulting image is reconstructed using a latent VAE decoder $\mathcal{D}$, expressed as $\tilde{x} = \mathcal{D}(\tilde{z})$. To generate a diverse dataset, ten text prompts were used to produce 20,000 PV image pairs. As illustrated in **Figure 4**, the generated images not only retain critical features but also achieve higher accuracy in extracting target annotations compared to the original reference images. This demonstrates the effectiveness of the pipeline in preserving essential details while enhancing annotation precision.

Mask generation, particularly in the context of solar energy applications, has received limited attention from researchers compared to image inference synthesis. However, recent advancements in Stable Diffusion models have sparked growing interest in this area. Several studies have explored the potential of these models for mask-generation tasks. Zhao et al. (51) developed the Visual Perception with a pre-trained Diffusion model (VPD), which utilizes pre-trained knowledge of objects (e.g.,

cats and tables) within the denoising UNet to offer semantic guidance for visual perception tasks. Building on this foundation, Wu et al. (52) expanded the application of pre-trained diffusion models to text-to-image synthesis for data generation. Further advancements were made with the introduction of the Dataset Diffusion network (53), which integrates three key components: class-prompt appending, class-prompt cross-attention, and self-attention exponentiation. A notable feature of this approach is its use of caption prompts from the COCO and VOC datasets to improve the annotation generation process, distinguishing it from the earlier DiffuMask (51).

Significant progress has been made in mask generation through various studies. However, these existing methods face specific limitations when applied to solar panel dataset generation. The unique characteristics of PV panel datasets, particularly their binary nature, require two critical capabilities: the generation of diverse backgrounds and precise perception of PV panel structures – aspects that have not been adequately addressed in prior research. To overcome these limitations, our study proposes a novel two-stage approach. First, we implement pre-generation training to develop an enhanced mask synthesizing network, referred to as the Mask Decoder. This network is specifically designed to handle the challenges of PV panel data. Subsequently, we employ supervised image and mask generation techniques to ensure the production of high-quality, comprehensive masks for PV datasets while substantially enhancing data diversity. The proposed methodology integrates the Mask Decoder with a text-to-image generation model, creating a robust framework for targeted mask-generation tasks. This integration enables the generation of diverse datasets tailored to specific requirements. The following section will provide a detailed explanation of the Mask Decoder's architecture and design principles.

3.2. Mask Decoder Network Based on Diffusion Model

Figure 5 presents the Mask Decoder network, which employs a diffusion model to manipulate latent space for generating PV image pairs. The process begins by transforming the input image $x \in \mathbb{R}^{H\times W\times 3}$ into its latent representation $z = \mathcal{E}(x) \in \mathbb{R}^{h\times w\times c}$. In the forward phase, Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$ is progressively introduced to z, producing a noised version z_K. During the reverse phase, the model learns to remove this noise step-by-step, reconstructing the original image through a fixed Markov Chain of length K. The diffusion model's learning process can be described as follows:

$$L_{SD} := \mathbb{E}_{z,\ \epsilon\sim\mathcal{N}(0,1),}\left[\ \| \epsilon - \epsilon_\theta(z_\hbar, \hbar) \|_2^2\right] \quad (1)$$

The decoding process leverages both the mask and the latent representation of the masked image for conditioning. During inference, the latent representation $\tilde{z}$ is derived from the diffusion model by iteratively reducing the predicted noise using the provided image, mask, and text. With the latent representative $\tilde{z}$ representing the real image and the corresponding text prompt S, multi-scale feature maps and cross-attention maps are extracted from the U-Net ϵ_θ. Here, $\hbar$ is uniformly sampled from the set $\{1, ..., K\}$.:

$$\{\mathrm{F}, \mathrm{A}\} = \epsilon_\theta(\tilde{z}, \hbar, \tau_\theta(S)) \quad (2)$$

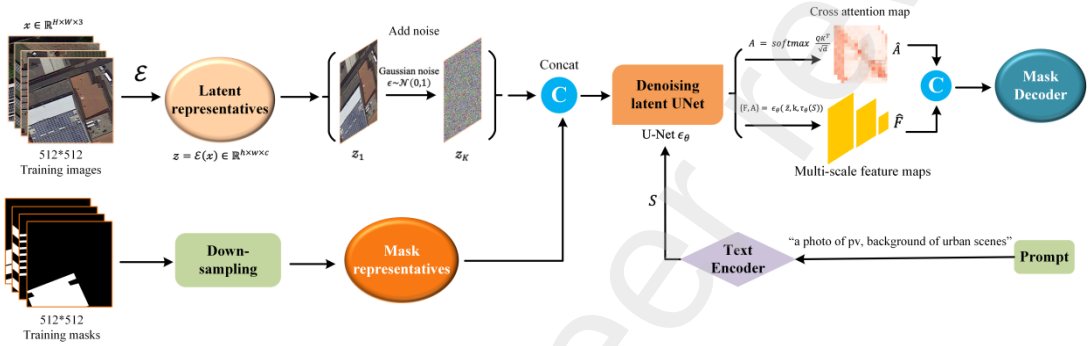


Figure 5. Detailed Mask Decoder network with less than 1% real-world PV data.

Where S for the training set is defined using the text prompt "a photo of pv, the background of urban scenes." The multi-scale feature maps, denoted by F , are derived from four distinct layers of the U-Net, each corresponding to a specific resolution: 8 × 8, 16 × 16, 32 × 32, and 64 × 64. On the other hand, A represents the cross-attention maps generated through text-to-image interactions across the 16 cross-attention layers in the U-Net. These maps are computed using the formula $A = \text{softmax}\ \frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d}}}$, where d stands for the latent projection dimension. The cross-attention maps are then organized into four groups based on their resolutions, and the average is calculated within each group, yielding the average cross-attention maps $\hat{\mathrm{A}}$. To obtain the final feature representation, the multi-scale feature maps F are concatenated with the average cross-attention maps $\hat{\mathrm{A}}$. This combined representation is further processed using a 1 × 1 convolution to fuse the features, resulting in the final output: $\hat{F} = Conv([F, \hat{\mathrm{A}}])$.

