Attention-Enhanced Diffusion With LLM-Driven Prompts for Controllable Defect Generation in Photovoltaic Cells

Ying He, Zechao Zhan, Jinxia Zhang, *Member, IEEE*, and Liping Chen

***Abstract*—Solar power is a vital clean energy source, and defects in photovoltaic cells critically impact power generation efficiency. While deep learning has advanced automated defect recognition, models face challenges due to scarcity and imbalance in defect samples. To overcome these challenges, a novel approach is proposed that exploits controllable image generation to produce diverse and realistic defect samples, while maintaining precise control over key attributes. First, to overcome the difficulty of generating accurate defect descriptions, a large language model is leveraged to automatically generate detailed text prompts. Next, to enable the model to learn domain-specific knowledge in photovoltaic cell defect recognition, the diffusion model is fine-tuned by DreamBooth. To further enhance the editability of these defects, a controllable mask, accompanied by the class name, is proposed to allow precise control over the type, size, and position of the defects. Finally, an attention enhancement module is proposed to improve both the precision and flexibility of defect generation, ensuring a more accurate and realistic simulation of defects in photovoltaic images. The quality of the generated images is assessed using diverse metrics, further demonstrating that the generated images significantly improve defect recognition performance.**

***Note to Practitioners*—Defect recognition in photovoltaic cells is crucial for maintaining the health of solar panels. In the field of automated photovoltaic cell defect recognition, although there is a significant need for extensive data to train models, acquiring high-quality images in the real world proves to be exceedingly difficult. Furthermore, the sample sizes between different types of defects are extremely imbalanced, leading to substantial disparities in detection performance across defect types. Additionally, due to the randomness of defect locations, defects can appear anywhere on the photovoltaic cells. This paper introduces an attention-enhanced diffusion with LLM-driven prompts, which enables practitioners to control defect generation by specifying the location, size, and type of defects through coordinates and textual prompts. This method has been validated on the UCF-EL and PVELAD datasets. The images generated significantly enhance the detection accuracy of the recognition network.**

***Index Terms*—Stable diffusion, Dreambooth, defect image generation, photovoltaic cell, class imbalance.**

I. Introduction

SOLAR energy, as a clean energy source [1], has seen rapid development in recent years. The economic viability of photovoltaic power generation has significantly improved, making it an increasingly attractive option for many countries and regions committed to energy transition [2]. In this context, the defect recognition and maintenance of photovoltaic cells are particularly vital as they directly impact the system's operational efficiency and return on investment. Photovoltaic cells are prone to various defects that can affect power output and usability due to transport and prolonged exposure to outdoor conditions.

Manual inspection of photovoltaic cells is costly, requiring high-quality defect samples and expert analysis to assess their impact on power generation efficiency [3]. Recent advances in deep learning and their successful industrial integration [4] have propelled the automatic recognition of photovoltaic cell defects into a focal point of current research [5]. This trend is evidenced by a growing body of literature, including [5], [6], [7], [8], [9], [10]. However, implementing this technology poses significant challenges: Defects in production are random and diverse, making it challenging to collect sufficient high-quality samples for training robust recognition models. Furthermore, in real-world power plants, the number of different types of defects is uneven, leading to missed recognition of rare classes.

To address these issues, traditional methods like data augmentation are commonly employed, altering the appearance of images without changing their labels to generate visually diverse yet consistent new images. Image transformations such as rotations [11], scaling, cropping, and flipping [12], as well as color adjustments that modify brightness, contrast, and saturation, simulate various lighting and environmental conditions. Introducing random noise such as Gaussian noise can mimic the effect of poor shooting quality. To tackle class imbalance, oversampling either replicates

Received 9 April 2025; revised 20 October 2025 and 11 January 2026; accepted 14 April 2026. Date of publication 20 April 2026; date of current version 24 April 2026. This article was recommended for publication by Associate Editor D. Wang and Editor D. Song upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Fund of China under Grant 62573123; in part by the Research Fund for Advanced Ocean Institute of Southeast University, Nantong, under Grant GP202411; and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Jinxia Zhang.)*

Ying He, Zechao Zhan, and Jinxia Zhang are with the Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China, and also with Advanced Ocean Institute of Southeast University, Nantong 226010, China (e-mail: heying719@seu.edu.cn; zhanzechao@seu.edu.cn; jinxiazhang@seu.edu.cn).

Liping Chen is with Das Solar Company Ltd., Wuxi 214021, China (e-mail: liping.chen1@das-solar.com).

