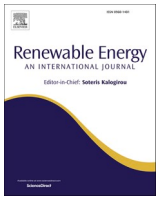
Renewable Energy 260 (2026) 125217



Contents lists available at ScienceDirect

Renewable Energy

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



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 [b,e], Junwei Liu [a,c], Haoran Zhang [f], Jinyue Yan [a,c,**]

[a] State Key Laboratory of Climate Resilience for Coastal Cities, Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[b] Zhejiang Key Laboratory of Industrial Intelligence and 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

ARTICLE INFO

Keywords:
Generative AI
Mask generation
Text-to-Image
SynthPV
PV segmentation
Generalizability

ABSTRACT

The robust detection of photovoltaic (PV) panels across multiple regions and backgrounds has traditionally relied on aligned remote-sensing imagery with manually labelled annotations, particularly in supervised learning. However, it faces challenges in collecting training datasets, including large volumes, accessibility concerns, and inconsistent quality. To efficiently produce paired PV images and masks, this study introduces 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 simultaneously generate highly aligned PV masks. The experimental results from three distinct scenarios using the Heilbronn, Jiaxing, and BDAPPV datasets demonstrate that synthetic image-mask pairs could significantly enhance PV segmentation performance. Notably, Intersection over Union (IoU) values increase by an average of 13.09 % 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, thereby enabling more efficient and scalable applications in solar energy system analysis and deployment.

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. Global PV installation capacity is projected to increase its share of the global renewable energy portfolio to over 40 % in 2028, and this growth could account for almost 80 % of renewable electricity expansion [1,2].

As PV deployment expands across urban and peri-urban environments, accurate detection and segmentation of PV modules have become increasingly crucial for solar potential assessment, energy planning, and infrastructure management. PV installations are now widely distributed across building rooftops, facades, water bodies, grasslands, mountains,

* Corresponding author. State Key Laboratory of Climate Resilience for Coastal Cities, Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China.
** Corresponding author. State Key Laboratory of Climate Resilience for Coastal Cities, Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China.
E-mail addresses: hongjun.tan@connect.polyu.hk (H. Tan), zhiling.guo@polyu.edu.hk (Z. Guo), lijiaze@emails.bjut.edu.cn (J. Li), ychen@eitech.edu.cn (Y. Chen), chenqi@cug.edu.cn (Q. Chen), junweei.liu@polyu.edu.hk (J. Liu), h.zhang@pku.edu.cn (H. Zhang), j-jerry.yan@polyu.edu.hk (J. Yan).

https://doi.org/10.1016/j.renene.2026.125217
Received 4 February 2025; Received in revised form 4 January 2026; Accepted 4 January 2026
Available online 6 January 2026
0960-1481/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

and roadsides, posing significant challenges for large-scale and heterogeneous urban mapping. Traditional methods for PV panel detection and energy capacity assessment often suffer from limited precision, data accessibility issues, and acquisition challenges [3,4].

Recent advances in deep learning (DL) have significantly improved the performance of PV detection using remote sensing imagery [5,6]. The process of collecting such data is not only time-consuming and costly but also hindered by issues of data diversity [7]. Additionally, technological constraints and a lack of diverse PV samples limit the effectiveness of DL methods in detecting urban PV modules [8]. The sensitivity of these models to heterogeneous environments also makes it challenging to transfer detection algorithms across different urban areas [9]. Consequently, despite continuous improvements in network architectures, the scalability and generalizability of city-scale PV segmentation remain constrained more by data availability and annotation bottlenecks than by model capacity itself.

To address these limitations, General Generative AI (GenAI) has emerged as a promising approach for improving PV data acquisition and processing. By generating synthetic images together with aligned annotations, GenAI offers the potential to significantly reduce manual labelling costs while enhancing data diversity and accessibility [10,11]. Such capabilities provide new opportunities to improve the robustness and scalability of urban PV segmentation, particularly in data-scarce environments. Furthermore, GenAI supports local data processing, ensuring data privacy and accessibility, which is particularly valuable for urban planners and developers.

1.2. Related works

Remote sensing techniques have become indispensable tools for detecting and segmenting photovoltaic (PV) images, supporting large-scale assessments of solar energy potential. Numerous studies have demonstrated the effectiveness of deep learning models in identifying PV installations across diverse urban surfaces and satellite platforms [5,12,13]. Parallel investigations [14] have revealed the effectiveness of combining deep learning architectures with multi-resolution remote sensing data for extracting information from various PV system configurations. Notable examples include DeepSolar [15], TransPV [16], and PVNet [17]. With the development of diverse network architectures equipped with various encoders [18], the integration of attention mechanisms to minimize background interference, and the implementation of sophisticated loss functions has further improved the detection of multi-level characteristics on urban surfaces, aiding in city-scale PV module identification [19].

Despite these methodological improvements, existing PV segmentation studies are predominantly model-centric, assuming the availability of extensive, high-quality labelled datasets. In practice, many state-of-the-art models are trained using tens of thousands of manually annotated image tiles or roof instances derived from limited geographic regions, which restricts their generalizability to heterogeneous urban environments [20]. To alleviate data scarcity, several studies have explored synthetic data generation and data augmentation techniques. Kasmi et al. [21] introduced a composite dataset combining aerial images, segmentation masks, and installation metadata. Chen et al. [22] utilized generative adversarial networks (GANs) [23] to create renewable energy scenarios based on weather conditions. Wang et al. [24] demonstrated the ability of GANs to generate high-quality samples for weather classification. Wen et al. [25] employed solar irradiance maps to train multi-scale GANs for image-based forecasting. Furthermore, Wang et al. [26] introduced a GAN and CNN-based weather classification model that enriched training datasets for diverse weather conditions. Other studies have combined numerical data with images to enrich training datasets or improve forecasting performance [27,28]. However, these approaches typically focus on improving data volume or realism and do not generate pixel-aligned image mask pairs, limiting their direct applicability to segmentation tasks.

More recently, diffusion-based and text-to-image generative models have shown strong potential in synthetic data generation for remote sensing applications [29]. Song et al. [20] utilized tools like Blender [30], Python, GPT-4 [31], and Stable Diffusion [32] to generate a synthetic remote sensing 3D dataset. The score-based diffusion model [33] was employed to forecast day-ahead solar irradiance. Meanwhile, Zhu et al. [34] developed an innovative scenario generation approach by merging conditional diffusion models [35] with few-shot learning, overcoming limitations in training datasets for weather classification tasks. Yuan et al. [36] further demonstrated the effectiveness of diffusion models in generating remote-sensing fake samples. These models offer improved control over image diversity and realism while reducing reliance on large labelled datasets. Nevertheless, to the best of our knowledge, no prior work has systematically explored text-to-image GenAI for the simultaneous generation of urban PV remote sensing images and their aligned segmentation masks, nor evaluated their effectiveness for downstream PV segmentation under low-data conditions.

1.3. Objective and contribution

Given the high cost, limited scalability, and geographic constraints of manually annotated PV datasets, there is a pressing need for data-centric approaches that can enhance PV segmentation performance while reducing dependence on large labelled datasets [37]. This study addresses this challenge by introducing SynthPV, a Text-to-Image based General Generative AI framework that aims at generating diverse urban PV images with aligned annotations to improve the adaptability, robustness, and scalability of PV detection algorithms. By integrating a small amount of real-world data with generated samples, the proposed framework enables practical model training even under conditions of minimal labelled data. The main contributions of this paper can be concluded as follows.