The Mask Decoder translates the feature representative $\hat{F}$ into perception annotations that are suitable for different downstream tasks. To segment PV from the background, the Mask Decoder employs a dual-component architecture: the pixel decoder and the transformer decoder. The pixel decoder, consisting of multiple up-sampling CNN layers, generates per-pixel embeddings. Meanwhile, the transformer decoder, which includes a series of transformer layers with cross-attention and self-

attention mechanisms, refines the queries and produces the final outputs. By taking the representation $\hat{F}$ and a set of K learnable queues $\{Q_0, Q_1, ..., Q_K\}$ as input, the Mask Decoder generates K binary masks $M = m_1, m_2, ..., m_K \in \{0,1\}^{K \times h \times w}$. This is accomplished by performing a straightforward matrix multiplication between the outputs of the transformer decoder and the pixel decoder.

3.3. Segmentation Model

The training process for segmentation tasks often involves managing extensive parameters and high computational costs, particularly when utilizing enhanced networks for refined PV segmentation. Among the various architectures, U-Net (54) is frequently selected due to its balance between accuracy, efficiency, and scalability. With a parameter range of approximately 7–31 million, it offers a manageable solution in terms of memory usage and computational demands.

In contrast, transformer-based models like SegFormer (55) deliver high accuracy but come with significantly larger parameter counts. For example, SegFormer B5 contains around 84 million parameters, with even larger versions further increasing the computational burden. U-Net's encoder-decoder design, featuring skip connections, stands out for its lightweight and efficient structure. This architecture excels in tasks requiring detailed feature extraction and spatial detail preservation, making it particularly suitable for segmenting objects with well-defined boundaries, such as PV panels.

A key advantage of U-Net is its ability to reduce computational complexity and parameter count in the decoder, optimizing both efficiency and processing speed. Within the MMSegmentation framework, U-Net's operations complexity mainly relies on convolutional and pooling layers, which are computed as follows:

$$FLOPS = 2 \times H \times W \times K^2 \times C_{input} \times C_{output} \quad (3)$$

Segformer employs self-attention layers, which have a computational complexity of Formula (4), where the cost scales with input size, making it less suitable for limited-resource environments typical of PV segmentation tasks.

$$FLOPS = 4 \times H \times W \times ({C_{input}}^2 + H \times W) \quad (4)$$

where H and W are the spatial dimensions, K is the kernel size, C_{input} is the input channels, and C_{output} is the output channels.

Our experiment seeks to evaluate the effectiveness of generated data in improving model performance under resource-constrained conditions. Specifically, we

demonstrate how generated data significantly enhances the accuracy of PV segmentation, delivering notable relative improvements without the high computational costs associated with larger architectures. To this end, this work utilizes the U-Net network as our segmentation framework, enabling comprehensive ablation studies to assess the generalizability and impact of the generated datasets.

3.4. Evaluation Metrics

To assess the performance of our network, a range of widely used classification metrics are utilized, including Intersection-over-Union (IoU), Precision, and Recall. These metrics provide a quantitative measure of accuracy, with higher values reflecting better results. IoU, in particular, evaluates the overlap between predicted outputs and ground truth labels. Additionally, a confusion matrix is used to delve deeper into the results, summarizing False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN).

Given that this study involves two classes, PV and background, the mean values of the metrics (mIoU, mPrecision, and mRecall) are calculated by averaging across the classes. Metrics nearing a value of 1.0 suggest that the model achieves close to ideal performance. For a more detailed understanding, the formulas used for these evaluation metrics are presented below:

$$IoU = \frac{area(PredictionResult \cap GroundTruth)}{area(PredictionResult \cup GroundTruth)} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

4. Results and Discussion

4.1. Synthesizing Data Analysis

This section comprehensively analyzes the quality of generated data by interpreting text prompts and highlighting key features, for example, PV, urban, etc. Through the visualization of attention maps, the SynthPV network could adaptively generate aligned images and masks simultaneously.

4.1.1. Generated Data Visualization

This work leverages the SynthPV network, integrating Mask Decoder and Text-to-Image generation model to create diverse datasets and targeted mask generation. Ten text prompts are used to generate 200k PV image samples, demonstrating the model's ability to synthesize realistic images and aligned masks. They are retrained to learn the local features of different cities and avoid image inconsistency. As is shown in Figure 6, the generated images and aligned masks are illustrated. To better compare with the true labels, manual labeling is conducted to test the competence of synthetic masks created by SynthPV. It is highlighted that synthetic masks could match the manually labeled masks over 80% by dividing the pixel number of the masks (b) by masks (c). **Figure 6** (d) showcases the overlapped images with mismatched pixels highlighted in red color, the uncovered gap areas mean that all generated masks could relatively cover the PV area, therefore, the SynthPV is effective enough to meet the labeling accuracy in segmentation tasks.