Digital Object Identifier 10.1109/TASE.2026.3685808

minority-class samples or synthesizes new ones by interpolation, exemplified by the Synthetic Minority Over-sampling Technique (SMOTE) [13], whereas undersampling reduces majority-class samples to balance the dataset. Although these traditional image enhancement methods increase data diversity to some extent, they are limited to simple transformations of existing images and cannot create new image details or textures.

In this work, Stable Diffusion is exploited to accurately generate diverse defect samples. While diffusion models have been extensively explored in various fields [14], as far as we know, this is the first attempt to apply diffusion in the field of electroluminescence (EL) defect generation. Compared with natural scene image generation, defect generation of photovoltaic cells is more challenging due to the diversity of defects and imbalanced number between different types of defects. To address the above challenges, a novel framework is proposed via attention-enhanced diffusion with LLM-driven prompts for controllable defect generation in photovoltaic cells. First, since abstract defect names are not interpretable by the model, a large language model is employed to automatically generate detailed feature descriptions for defect images. Next, a pretrained diffusion model is fine-tuned by DreamBooth, enabling domain-specific adaptation to photovoltaic defect features. To further enhance editability, the defect name and a controllable mask mechanism are exploited to allow pixel-level manipulation of defect type, size, and position. Finally, an attention enhancement module is proposed to refine the focus of the model on defect-related regions during generation.

The main contributions of this paper are as follows:

- To address the scarcity and imbalance of defect samples, a controllable EL defect generator is proposed that combines prompted diffusion with an attention-enhanced consistency module. It supports multiple defect classes within a single backbone, enables control over defect type and position, and preserves EL background regularity.
- Considering the abstract nature of defect names, which alone cannot be fully understood by the model, a large language model is utilized to automatically generate accurate text prompts. These prompts are then used to fine-tune the diffusion model, enhancing its capability to generate highly precise and contextually relevant defect images.
- To address the high variability of defect characteristics, the method leverages defect names and a controllable mask mechanism that provides pixel-level control over defect type, size, and position. This approach achieves seamless integration of generated defects into background photovoltaic structures, ensuring both diversity and geometric consistency. To evaluate the alignment between generated defects and specified positions, and thereby verify the controllability of defect positions, we propose the metric of the Mean Spatial Offset (MSO).
- In order to enhance editing capability, an attention enhancement module is proposed to amplify the model's focus on defect-related regions during the diffusion process, improving the fidelity and editability of generated defects while preserving background integrity.
- Photovoltaic images exhibit specific structural priors, and existing evaluation metrics fall short of capturing domain-specific quality aspects. Therefore, the Busbar Spacing Assessment (BSA) metric is proposed, specifically designed to evaluate the structural integrity of photovoltaic images.

The remainder of this paper is organized as follows. Section II reviews related work on class imbalance and industrial defect image expansion. Section III introduces a controllable diffusion method that integrates several innovations for the generation of high-quality EL defect images. Section IV employs diverse methods to evaluate the generated images. Section V concludes the work.

II. Related Work

Class imbalance [15] is a well-known challenge in machine learning, particularly in domains like industrial defect recognition where some defect types or failure conditions are underrepresented. Addressing this issue is crucial for improving model performance and ensuring that classifiers generalize well across all classes, especially when the minority class is of high importance.

In this section, the existing remedies for class imbalance in inspection of industrial defects are introduced and summarized in detail: (i) traditional class-imbalance handling, (ii) GAN-based synthetic defects and (iii) Diffusion-based anomaly/defect generation. Additionally, to provide a broader perspective, we include (iv) cross-domain insights from medical imaging, focusing on synthetic data generation for medical applications.

A. Traditional Class-Imbalance Handling

A variety of strategies mitigate class imbalance. At the data level, SMOTE [13] interpolates minority samples, and DeepSMOTE [16] extends this idea with deep models to produce higher-quality synthetic data, though such manipulations may not guarantee domain-faithful realism. At the objective level, cost-sensitive learning adjusts the loss to emphasize rare classes—focal loss [17] down-weights easy cases to focus on hard examples, and Hellinger Net [18] leverages a skew-insensitive distance for improved imbalanced prediction. Model-level remedies include ensemble methods such as boosting [19], which iteratively reweight hard instances. In practice, hybrids that combine oversampling with ensembles (e.g., DeepSMOTE [16]) often outperform single techniques.

On EL images, classical data-level remedies such as SMOTE have been explicitly validated. Mantel et al. [20] combine SMOTE with Support Vector Machine (SVM) to improve cell-level defect prediction under severe imbalance. On the loss side, cost-sensitive objectives, notably focal-style losses, emphasize rare and hard examples in EL defect detection [21]. Moreover, ensemble methods remain strong baselines on EL datasets like ELPV [22]. Recent EL pipelines also integrate data augmentation with class weighting to further stabilize training under limited data [23].