(1) The SynthPV is introduced to simultaneously generate urban PV images and aligned masks to address the annotation bottleneck in city-scale PV mapping directly.
(2) The work leverages a Text-to-Image-based GenAI framework to generate remote-sensing image pairs, significantly enhancing the diversity in PV appearance, background complexity, and urban context.
(3) By training with a combination of generated and integrated data from three datasets, the approach achieves higher 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

Fig. 1 illustrates the research framework of this study, addressing challenges like data scarcity, diversity, labelling costs, and accessibility concerns in the PV segmentation process. 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 generating 20,000 PV image pairs per city, ensuring that synthetic data closely aligns 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 background, reviews related work, and highlights the study's contributions. Section 2 introduces the study areas and data-processing procedures and

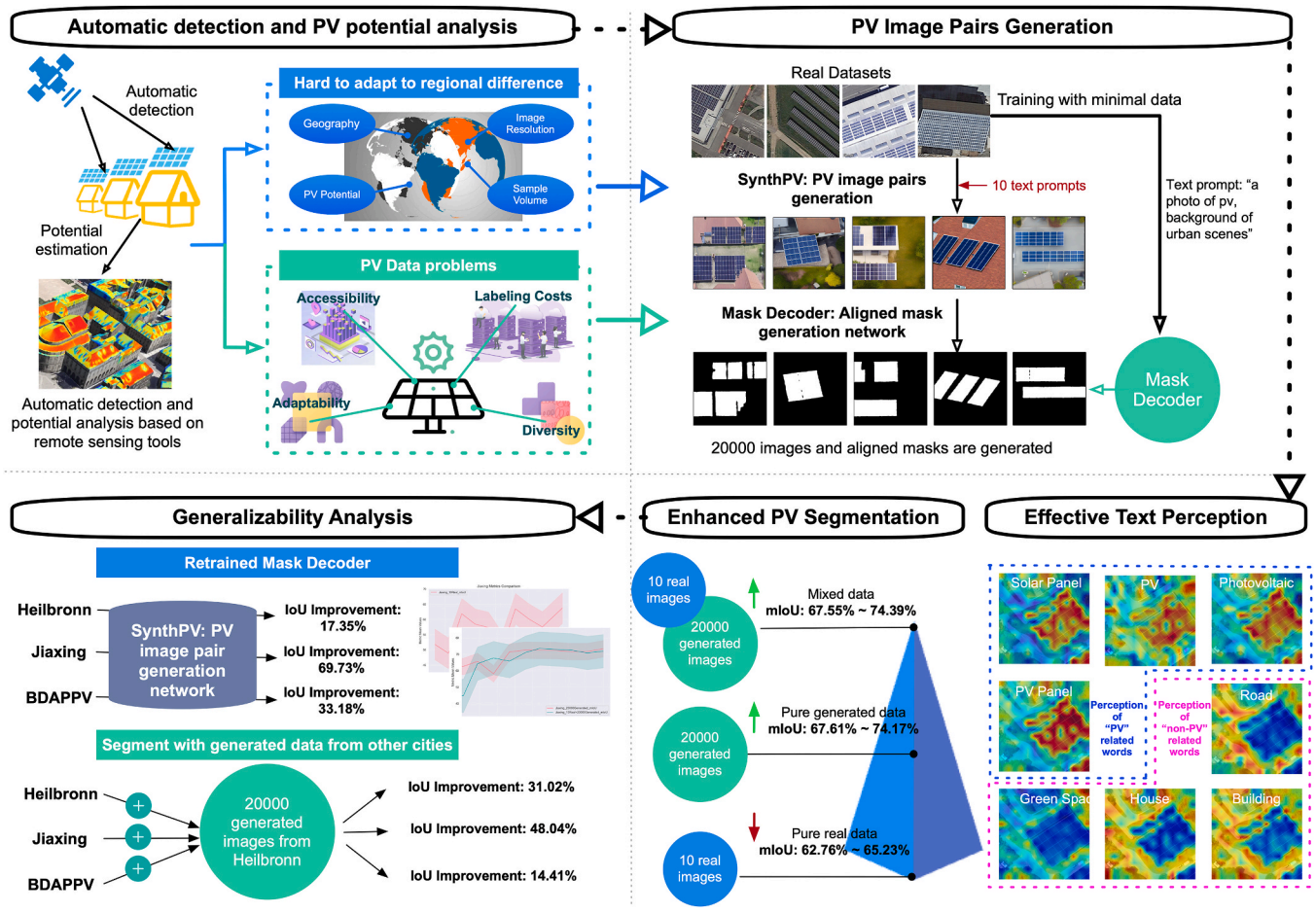


Fig. 1. The research framework for enhancing PV segmentation using SynthPV and Mask Decoder networks across multi-regional datasets.

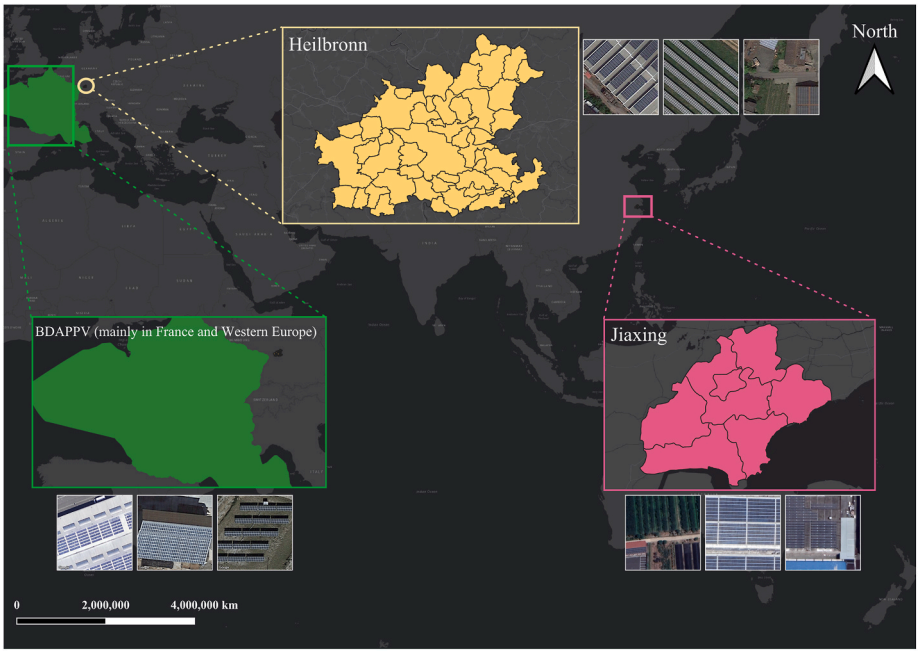


Fig. 2. The satellite imagery of the three cities.

provides 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 analyses the results of the proposed approach, discusses its text-perception capabilities for distinguishing PV from non-PV-related features, evaluates its generalizability across different datasets, and highlights its potential for broader applications. Section 5 concludes this work with solid results.

2. Data and materials

2.1. Study area

This study adopts three datasets from diverse regions to investigate the transformability and generalizability of synthetic datasets (Fig. 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 and offers moderate spatial resolution (1.34 m/px) and 512 × 512 tiles. The third dataset, BDAPPV [21], primarily covers France and Western Europe and showcases PV systems in coastal plains and areas with variable weather patterns. It has a lower resolution (2.0 m/px) and uses 400 × 400 tiles. Table 1 shows the detailed differences between the three cities. During training, 10 real images from each dataset are selected for comparison with the generated image pairs and the training images. SynthPV is pretrained on 50 real images, and 20,000 image pairs are generated using a Text-to-Image diffusion model that incorporates a PV-specified Mask Decoder, 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.

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. Fig. 3 illustrates representative images from the three regions, depicting varied backgrounds, including 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.