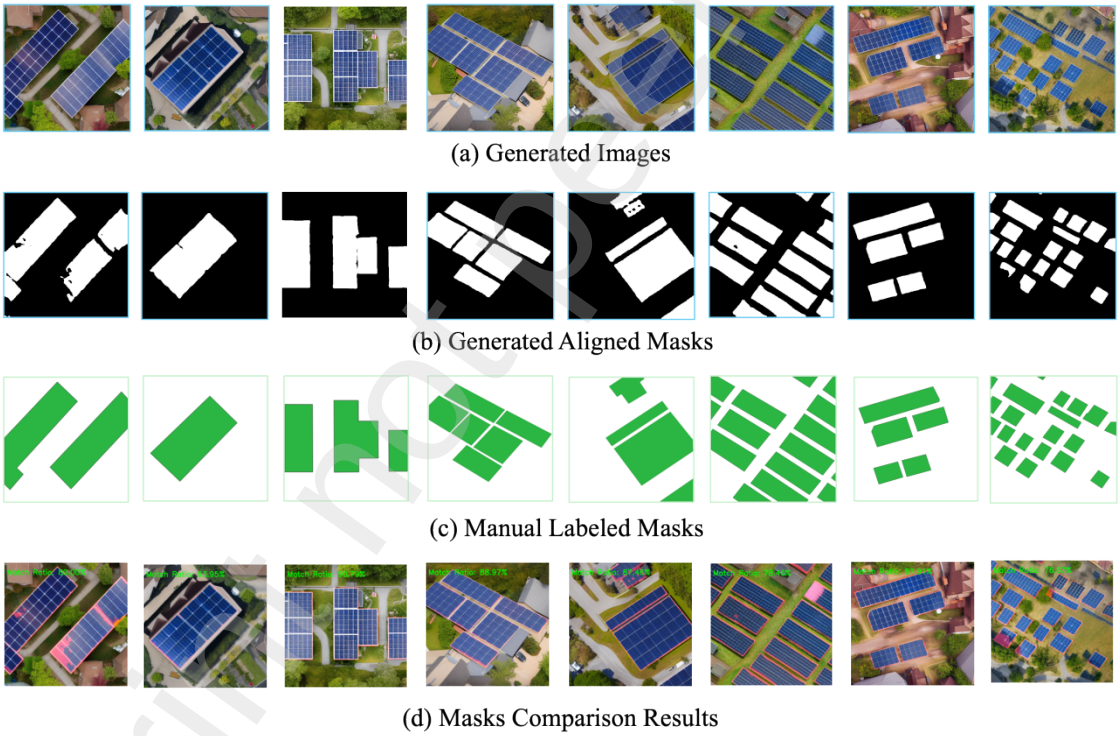


Figure 6. The (a) generated reference images, (b) aligned masks, (c) manual labeled masks, and (d) mask comparison results with match ratio.

4.1.2. Text Perception Analysis Based on CLIP Surgery

To evaluate the quality of the generated images, the CLIP Surgery method (56) was employed to visualize similarities between image data and textual descriptions. Utilizing the ViT-B/16 backbone within the CLIP architecture ensured robust feature

representation for images and texts. All input images were resized to 224×224 pixels to comply with the CLIP model's input specifications.

Building upon the standard CLIP model, the approach integrates several key modifications to enhance explainability. Firstly, consistent self-attention is implemented by utilizing homogeneous parameters in the self-attention layers, promoting semantically coherent relationships between image tokens. This ensures that the attention mechanism maintains consistent semantic connections across different regions of the image. Additionally, a dual paths architecture is adopted, bypassing Feed-Forward Networks (FFNs) in deeper layers to maintain feature affinity and mitigate the adverse effects of FFNs on explainability. Furthermore, CLIP Feature Surgery is introduced, a technique designed to eliminate redundant features that contribute to noisy activations in the similarity maps. This process involves identifying and removing features that do not significantly contribute to target class recognition by averaging feature activations across classes and suppressing those that do not align with the target text prompts, resulting in cleaner and more accurate similarity visualizations.

Text descriptions representing various positive features, such as "solar panels," "buildings," and "roads," were selected and encoded alongside images into a shared feature space. The similarity between each image and text prompt was calculated using cosine similarity, and regions exhibiting high similarity were highlighted using a color map, where the "red" color shows the text token with the highest similarity, "yellow" is the second, and "blue" means the last. The final similarity maps were overlaid on the original images, providing clear and interpretable visual explanations of the model's associations between textual concepts and image regions.

From **Figure 7**, the texts related to "PV," "Solar Panel," "PV Panel," and "Photovoltaic" were effectively perceived, with the highest probability areas correctly located on the PV panel regions. This demonstrates the robustness of the CLIP Surgery method in aligning textual descriptions with the corresponding image regions, further validating the accuracy of SynthPV in generating semantically consistent masks. The integration of homogeneous self-attention parameters and the dual-path architecture ensures semantically coherent attention distribution, enhancing the explainability and precision of similarity maps. This supports the conclusion that SynthPV is a reliable tool for capturing image-text associations with high fidelity.