The PV datasets consist of binary semantic labels identifying PV installations and non-PV areas [38]. 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 images of Heilbronn and Jiaxing. The annotation process applied a consistent cropping approach to ensure uniformity across datasets. The BDAPPV dataset contains open-source images and annotations. In total, this experiment datasets include 27,945 tiles for Heilbronn City, 5484 for Jiaxing City, and 13,300 for BDAPPV.

3. Method

In this section, the detailed design of the method is explained. As depicted in Fig. 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 in 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 Fig. 4, the SynthPV network demonstrates the generation of images and masks. Within this framework, the VAE employs an encoder $\mathscr{E}$ to compress image inputs into a latent space, while a decoder $\mathscr{D}$ 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.

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 $\widetilde{z}$ 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 $\mathscr{D}$, denoted as $\widetilde{x} = \mathscr{D}(\widetilde{z})$. To generate a diverse dataset, ten text prompts were used to produce 20,000 PV image pairs. As illustrated in Fig. 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 pipeline's effectiveness in preserving essential details while enhancing annotation precision.

Mask generation, particularly in the context of solar energy applications, has received limited attention from researchers relative 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. [39] 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. [40] expanded the application of pre-trained diffusion

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	Southern Germany	Growing PV installations	27,945 (10)	50	20,000	0.15	512 * 512
Jiaxing, China	Eastern China	Pioneer in PV installation	5484 (10)	50	20,000	1.34	512 * 512
BDAPPV	Mainly in France and Western Europe	Southern regions with high solar irradiation	13,300 (10)	50	20,000	2.0	400 * 400

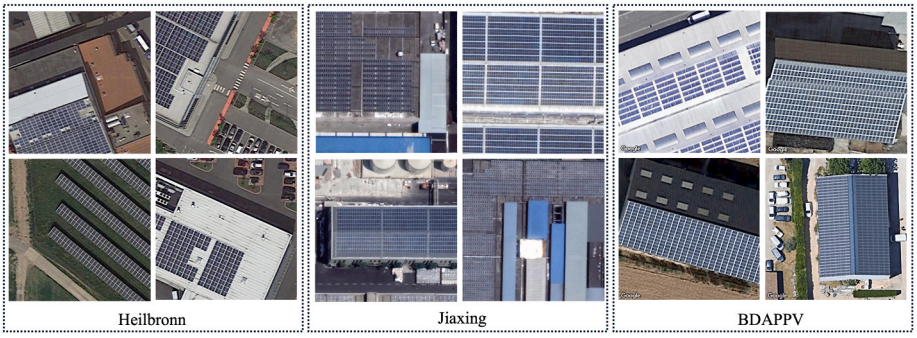


Fig. 3. Samples of the selected pre-generating training dataset.

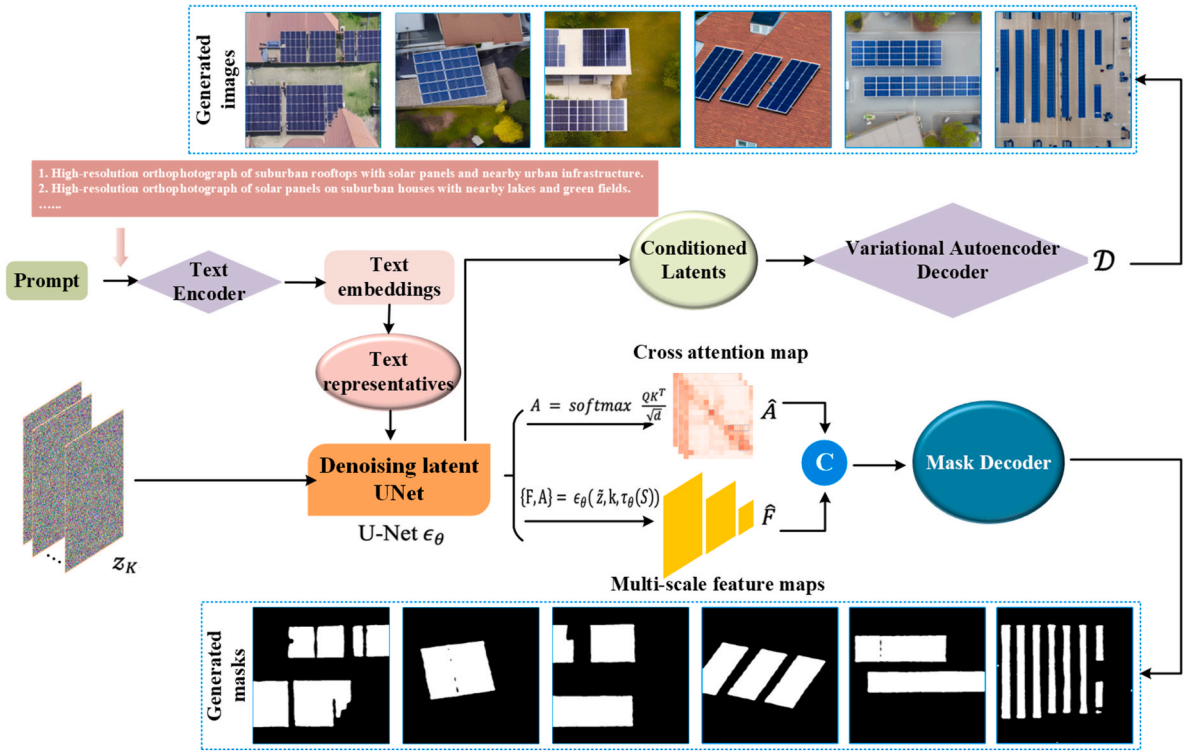


Fig. 4. Detailed SynthPV network, a Text-to-Image based Generative AI model, for generating paired PV images and masks, and illustrated with five reference examples.

models to text-to-image synthesis for data generation. Further advancements were made with the introduction of the Dataset Diffusion network [41], 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 annotation generation, distinguishing it from the earlier DiffuMask [39].

Significant progress has been made in mask generation through various studies. However, these existing methods have specific limitations when applied to the generation of solar panel datasets. 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 produce 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, thereby 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

Fig. 5 presents the Mask Decoder network, which employs a diffusion model to manipulate the 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:

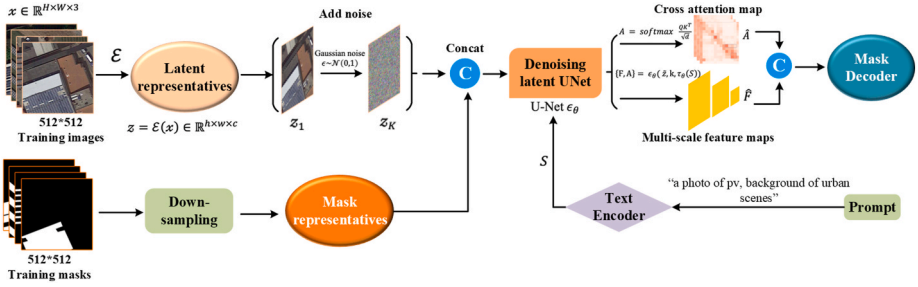


Fig. 5. Detailed Mask Decoder network with less than 1 % real-world PV data.

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

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, $\mathscr{k}$ is uniformly sampled from the set $\{1,\ldots,K\}$.:

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

The S means 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 $A = softmax\ \frac{QK^T}{\sqrt{d}}$, 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 $\widehat{A}$. To obtain the final feature representation, the multi-scale feature maps F are concatenated with the average cross-attention maps $\widehat{A}$. This combined representation is further processed using a 1×1 convolution to fuse the features, resulting in the final output $\widehat{F} = Conv([F,\widehat{A}])$.