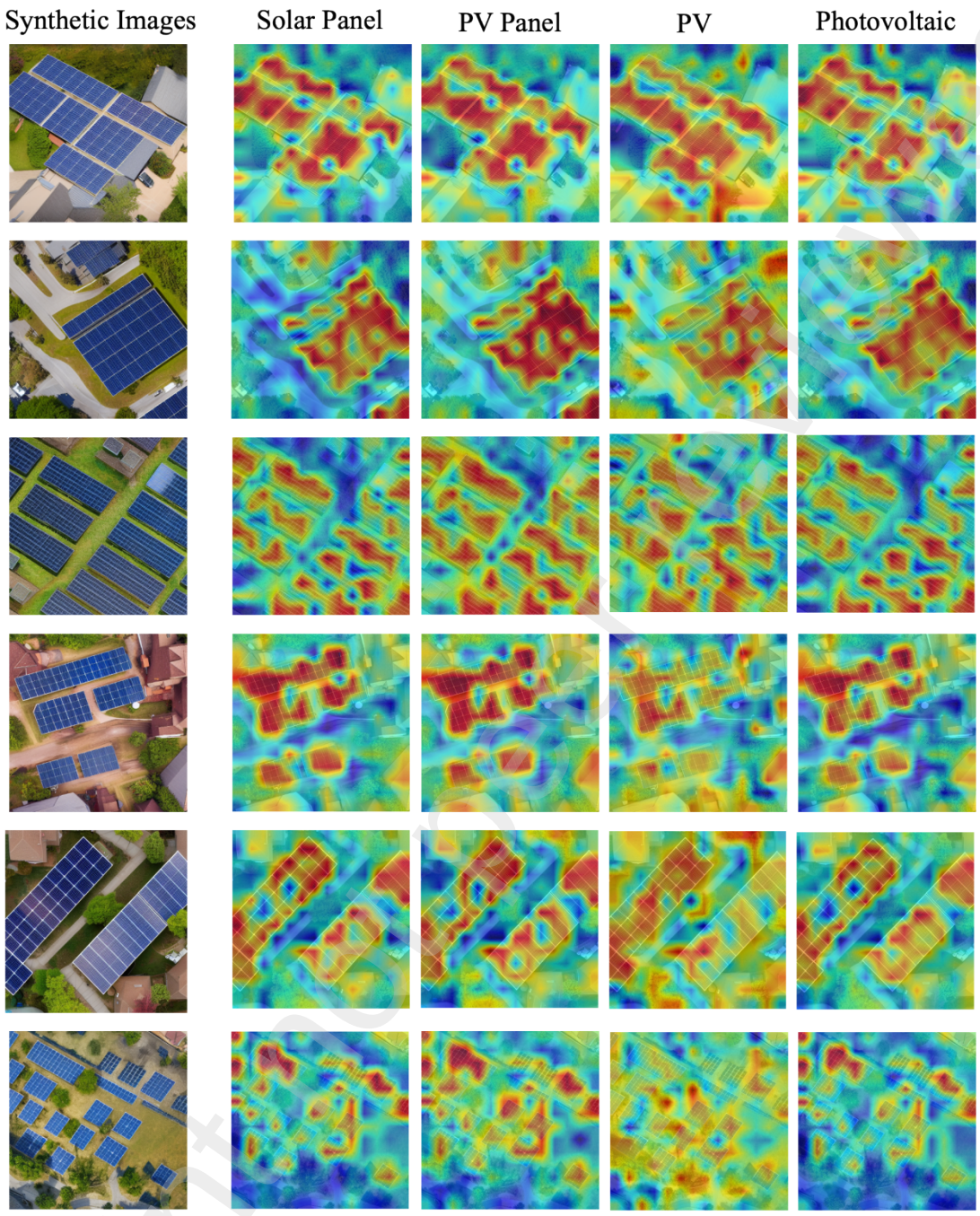


Figure 7. Text perception for related words to solar panels of generated data.

Figure 8 highlights the capability of the proposed method to perceive background information. Common prompts like "road," "green space," "building," and "house" were successfully detected, demonstrating the model's ability to recognize diverse background elements. However, rare prompts such as "park", "water", and "lake" were minimal and challenging to detect, likely due to their infrequent representation in the dataset. These results underscore the importance of robust feature pruning techniques like CLIP Feature Surgery, which eliminate noise while preserving meaningful associations. The similarity maps provide an intuitive and interpretable visualization of

how textual concepts correspond to various regions in the image, making the model's decision-making process transparent and reliable.

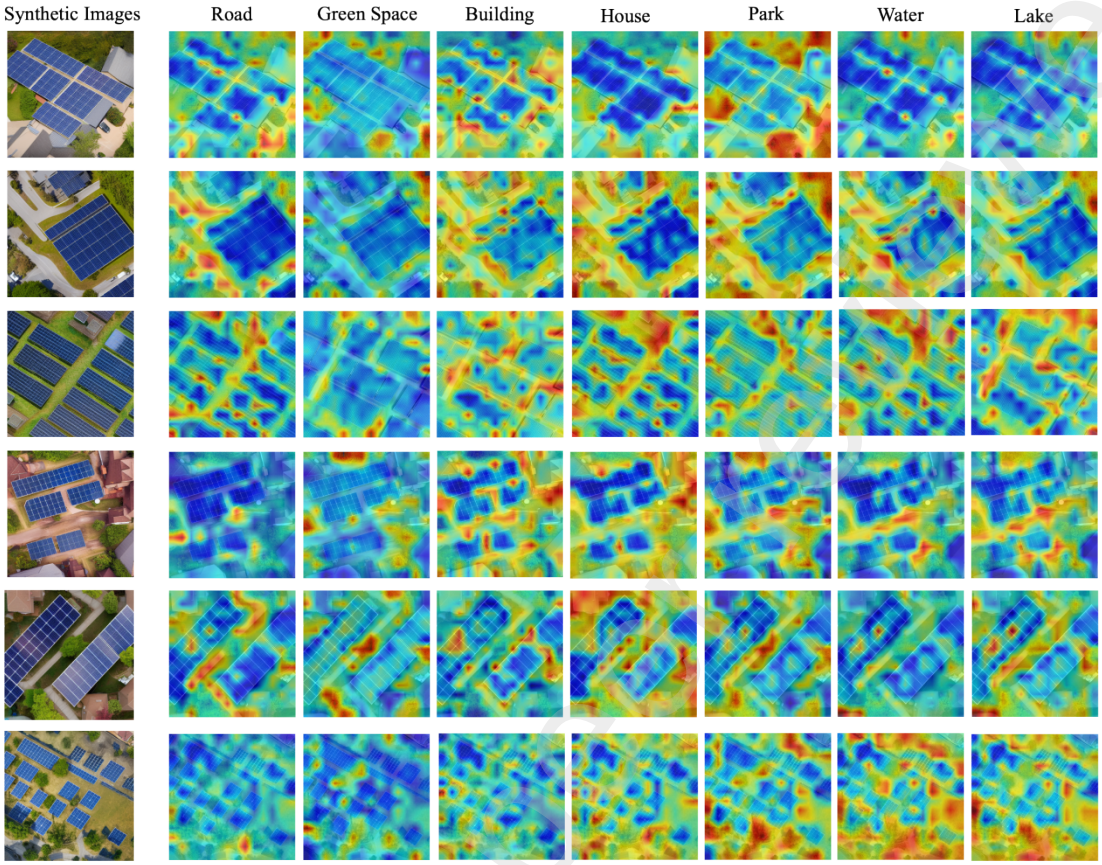


Figure 8. Text perception for words related to backgrounds of generated data.

4.2. Segmentation Results Analysis

4.2.1. Experiment Setup

This subsection analyzes the segmentation results with pure real data, generated data, and integrated data. Three real datasets with 10 real image tiles and their generated 20000 image tiles are tested comprehensively to compare the outputs and verify the effectiveness of generated data. Both the real and generated datasets comprise remote-sensing PV images and binarized labels, which are all processed into 512 × 512-pixel tiles.

The semantic segmentation models are developed using PyTorch, with all computations performed on a server featuring an NVIDIA A100-80GB-PICE. For the segmentation experiments, the open-source platform MMSegmentation (57) is utilized. During training, specific hyperparameters are configured: a batch size of 2, a learning rate of 0.01, a weight decay of 0.005, and a momentum of 0.9. To ensure a balanced comparison, the models are trained for 2000 iterations on real data and 160000

iterations on generated data, maintaining relatively equivalent training steps between the two datasets.

4.2.2. Results Comparison

In the segmentation part, only ten pictures from each dataset are chosen to be trained to simulate the scenario of data scarcity in the real world. The test packages also incorporate typical image backgrounds, such as rooftop PV and grassland PV. The IoU values of PV and mean evaluation metrics for each class are collected for comparison. From **Table 2**, the segmentation results reveal that training with just 10 real images yields suboptimal overall accuracy for PV detection. Among the datasets, the original Heilbronn data achieves the highest mIoU value of 65.23%, outperforming the other datasets, which attain mIoU values of 63.77% and 62.76%, respectively. Jiaxing City data records the highest mRecall value at 79.18%, while the BDAPPV dataset achieves the best mPrecision value of 72.65%. The results show the significant challenges posed by limited image availability in this domain. Utilizing generated PV data may offer a promising solution for improving detection accuracy under data scarcity and constrained resources.

Table 2. Segmentation evaluation comparison with three training modes of three datasets. RD: Pure real data training; GD: Pure generated data training; MD: Mixed data training.

City	Training Modes	Evaluation Values (%)			
		IoU (PV)	mIoU	mPrecision	mRecall
Germany Heilbronn	RD	30.95	65.23	72.49	74.63
	GD	36.32 ↑	68.01 ↑	87.63 ↑	70.56
	MD	40.55 ↑	70.13 ↑	87.89 ↑	73.22
Zhejiang Jiaxing	RD	28.87	63.77	67.81	79.18
	GD	49.00 ↑	74.17 ↑	80.79 ↑	84.92 ↑
	MD	49.42 ↑	74.39 ↑	81.47 ↑	84.50 ↑
BDAPPV	RD	27.49	62.76	72.65	69.70
	GD	36.61 ↑	67.61 ↑	85.22 ↑	71.27 ↑
	MD	36.62 ↑	67.55 ↑	80.54 ↑	73.31 ↑

To evaluate the effectiveness of these generated images in segmentation tasks, the models were trained on the synthetic data and tested on real-world datasets. For instance, testing involved 27,945 image tiles from Heilbronn, 5,484 from Jiaxing, and 13,300 from BDAPPV, using weights derived from the generated data. Each city generated 20,000 PV images to enhance the diversity and volume of the dataset. **Table 2** summarizes the test results for synthetic datasets. Jiaxing demonstrated a significant

improvement in the IoU value for PV panel segmentation, achieving a more than 20% increase compared to training with only 10 real images. Additionally, it gained notable mIoU and mRecall values of 74.17% and 84.92%, respectively. Similarly, Heilbronn and BDAPPV also showed substantial enhancements in evaluation metrics. IoU values increased by 5.37% and 9.12%, respectively, underscoring the effectiveness of the generated datasets in improving real-world segmentation performance.

The mixed data are tested to verify the robustness of generated data further. To learn better the texture features of real images, we added 10 real images to the generated ones. A total of 10 real images plus 20000 generated images images are trained in this section. The table shows that all three datasets reach higher mIoU values, which are 70.13%, 74.39%, and 67.55% respectively, each increasing more than 4.0%. The mPrecision values are all higher than 80.0% and demonstrate that generated datasets-integrated segmentation reaches an effective and robust segment of PV panels, augmenting the data diversity and enhancing the segmentation, with only 10 real data inputs. The visualized reference images in **Figure 9**, comparing the results of three experiments, show a reduction in false negatives (highlighted in red) and false positives (in blue) as the integration of generated data increases. This progress highlights the beneficial impact of synthetic data on PV segmentation tasks across diverse cities and regions.

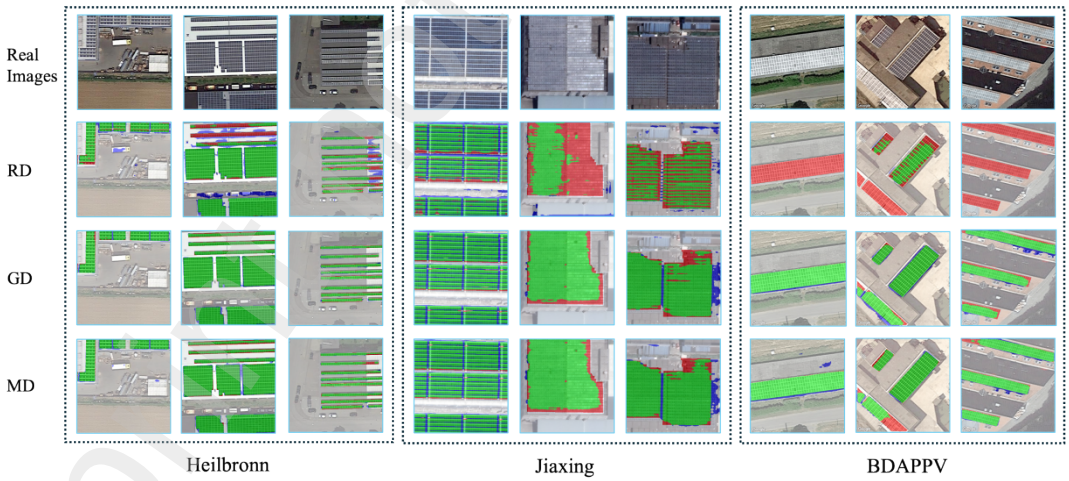


Figure 9. Visualization results comparison of three experiments. RD: Pure real data training; GD: Pure generated data training; MD: Mixed data training.