The Mask Decoder translates the feature representative $\widehat{F}$ into perception annotations suitable for various downstream tasks. To segment PV from the background, the Mask Decoder employs a dual-branch architecture comprising a pixel decoder and a 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 $\widehat{F}$ and a set of K learnable queues $\{\mathrm{Q}_0,\mathrm{Q}_1,\ldots,\mathrm{Q}_\mathrm{K}\}$ as input, the Mask Decoder generates K binary masks $M = m_1,m_2,\ldots,m_K \in \{0,1\}^{K\times h\times w}$. 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 [42] is frequently selected for its balance among 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 such as SegFormer [43] achieve high accuracy but entail significantly larger parameter counts. For example, SegFormer B5 contains approximately 84 million parameters, and larger versions further increase the computational burden. U-Net's encoder-decoder architecture, with skip connections, is notable 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, the 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.

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

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.

In our implementation, the segmentation model is trained using Cross-Entropy Loss ($\mathscr{L}_{CE}$), which is a widely adopted objective function for classification tasks, including pixel-wise semantic segmentation. For binary segmentation (PV vs. background), the loss is computed as:

$$\mathscr{L}_{CE} = -\frac{1}{N}\sum_i q_i \log\,(p_i) + (1-q_i)\log\,(1-p_i) \quad (5)$$

where N is the number of pixels, $q_i\in\{0,1\}$ is the ground truth label, and $p_i\in[0,1]$ is the predicted probability for the PV class.

To optimize the model parameters, the network uses Stochastic Gradient Descent (SGD) with momentum. The parameter update rule is defined as:

$$\nu_{t+1} = \mu\cdot\nu_t + \eta\cdot\nabla_\theta\mathscr{L}(\theta_t) \quad (6)$$

$$\theta_{t+1} = \theta_t - \nu_{t+1} \quad (7)$$

where θ_t denotes the model parameters at iteration t, η is the learning rate, μ is the momentum factor (set to 0.9), and $\mathscr{L}$ represents the loss function. The weight decay of 0.005 is also applied to prevent overfitting.

Our experiment evaluates the effectiveness of generated data in improving model performance under resource-constrained conditions. Specifically, we demonstrate that generated data significantly improves PV segmentation accuracy, delivering notable relative gains without the high computational costs associated with larger architectures. To this end, this work utilizes the U-Net network as the 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 is utilized, including Intersection-over-Union (IoU), Precision, and Recall. These metrics provide quantitative measures of accuracy, with higher values indicating better performance. IoU, in particular, quantifies 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 approaching 1.0 indicate that the model achieves near-ideal performance. For a more detailed understanding, the formulas used for these evaluation metrics are presented below:

$$IoU = \frac{area(PredictionResult \bigcap GroundTruth)}{area(PredictionResult \bigcup GroundTruth)} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

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 a Mask Decoder and a Text-to-Image generation model to create diverse datasets and targeted mask generation. Ten text prompts were used to generate 20,000 PV image samples, demonstrating the model's ability to synthesize realistic images and aligned masks. The models were retrained to learn local features across different cities and to reduce image inconsistencies. Fig. 6 illustrates the generated images and the corresponding aligned masks. To better compare with the ground truth, manual labelling was conducted to evaluate the quality of the synthetic masks generated by SynthPV. The results indicate that the synthetic masks matched the manually labelled masks with over 80 % overlap, calculated by dividing the number of pixels in the synthetic masks (b) by those in the manual masks (c). Fig. 6 (d) displays the overlaid images, with mismatched pixels highlighted in red. The remaining uncovered areas indicate that the generated masks generally cover the PV regions. Therefore, SynthPV is effective in achieving acceptable labelling accuracy for segmentation tasks.

4.1.2. Text perception analysis based on CLIP surgery

To evaluate the quality of the generated images, the CLIP Surgery method [44] 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 image regions. Additionally, a dual-path 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,

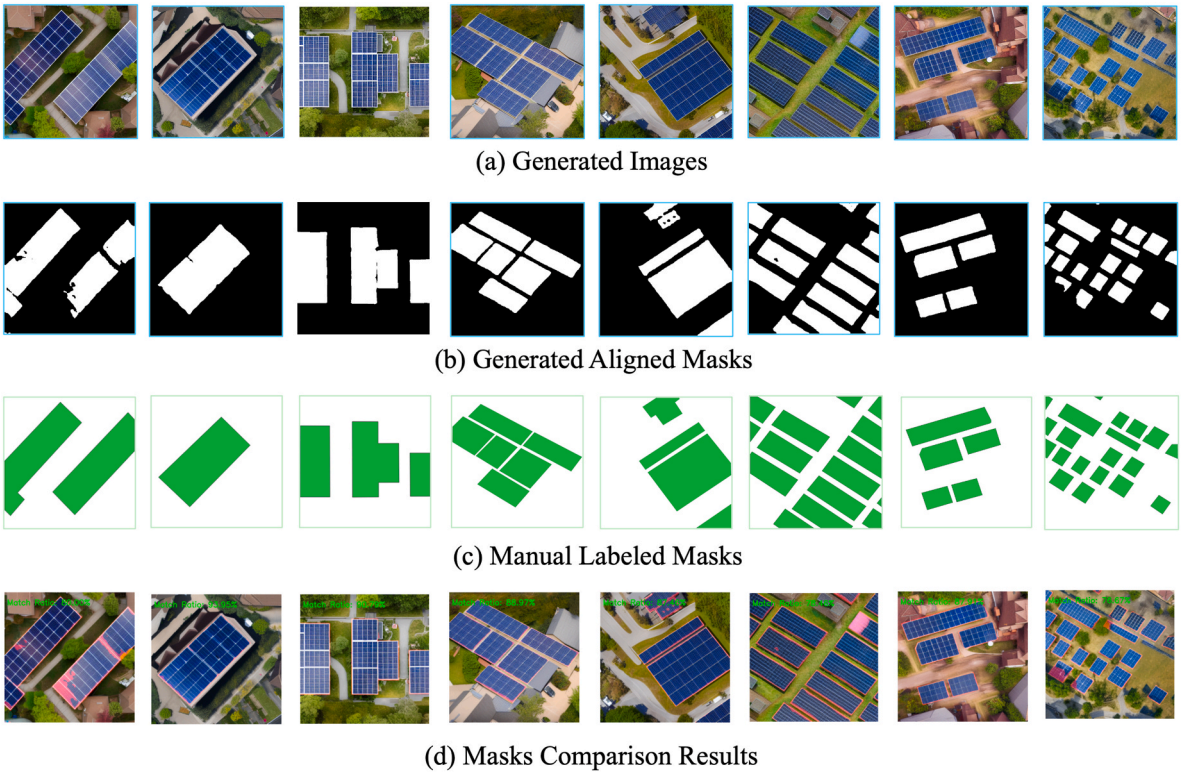


Fig. 6. The (a) generated reference images, (b) aligned masks, (c) manually labelled masks, and (d) mask comparison results with match ratio.