4.3. Transformative Learning of SynthPV

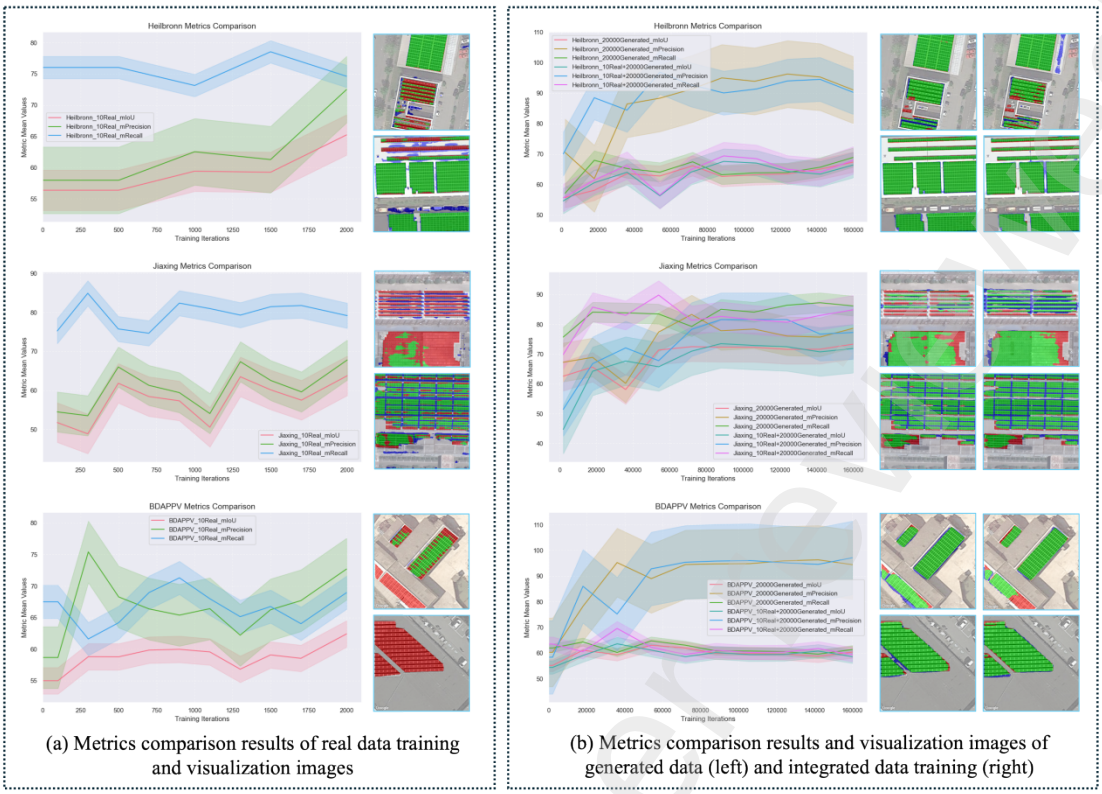
The generalizability of the generating network is crucial for ensuring the effectiveness of transformative learning and detection tasks across different urban

environments with varying characteristics. In this study, Heilbronn City, Jiaxing City, and a subset of the BDAPPV dataset (27) were selected as representative case studies for conducting transferred learning experiments.

The experimental setup mirrors the configurations detailed in previous sections, where the SynthPV is retrained to adapt the model to diverse datasets. This retraining accounts for differences in imaging backgrounds, resolutions, and urban morphologies. As shown in Table 3, the retraining process significantly improves the model's overall evaluation metrics, with the IoU values improvement of 17.35% for Heilbronn, 69.73% for Jiaxing, and 33.18% for the BDAPPV subset than real data test. Notably, Jiaxing exhibits the most substantial improvement, possibly due to its unique characteristics or higher baseline discrepancies than other datasets. These results highlight the importance of multi-context training for enhancing the model's robustness and cross-domain applicability. **Figure 10** (b) depicts the visualization results after retraining, it can be concluded that all three generated datasets could improve the detection accuracy of PV panels in the real world and highlight the efficacy of generated data.

Table 3. Retraining results of generated data from each city. IoU improvement = [IoU (MD) - IoU(RD)] / IoU(RD).

City	Training with 10 real images (%)				Training with generated 20000 images (%)				IoU improvement
	IoU(PV)	mIoU	mPrecision	mRecall	IoU(PV)	mIoU	mPrecision	mRecall	
Germany Heilbronn	30.95	65.23	72.49	74.63	36.32	68.01	87.63	70.56	17.35%
Zhejiang Jiaxing	28.87	63.77	67.81	79.18	49.00	74.17	80.79	84.92	69.73%
BDAPPV	27.49	62.76	72.65	69.70	36.61	67.61	85.22	71.27	33.18%



(a) Metrics comparison results of real data training and visualization images

(b) Metrics comparison results and visualization images of generated data (left) and integrated data training (right)

Figure 10. Evaluation results of retraining and metrics comparison results.

4.4. Generalizability of Generated Data

Heilbronn, known for its diverse urban patterns and environmental variability, was chosen as a representative testbed to assess the generated data's robustness and generalizability. To simulate realistic conditions, 10 real images from each city were mixed with the generated data. This approach aimed to evaluate the efficacy of data augmentation in improving model generalization. As is shown in **Table 4**. the testing results demonstrated significant improvements in IoU values, with 31.02%, 48.04%, and 14.41% respectively for the three modes. These findings underscore the potential of the generated data to enhance model performance, with training with Jiaxing real data displaying the greatest adaptability to mixed datasets, likely due to its superior handling of complex spatial features.

Table 4. Mixing with 20000 images generated from Heilbronn. IoU improvement = [IoU (MD) - IoU(RD)] / IoU(RD).

City	Training with 10 real images (%)				Training with 20000 images generated from Heilbronn (%)				IoU improvement
	IoU(PV)	mIoU	mPrecision	mRecall	IoU(PV)	mIoU	mPrecision	mRecall	
Germany Heilbronn	30.95	65.23	72.49	74.63	40.55	70.13	87.89	73.22	31.02%
Zhejiang Jiaxing	28.87	63.77	67.81	79.18	42.74	70.97	77.13	82.93	48.04%
BDAPPV	27.49	62.76	72.65	69.70	31.45	65.02	88.17	67.20	14.41%

4.5. Limitations and Future Study

The segmentation results of PV panels highlight the effectiveness and robustness of the generated images and masks. However, certain limitations warrant attention in future studies. Enhancements to the SynthPV model could improve perception accuracy across diverse data sources. When training it on different datasets, the decoder applies unique down-sampling patterns to the masks. The concern lies in the synthetic data's background generation, which often lacks realism and fails to reflect real-world conditions accurately. This common challenge in generative AI image creation remains a focus of ongoing research, with efforts directed toward improving the authenticity of generated visuals. Regarding generalizability tests, this study utilized data exclusively from Heilbronn. Future work could involve cross-experiments using a mix of multi-source generated datasets to further validate the approach.

4.6. Potential Applications

The SynthPV method of this work provides a scalable and efficient approach for generating paired PV images and masks, making it suitable for various applications in solar energy system analysis and deployment. For practical implementation, the method can be used to enhance segmentation models in different regions with limited or inconsistent real-world datasets. It is also effective in cross-regional analysis, where synthetic datasets generated for one region can improve detection performance in another, as demonstrated by integrating generated data from Heilbronn into other datasets. SynthPV's integration into remote sensing workflows can streamline the planning and monitoring of PV installations by reducing reliance on manual annotations and improving segmentation accuracy in diverse environments. Additionally, its adaptability to varying geographic and background conditions makes it ideal for large-scale renewable energy assessments, urban planning, and automated PV system mapping, ensuring cost-effective and high-precision outcomes.

5. Conclusion

PV segmentation tasks face challenges due to the limited accessibility and diversity of PV data. Remote sensing tools encounter obstacles such as high labeling costs and reliance on supervised training methods. To address these issues, this study leverages the SynthPV network, integrating Mask Decoder and Text-to-Image based GenAI model, to produce PV images and corresponding masks, aiming to enhance model generalizability, data diversity, and segmentation robustness. The SynthPV network

was utilized to create multi-background PV panel samples and aligned masks, and the Mask Decoder was integrated and trained with less than 1% real-world data. After experiments with real, generated, integrated data training of three cities, results show that evaluation values improved significantly regarding the pure synthetic training, and integrated training process, with IoU values increasing by 9.60%, 9.13%, and 20.55%. The transformative learning and generalizability tests were operated to verify the effectiveness and robustness of the proposed method. The dataset generated from Heilbronn demonstrated successful transferability in detecting PV panels in other cities and could be seamlessly integrated with existing datasets. The results underscore the transformative potential of generated data in segmentation tasks, paving the way for broader applications in industry and aiding policy development.

Data Availability

Data and code will be made available on request.

Acknowledgments

The Hong Kong Polytechnic University supported this work through projects, P0043885 - Flexibility of Urban Energy Systems (FUES), P0047700 - International Research Centre of Urban Energy Nexus, and P0042845 - Data-driven solutions for decarbonizing transportation sector by coupling renewable energy, energy storage, and smart EV-charging. This work is also supported by the High Performance Computing Center at Eastern Institute of Technology and Ningbo Institute of Digital Twin, Ningbo.

Declaration of Competing Interest

The authors declare no competing interests.

Data Availability

Data and code will be made available on request.

References

1. IEA. Renewables 2023. IEA, Paris. 2024.
2. Chen K, Hu K, Li H, Chan S, Chen J, Pei Y, et al. Scalable spectrally selective solar cell for highly efficient photovoltaic thermal conversion. Advances in Applied Energy. 2024;16:100199.
3. Saint-Drenan YM, Good GH, Braun M. A probabilistic approach to the estimation of regional photovoltaic power production. Solar Energy. 2017;147:257-76.
4. Burke M, Driscoll A, Lobell DB, Ermon S. Using satellite imagery to understand and promote sustainable development. Science. 2021;371(6535):eabe8628.
5. Chen Q, Li X, Zhang Z, Zhou C, Guo Z, Liu Z, et al. Remote sensing of photovoltaic scenarios: Techniques, applications and future directions. Applied Energy. 2023;333:120579.

6. Clark CN, Pacifici F. A solar panel dataset of very high resolution satellite imagery to support the Sustainable Development Goals. Sci Data. 2023;10(1).
7. Li P, Zhang H, Guo Z, Lyu S, Chen J, Li W, et al. Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning. Advances in Applied Energy. 2021;4:100057.
8. Uzun B, Atasoy BA, Celik Simsek N. Unmanned Aerial Vehicle (UAV) support for subdivision phase of land readjustment: A case study from Turkey. Land Use Policy. 2022;120:106301.
9. Zech M, Tetens H-P, Ranalli J. Toward global rooftop PV detection with Deep Active Learning. Advances in Applied Energy. 2024;16:100191.
10. De Jong T, Bromuri S, Chang X, Debusschere M, Rosenski N, Schartner C, et al. Monitoring spatial sustainable development: semi-automated analysis of satellite and aerial images for energy transition and sustainability indicators. arXiv preprint arXiv:200905738. 2020.
11. Tan HJ, Guo ZL, Lin ZY, Chen YT, Huang D, Yuan W, et al. General generative AI-based image augmentation method for robust rooftop PV segmentation. Applied Energy. 2024;368.
12. Zhang Y, Schlueter A, Waibel C. SolarGAN: Synthetic annual solar irradiance time series on urban building facades via Deep Generative Networks. Energy and AI. 2023;12:100223.
13. Xun S, Li D, Zhu H, Chen M, Wang J, Li J, et al. Generative adversarial networks in medical image segmentation: A review. Computers in Biology and Medicine. 2022;140:105063.
14. Zhang C, Chen X, Ji S. Semantic image segmentation for sea ice parameters recognition using deep convolutional neural networks. International Journal of Applied Earth Observation and Geoinformation. 2022;112:102885.
15. Jurakuziev D, Jumaboev S, Lee M. A framework to estimate generating capacities of PV systems using satellite imagery segmentation. Engineering Applications of Artificial Intelligence. 2023;123:106186.
16. Jiang H, Yao L, Lu N, Qin J, Liu T, Liu Y, et al. Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery. Earth Syst Sci Data. 2021;13(11):5389-401.
17. Zhu R, Kwan M-P, Perera ATD, Fan H, Yang B, Chen B, et al. GIScience can facilitate the development of solar cities for energy transition. Advances in Applied Energy. 2023;10:100129.
18. Wang M, Cui Q, Sun Y, Wang Q. Photovoltaic panel extraction from very high-resolution aerial imagery using region–line primitive association analysis and template matching. ISPRS Journal of Photogrammetry and Remote Sensing. 2018;141:100-11.
19. Yu J, Wang Z, Majumdar A, Rajagopal R. DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. Joule. 2018;2(12):2605-17.
20. Zhu R, Guo D, Wong MS, Qian Z, Chen M, Yang B, et al. Deep solar PV refiner: A detail-oriented deep learning network for refined segmentation of photovoltaic areas from satellite imagery. International Journal of Applied Earth Observation and Geoinformation. 2023;116:103134.
21. Guo ZL, Lu JY, Chen Q, Liu ZG, Song CC, Tan HJ, et al. TransPV: Refining photovoltaic panel detection accuracy through a vision transformer-based deep learning model. Applied Energy. 2024;355.
22. Wang J, Chen X, Jiang W, Hua L, Liu J, Sui H. PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery. International Journal of Applied Earth Observation and Geoinformation. 2023;119:103309.
23. Qian Z, Chen M, Zhong T, Zhang F, Zhu R, Zhang Z, et al. Deep Roof Refiner: A detail-oriented deep learning network for refined delineation of roof structure lines using satellite imagery. International Journal of Applied Earth Observation and Geoinformation. 2022;107:102680.
24. Tan HJ, Guo ZL, Zhang HR, Chen Q, Lin ZJ, Chen YT, et al. Enhancing PV panel segmentation in remote sensing images with constraint refinement modules. Applied Energy. 2023;350.
25. Paletta Q, Terrén-Serrano G, Nie Y, Li B, Bieker J, Zhang W, et al. Advances in solar forecasting: Computer vision with deep learning. Advances in Applied Energy. 2023;11:100150.
26. Song J, Chen H, Xuan W, Xia J, Yokoya N. Synrs3d: A synthetic dataset for global 3d semantic understanding from monocular remote sensing imagery. arXiv preprint arXiv:240618151. 2024.
27. Kasmi G, Saint-Drenan YM, Trebosc D, Jolivet R, Leloux J, Sarr B, et al. A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. Sci Data. 2023;10(1).
28. Chen YZ, Wang YS, Kirschen D, Zhang BS. Model-Free Renewable Scenario Generation Using Generative Adversarial Networks. Ieee T Power Syst. 2018;33(3):3265-75.
29. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Communications of the ACM. 2020;63(11):139-44.
30. Wang F, Zhang ZY, Liu C, Yu YL, Pang SL, Duic N, et al. Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting. Energy Conversion and Management. 2019;181:443-62.