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 Fig. 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

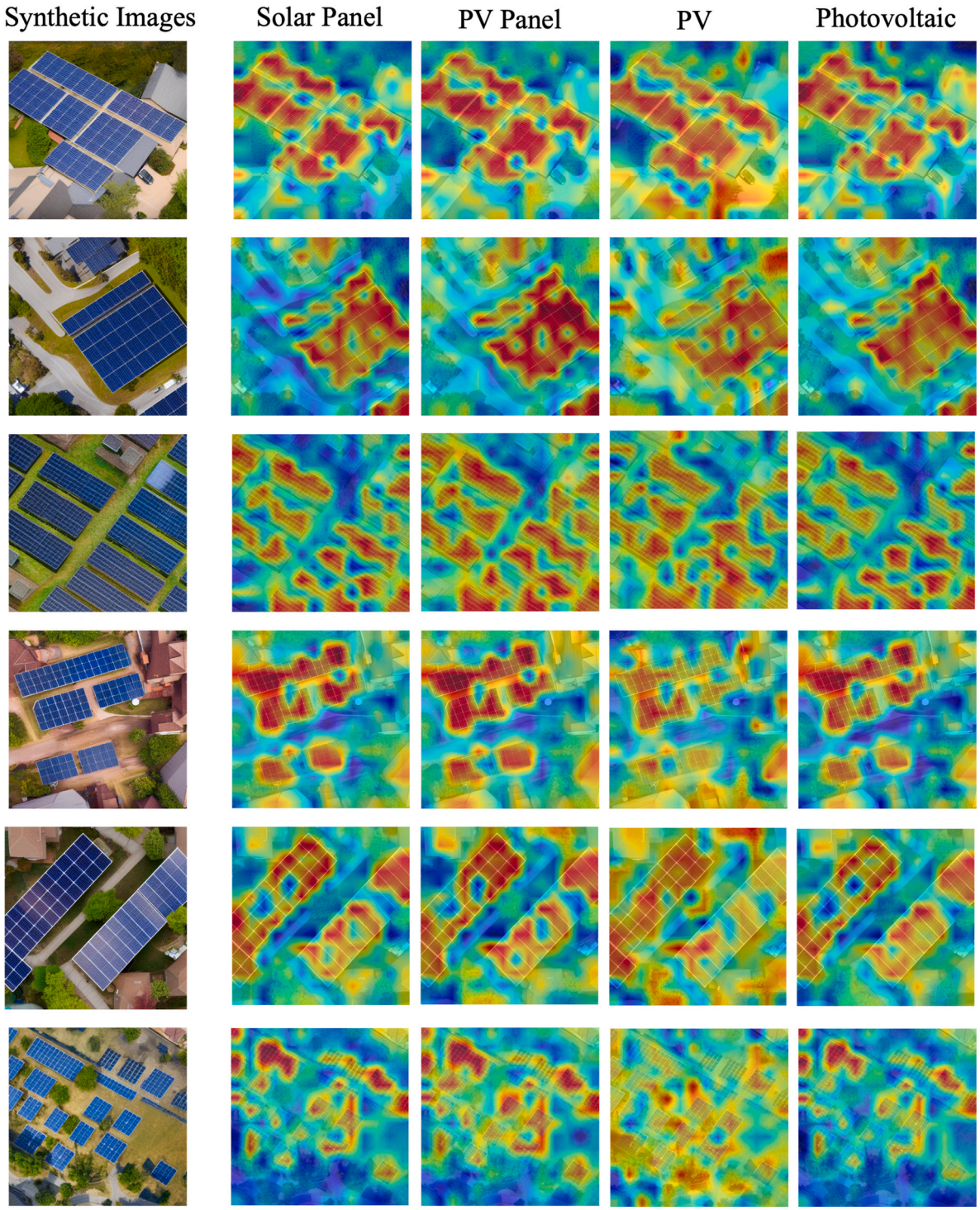


Fig. 7. Text perception for related words to solar panels of generated data.

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.

Fig. 8 illustrates the proposed method's ability 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 scarce and difficult to detect, likely due to their low frequency in the dataset. These results underscore the importance of robust feature-pruning techniques, such as 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.

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, each comprising 10 real image tiles and their corresponding 20,000 generated image tiles, were comprehensively tested to compare the outputs and verify the effectiveness of the 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.

To ensure that the 10 real image tiles effectively represent the dataset's diversity, the t-distributed Stochastic Neighbor Embedding (T-SNE) [45] was used for strategic sample selection. Specifically, RGB image data were used as input features and projected into a two-dimensional space via t-SNE to preserve local and global similarity structures among the full set of real images. This visualization facilitated manual clustering of the feature space, from which we identified representative regions corresponding to varying PV structures and background complexities.

From these representative regions, 10 samples were strategically selected, favoring those with larger PV panel areas and diverse background features, to ensure broad coverage of key visual patterns. A degree of controlled randomness was introduced in the final selection to simulate real-world constraints and avoid selection bias. As a result, the baseline real dataset, although limited in size, captures a wide range of image variation across cities. This selection approach was uniformly applied to the Heilbronn, Jiaxing, and BDAPPV datasets. Through this careful design, the experiments aim to assess the value of the generated data in improving segmentation accuracy under low-resource supervision.

The semantic segmentation models are developed in PyTorch, with all computations performed on a server equipped with an NVIDIA A100-80 GB-PICE. For the segmentation experiments, the open-source platform MMSegmentation [46] was utilized. During training, the following hyperparameters were configured: batch size 2, learning rate 0.01, weight decay 0.005, and momentum 0.9. To ensure appropriate learning while considering dataset scale and training efficiency, the models were trained for 2000 iterations on the real datasets and 160000 iterations on the generated dataset. This configuration allows sufficient learning

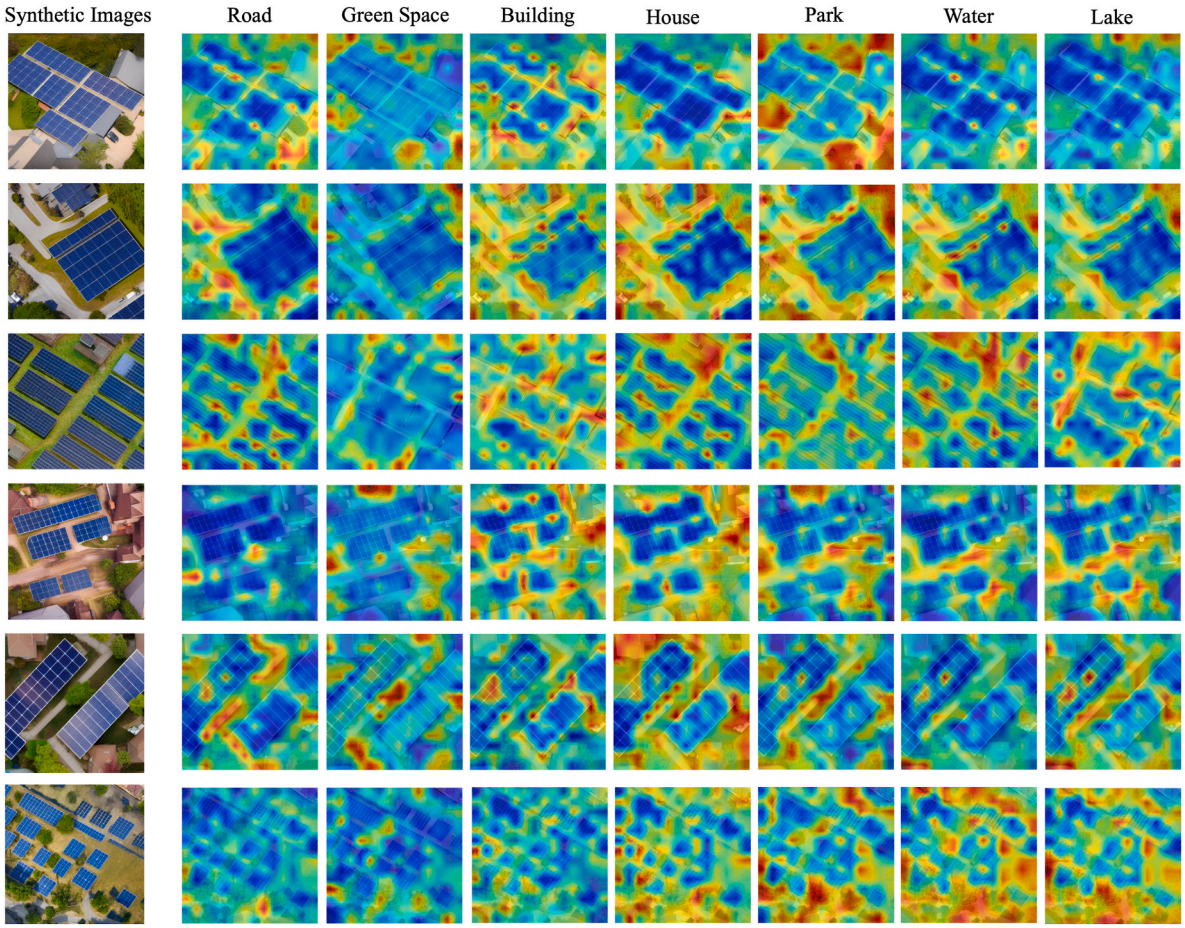


Fig. 8. Text perception for words related to the backgrounds of the generated data.