31. Dong W, Chen XQ, Yang Q. Data-driven scenario generation of renewable energy production based on controllable generative adversarial networks with interpretability. Applied Energy. 2022;308.
32. Wen HR, Du Y, Chen XY, Lim EG, Wen HQ, Yan K. A regional solar forecasting approach using generative adversarial networks with solar irradiance maps. Renewable Energy. 2023;216.
33. Paletta Q, Arbod G, Lasenby J. Benchmarking of deep learning irradiance forecasting models from sky images – An in-depth analysis. Solar Energy. 2021;224:855-67.
34. Bright JM, Babacan O, Kleissl J, Taylor PG, Crook R. A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration. Solar Energy. 2017;147:83-98.
35. Wen H, Du Y, Chen X, Lim EG, Wen H, Yan K. A regional solar forecasting approach using generative adversarial networks with solar irradiance maps. Renewable Energy. 2023;216:119043.
36. Wang F, Zhang Z, Liu C, Yu Y, Pang S, Duić N, et al. Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting. Energy Conversion and Management. 2019;181:443-62.
37. Sappakit T, Limsila T, Chammanard K, Jobsri N, Laomahamek S, Worakulpisut TC, et al. Automated Object Keypoints Dataset Generation Using Blender. Elec Eng Electron Co. 2024.
38. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.
39. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. 2022 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr). 2022:10674-85.
40. Hatanaka Y, Glaser Y, Galgon G, Torri G, Sadowski P. Diffusion models for high-resolution solar forecasts. arXiv preprint arXiv:230200170. 2023.
41. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems. 2022;35:36479-94.
42. Zhu Y, Zhou Y, Xia Y, Wei W. Resilience-Oriented Extreme Weather Conditional Renewable Scenario Generation Based on Diffusion Models and Few-shot Learning. 2023.
43. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems. 2020;33:6840-51.
44. Yuan Z, Hao C, Zhou R, Chen J, Yu M, Zhang W, et al. Efficient and Controllable Remote Sensing Fake Sample Generation Based on Diffusion Model. IEEE Transactions on Geoscience and Remote Sensing. 2023;61:1-12.
45. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.
46. Zhu J-Y, Park T, Isola P, Efros AA, editors. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision; 2017.
47. Nie Y, Zelikman E, Scott A, Paletta Q, Brandt A. SkyGPT: Probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained VideoGPT. Advances in Applied Energy. 2024;14:100172.
48. Browne A. Using Stable Diffusion to Improve Image Segmentation Models. 2023.
49. Lin Z, Guo Z, Huang D, Song C, Tan H, Song X, et al. Leveraging generative AI for renewable energy: photovoltaic panel semantic segmentation case study. Energy proceedings. 2023;36.
50. Meiser M, Duppe B, Zinnikus I. Generation of meaningful synthetic sensor data - Evaluated with a reliable transferability methodology. Energy and Ai. 2024;15.
51. Wu WJ, Zhao YZ, Shou MZ, Zhou H, Shen CH. DiffuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segmentation Using Diffusion Models. Ieee I Conf Comp Vis. 2023:1206-17.
52. Wu WJ, Zhao YZ, Chen H, Gu YC, Zhao R, He YF, et al. DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models. Adv Neur In. 2023.
53. Nguyen Q, Vu T, Tran A, Nguyen K. Dataset Diffusion: Diffusion-based Synthetic Dataset Generation for Pixel-Level Semantic Segmentation. Adv Neur In. 2023.
54. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lect Notes Comput Sc. 2015;9351:234-41.
55. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez J, Luo P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers2021.
56. Li Y, Wang H, Duan Y, Li X. Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv preprint arXiv:230405653. 2023.
57. open-mmlab. mmsegmentation. 2023.