epochs from the small real dataset through repeated exposure, while avoiding excessive computation on the much larger generated dataset.

4.2.2. Results comparison

In the segmentation phase, only ten images from each dataset were selected for training to simulate the scenario of data scarcity encountered in the real world. The test sets also included typical image backgrounds, such as rooftop PV and grassland PV. The IoU values for PV and the mean evaluation metrics for each category were gathered for comparison. According to Table 2, the segmentation results indicated that training with just 10 real images led to suboptimal overall accuracy for PV detection. Among the datasets, the original Heilbronn data yielded the highest mIoU value of 65.23 %, surpassing the other datasets, which achieved mIoU values of 63.77 % and 62.76 %, respectively. The Jiaxing City dataset recorded the highest mRecall value at 79.18 %, while the BDAPPV dataset achieved the best mPrecision value of 72.65 %. These findings highlight the significant challenges posed by limited image availability in this field. Employing generated PV data is suggested as a promising solution for enhancing detection accuracy in scenarios of data scarcity and constrained resources.

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, 5484 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 were tested to further verify the robustness of the generated data. To better capture the texture features of real images, 10 real images were added to the generated ones. A total of 10 real images plus 20,000 generated images were used for training in this section. The results, as shown in the table, indicate that all three datasets achieved higher mIoU values of 70.13 %, 74.39 %, and 67.55 %, respectively, each increasing by more than 4.0 %. The mean precision values exceeded 80.0 %, demonstrating that segmentation based on the generated datasets effectively improved the detection of PV panels by augmenting data diversity, even with only 10 real data inputs. The visualized reference images in Fig. 9, which compare the results of three experiments, revealed a reduction in false negatives (highlighted in red) and false positives (in blue) as the proportion of generated data increased. This progress underscores the beneficial impact of synthetic data on PV segmentation tasks across diverse cities and regions (see Fig. 10).

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	RD	30.95	65.23	72.49	74.63
Heilbronn	GD	36.32 ↑	68.01 ↑	87.63 ↑	70.56
	MD	40.55 ↑	70.13 ↑	87.89 ↑	73.22
Zhejiang	RD	28.87	63.77	67.81	79.18
Jiaxing	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 ↑

4.3. Transformative learning of SynthPV

The generalizability of the generating network is crucial to ensuring the effectiveness of transformative learning and detection tasks across diverse urban environments. In this study, Heilbronn City, Jiaxing City, and a subset of the BDAPPV dataset [21], which primarily covers France and Western Europe, were selected as representative case studies for conducting transfer learning experiments.

The experimental setup mirrors the configurations detailed in the previous sections, in which the SynthPV was retrained to adapt to diverse datasets. This retraining accounts for differences in imaging backgrounds, resolutions, and urban morphologies. As shown in Table 3, the retraining process significantly improved the model's overall evaluation metrics, with IoU improvements of 17.35 % for Heilbronn, 69.73 % for Jiaxing, and 33.18 % for the BDAPPV subset relative to the real-data test. Notably, Jiaxing exhibits the most significant improvement, possibly due to its unique characteristics or greater baseline discrepancies relative to other datasets. These results highlight the importance of multi-context training for enhancing the model's robustness and cross-domain applicability. Fig. 10 (b) depicts the visualization results after retraining. These findings suggest that all three datasets improved the detection accuracy of PV panels in real-world scenarios, further demonstrating the efficacy of the generated data.

4.4. Generalizability of generated data

To further assess the generalizability and robustness of the generated data beyond a single source domain, an experiment that differed in objective and setup from that in Section 4.2 was designed. This section first investigates whether data generated in one city (Heilbronn) can effectively augment real-world datasets from other cities (e.g., Jiaxing and BDAPPV), particularly under data-scarce conditions. Specifically, the 20,000 image pairs were generated solely from Heilbronn's data and were tested for generalizability by integrating them with 10 real image tiles from the Jiaxing, BDAPPV, and Heilbronn datasets. Furthermore, a new real-world dataset, PV01 [47], comprising 10 authentic images, was used to train with three synthetic datasets. PV01 was obtained from Unmanned Aerial Vehicle (UAV) orthophotos with 0.1m spatial resolution; a total of 1152 images were tested. All experiments were under the same training environment and setup. These experiments simulated realistic scenarios in which limited labelled data are available from target domains and evaluated whether synthetic data from a different domain can improve model performance.

As shown in Table 4, the test results demonstrate significant improvements in IoU, with increases of 31.02 %, 48.04 %, and 14.41 % across the three modes. Among these, the Jiaxing dataset shows the greatest adaptability to the mixed dataset, which may be attributed to its capability to capture complex spatial textures and edge characteristics. Table 5 presents the test results on the new data after training, along with those on the synthetic data. All synthetic datasets performed well across different scenarios, with the highest IoU for PV detections reaching 73.95 %, even surpassing the prior tests in Table 4. These findings underscore that the proposed PV image generation approach can contribute not only to model performance within the same domain but also to cross-domain augmentation, thereby extending the utility of synthetic data across geographically diverse scenarios.

4.5. Limitations and future study

The segmentation results of PV panels highlight the effectiveness and robustness of the generated images and masks. However, a key limitation of using synthetic data is the potential for distributional bias. Since the generated images are based on limited real data, they may over-represent typical cases while underrepresenting rare or complex visual

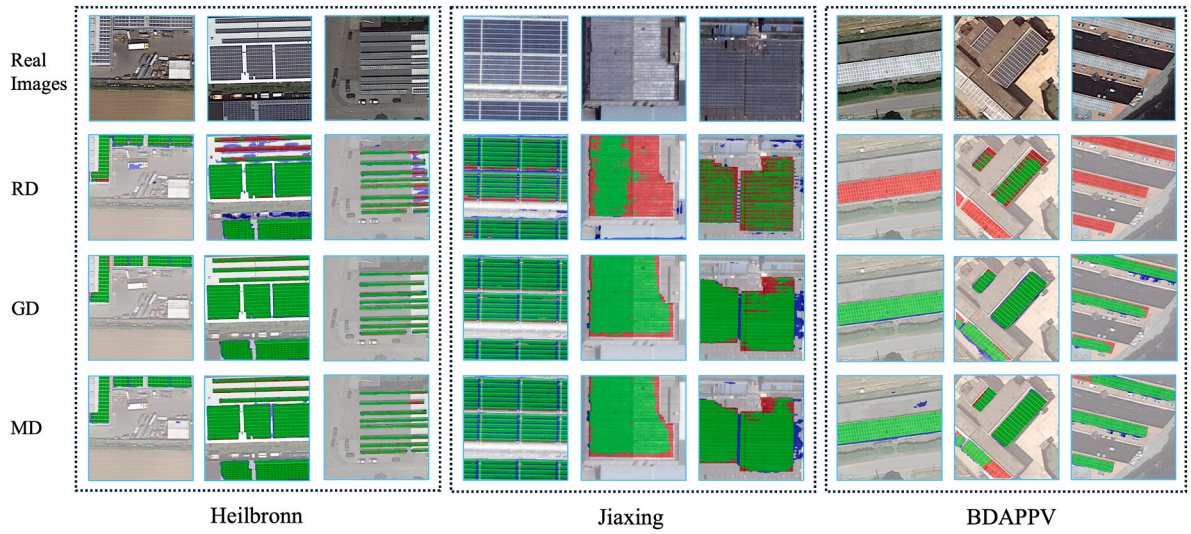


Fig. 9. Visualization results comparison of three experiments. RD: Pure real data training; GD: Pure generated data training; MD: Mixed data training.

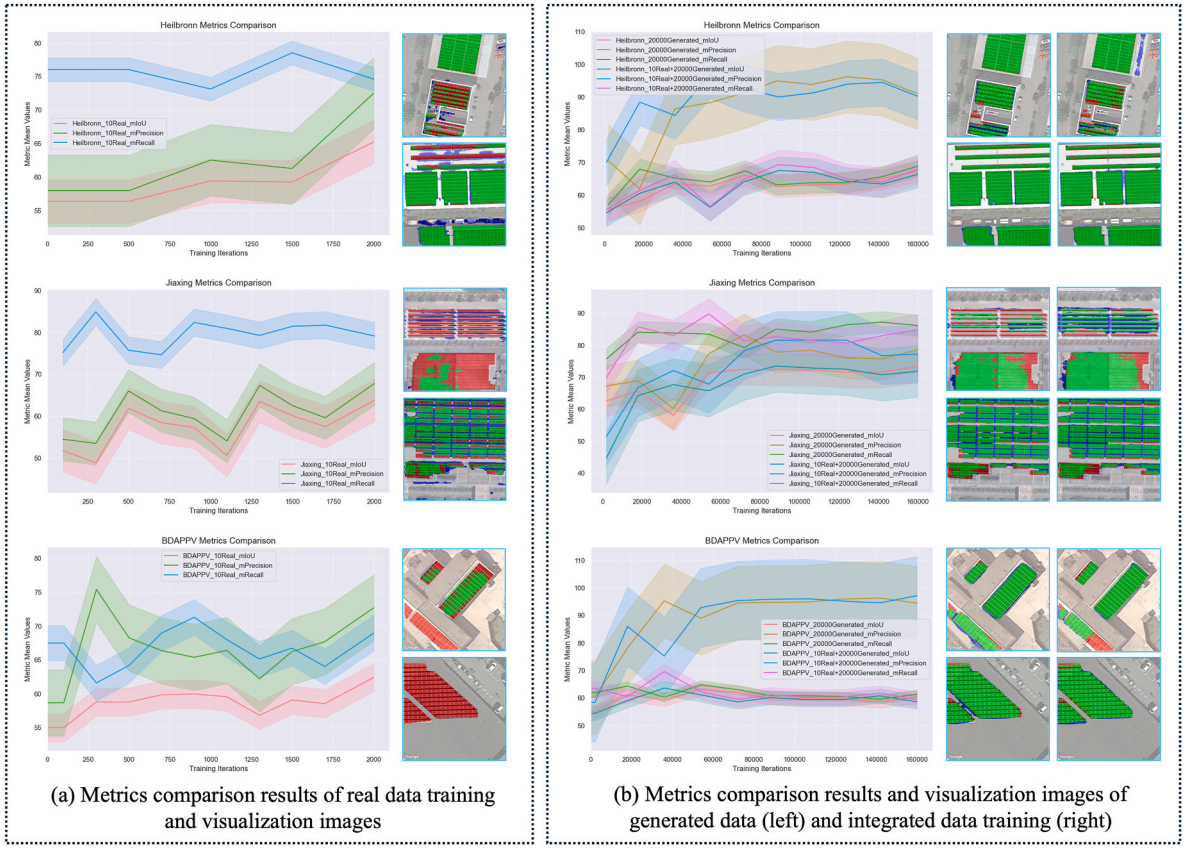


Fig. 10. Evaluation results of retraining and metrics comparison results.

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

City	Training with 10 real images (%)				Training with generated 20,000 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 %

Table 4
Evaluation of generalizability using generated data from Heilbronn combined with limited real data from three cities. IoU improvement = [IoU (MD) - IoU(RD)]/IoU (RD).

City	Training with 10 real images (%)				Training with 20,000 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 %

Table 5
Evaluation of generalizability using 10 real images from the PV01 dataset, trained with 20,000 generated data from Heilbronn, Jiaxing, and BDAPPV, respectively.

Training with 20,000 images generated from	Metrics (%)			
	IoU (PV)	mIoU	mPrecision	mRecall
Germany Heilbronn	73.95	81.70	93.01	87.58
Zhejiang Jiaxing	55.46	68.30	83.65	78.63
BDAPPV	70.88	79.01	89.16	87.12

features such as occlusions, inconsistent lighting, or background clutter. This limitation may reduce the model's generalizability across diverse real-world scenes. Another concern concerns the realism of the generated backgrounds. Despite efforts to diversify visual texture, synthetic backgrounds often lack the environmental complexity and natural variability of real remote-sensing imagery. This may lead to a domain gap between synthetic and real data, potentially hindering segmentation performance. Future work may involve integrating adversarial learning or hybrid generation techniques to more seamlessly blend synthetic foregrounds with real-world backgrounds. Additionally, as this study focused on cross-experiments on Heilbronn data, further validation through cross-region experiments using a broader range of real and synthetic datasets is essential to confirm the scalability and robustness of the proposed 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 Heilbronn-generated data 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 diverse geographic and background conditions makes it well-suited for large-scale renewable energy assessments, urban planning, and automated PV system mapping, yielding cost-effective and high-precision outcomes.

5. Conclusion

The limited accessibility and diversity of PV data hinder PV segmentation tasks. Remote sensing tools encounter obstacles such as high labelling costs and reliance on supervised training methods. To address these issues, this study leverages the SynthPV network, which integrates a Mask Decoder and a Text-to-Image based GenAI model, to generate PV images and corresponding masks. This aims to enhance model generalizability, data diversity, and segmentation robustness. The SynthPV network was utilized to create multi-background PV panel samples and aligned masks. The Mask Decoder was integrated and trained with less than 1 % real-world data. After experiments using real, synthetic, and integrated training data from three cities, the results showed significant improvements over purely synthetic training. The integrated training process increased IoU values by 9.60 %, 9.13 %, and 20.55 %, respectively. Transformative tests further confirm the effectiveness and robustness of the proposed method. Notably, synthetic datasets generated from Heilbronn exhibit strong generalizability in PV detection tasks when seamlessly integrated with other existing datasets. These results highlight the practical value of synthetic data generation as a scalable and cost-efficient solution to improving PV segmentation performance, thereby enabling more reliable PV mapping and supporting real-world applications in urban energy analysis and solar infrastructure deployment.

CRediT authorship contribution statement

Hongjun Tan: Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Zhiling Guo:** Writing – review & editing, Validation, Supervision, Conceptualization. **Jiaze Li:** Visualization, Methodology, Formal analysis. **Yuntian Chen:** Supervision, Software. **Qi Chen:** Software, Data curation. **Junwei Liu:** Writing – review & editing, Investigation. **Haoran Zhang:** Writing – review & editing, Supervision. **Jinyue Yan:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

We are grateful for the support and funding from the following projects: the 2024YFE0208700 - MOST: Urban Photovoltaic System Planning Method Considering Carbon Footprint and Environmental Benefits; P0052733 - RISUD: Cutting-edge Solar Synergies Integrated with 3D Urban Environments towards Carbon Neutral City; P0047700 - PolyU: International Centre of Urban Energy Nexus (UEX); P0056532 - IPT4U: Intelligent Platform & Toolbox for Urban Infrastructure Resilience; the High Performance Computing Centers at Eastern Institute of Technology, Ningbo, and Ningbo Institute of Digital Twin.

Data availability

Data and code will be made available on request.

References

[1] K. Chen, et al., Scalable spectrally selective solar cell for highly efficient photovoltaic thermal conversion, Adv. Appl. Energy 16 (2024) 100199.
[2] IEA, Renewables 2024, International Energy Agency, Paris, 2024.
[3] Y.M. Saint-Drenan, G.H. Good, M. Braun, A probabilistic approach to the estimation of regional photovoltaic power production, Sol. Energy 147 (2017) 257–276.
[4] M. Burke, et al., Using satellite imagery to understand and promote sustainable development, Science 371 (6535) (2021) eabe8628.

[5] J. Yang, et al., A novel method for assessment rooftop PV potential based on remote sensing images, Renew. Energy 237 (2024) 121810.
[6] S. Tian, et al., An innovative method for evaluating the urban roof photovoltaic potential based on open-source satellite images, Renew. Energy 224 (2024) 120075.
[7] P. Li, et al., Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning, Adv. Appl. Energy 4 (2021) 100057.
[8] M. Zech, H.-P. Tetens, J. Ranalli, Toward global rooftop PV detection with deep active learning, Adv. Appl. Energy 16 (2024) 100191.
[9] T. De Jong, et al., Monitoring spatial sustainable development: semi-automated analysis of satellite and aerial images for energy transition and sustainability indicators, arXiv preprint (2020) arXiv:2009.05738.
[10] Y. Zhang, A. Schlueter, C. Waibel, SolarGAN: synthetic annual solar irradiance time series on urban building facades via deep generative networks, Energy AI 12 (2023) 100223.
[11] S. Xun, et al., Generative adversarial networks in medical image segmentation: a review, Comput. Biol. Med. 140 (2022) 105063.
[12] K.C. Dong, et al., Advancing building facade solar potential assessment through AIoT, GIS, and meteorology synergy, Adv. Appl. Energy 17 (2025).
[13] Q. Chen, et al., Remote sensing of photovoltaic scenarios: techniques, applications and future directions, Appl. Energy 333 (2023) 120579.
[14] H. Jiang, et al., Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery, Earth Syst. Sci. Data 13 (11) (2021) 5389–5401.
[15] J. Yu, et al., DeepSolar: a machine learning framework to efficiently construct a solar deployment database in the United States, Joule 2 (12) (2018) 2605–2617.
[16] Z.L. Guo, et al., TransPV: refining photovoltaic panel detection accuracy through a vision transformer-based deep learning model, Appl. Energy 355 (2024).
[17] J. Wang, et al., PVNet: a novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery, Int. J. Appl. Earth Obs. Geoinf. 119 (2023) 103309.
[18] D. Jurakuziev, S. Jumaboev, M. Lee, A framework to estimate generating capacities of PV systems using satellite imagery segmentation, Eng. Appl. Artif. Intell. 123 (2023) 106186.
[19] Z. Qian, et al., Deep Roof Refiner: a detail-oriented deep learning network for refined delineation of roof structure lines using satellite imagery, Int. J. Appl. Earth Obs. Geoinf. 107 (2022) 102680.
[20] J. Song, et al., Synrs3d: a synthetic dataset for global 3d semantic understanding from monocular remote sensing imagery, Adv. Neural Inf. Process. Syst. 37 (2024) 117388–117425.
[21] G. Kasmi, et al., A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata, Sci. Data 10 (1) (2023).
[22] Y.Z. Chen, et al., Model-free renewable scenario generation using generative adversarial networks, IEEE Trans. Power Syst. 33 (3) (2018) 3265–3275.
[23] I. Goodfellow, et al., Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[24] F. Wang, et al., Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting, Energy Convers. Manag. 181 (2019) 443–462.
[25] H.R. Wen, et al., A regional solar forecasting approach using generative adversarial networks with solar irradiance maps, Renew. Energy (2023) 216.
[26] F. Wang, et al., Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting, Energy Convers. Manag. 181 (2019) 443–462.
[27] Q. Paletta, G. Arbod, J. Lasenby, Benchmarking of deep learning irradiance forecasting models from sky images – an in-depth analysis, Sol. Energy 224 (2021) 855–867.
[28] J.M. Bright, et al., A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration, Sol. Energy 147 (2017) 83–98.
[29] A. Browne, Using Stable Diffusion to Improve Image Segmentation Models, 2023.
[30] T. Sappakit, et al., Automated object keypoints dataset generation using blender. 2024 21st International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Ecti-Con 2024, 2024.
[31] J. Achiam, et al., Gpt-4 Technical Report, 2023 arXiv preprint arXiv:2303.08774.
[32] R. Rombach, et al., High-Resolution image synthesis with latent diffusion models. 2022 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr), 2022, pp. 10674–10685.
[33] C. Saharia, et al., Photorealistic text-to-image diffusion models with deep language understanding, Adv. Neural Inf. Process. Syst. 35 (2022) 36479–36494.
[34] Y. Zhu, et al., Resilience-Oriented Extreme Weather Conditional Renewable Scenario Generation Based on Diffusion Models and few-shot Learning, 2023.
[35] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[36] Z. Yuan, et al., Efficient and controllable remote sensing fake sample generation based on diffusion model, IEEE Trans. Geosci. Rem. Sens. 61 (2023) 1–12.
[37] M. Meiser, B. Duppe, I. Zinnikus, Generation of meaningful synthetic sensor data - evaluated with a reliable transferability methodology, Energy AI 15 (2024).
[38] R. Zhu, et al., Deep solar PV refiner: a detail-oriented deep learning network for refined segmentation of photovoltaic areas from satellite imagery, Int. J. Appl. Earth Obs. Geoinf. 116 (2023) 103134.
[39] W.J. Wu, et al., DiffuMask: synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. 2023 Ieee/Cvf International Conference on Computer Vision, Iccv, 2023, pp. 1206–1217.
[40] W.J. Wu, et al., DatasetDM: synthesizing data with perception annotations using diffusion models, Adv. Neural Inf. Process. Syst. 36 (2023). Neurips 2023.
[41] Q. Nguyen, et al., Dataset diffusion: diffusion-based synthetic dataset generation for pixel-level semantic segmentation, Adv. Neural Inf. Process. Syst. 36 (Neurips 2023) (2023).
[42] O. Ronneberger, P. Fischer, T. Brox, *U-Net: Convolutional networks for biomedical image segmentation.* Medical image computing and computer-assisted intervention, Pt Iii 9351 (2015) 234–241.
[43] E. Xie, et al., Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers, 2021.
[44] Y. Li, et al., Clip surgery for better explainability with enhancement in open-vocabulary tasks, arXiv preprint (2023) arXiv:2304.05653.
[45] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[46] MMSegmentation MmsegmentationContributors, OpenMMLab Semantic Segmentation Toolbox and Benchmark, 2020.
[47] H. Jiang, et al., Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery, Earth Syst. Sci. Data Discuss. 2021 (2021) 1–17.