Despite these advances, many approaches still rely on raw image manipulation or fixed reweighting and struggle to

generate diverse, domain-faithful samples, risking overfitting or limited rare-class gains. In contrast, our method uses controllable image synthesis via an attention-enhanced diffusion model with LLM-driven prompts to produce realistic, diverse defect images with fine control over key attributes, yielding more effective imbalance remediation and stronger classifier robustness.

B. GAN-Based Synthetic Defects

Generative Adversarial Networks (GANs) [24] have been widely used to generate synthetic defect samples. GANs consist of a generator that creates synthetic data and a discriminator that distinguishes between real and generated data. Earlier, Shou et al. [25] explore GANs for EL-based defect detection and anomaly segmentation, demonstrating the feasibility of GAN-driven synthesis and detection on EL imagery. Romero et al. [26] construct a DCGAN-based synthetic EL dataset covering multiple defect categories to alleviate severe class imbalance and also survey prior EL-GAN. Building on this line, Drir et al. [27] integrate a perceptual loss (VGG features) into DCGAN to improve the texture fidelity of synthesized EL images and report consistent gains on downstream defect classification. He et al. [28] introduce a progressive GAN-based approach for surface defect image generation, leveraging adversarial learning, cycle consistency, and self-attention to alleviate data scarcity and improve downstream defect detection accuracy. These EL-focused results support the usefulness of GAN augmentation for PV diagnostics while still inheriting typical GAN caveats.

However, GANs often struggle to generate truly diverse defect samples and may simply reassemble or manipulate existing training data, rather than creating completely new realistic defects. This limitation can affect the diversity of generated defect samples, which is crucial for training robust classifiers.

C. Diffusion-Based Anomaly/Defect Generation

Recent diffusion-based approaches have been explored for synthetic anomaly/defect data. AnomalyDiffusion [29] leverages a latent diffusion prior for few-shot anomaly synthesis with spatial guidance, disentangling anomaly appearance and location to improve controllability. StableSDG [30] adapts Stable Diffusion v1.5 to steel-surface defects via finetuning and image-oriented sampling that starts from partially noised real images to better preserve background structure. DualAnoDiff [31] couples global and local diffusion branches with cross-branch interaction to enhance image–mask coherence during denoising. AnomalyAny [32] performs test-time synthesis from a single normal image plus text prompts, using attention-guided optimization and CLIP-aligned refinement to improve semantic fidelity.

The aforementioned methods are proposed for other defect detection domains rather than the photovoltaic electroluminescence domain. To the best of our knowledge, diffusion-based approaches specifically tailored to EL defect synthesis have not yet been reported. While the above methods can generate realistic anomalies, they lack fine-grained control over the type, location, and scale of defects, as well as the ability to achieve seamless background integration. These limitations restrict their application in the PV EL domain. To better generate EL images, we integrate user-defined spatial priors with attention-driven, text-guided diffusion that aligns with PV geometric characteristics. This approach enables the controllable generation of defects while maintaining consistency with the inherent backgrounds of EL images.

D. Cross-Domain Insights From Medical Imaging

In the medical field, anomaly generation is crucial for tasks like lesion detection and diagnosis, especially when labeled data is scarce. One such innovation is LEGAN [33], which is introduced to address intraclass imbalance in GAN-based medical image augmentation. Furthermore, MINIM [34], a medical image-text generative model, combines the power of Stable Diffusion models with text inputs to generate medical images across different modalities, such as chest X-rays, MRIs, and OCT images. SegGuidedDiff [35] introduces a method for generating medical images with anatomical constraints by conditioning the generation process on anatomical segmentation masks.

While these medical image generation techniques offer robust solutions for constrained and condition-guided generation tasks, their direct applicability to PV defect analysis remains limited primarily due to substantial domain shifts between the two fields. The anatomical priors, imaging protocols and text vocabularies in medical datasets differ from electroluminescence physics and defect taxonomies. As a result, models tuned for organ structures and radiology semantics fail to capture busbars, metallization patterns, and the fine texture cues that govern photovoltaic defects and background continuity.

III. Method

To address the scarcity and imbalance of real samples, a controllable photovoltaic defect sample generation method based on attention-enhanced stable Diffusion with LLM-driven prompts is proposed. The overview of our proposed method is illustrated in Fig. 1. Initial image samples are processed to obtain high-quality images, and their descriptions are acquired through a large language model. Subsequently, DreamBooth is utilized to fine-tune the U-Net and text encoder of Stable Diffusion model, enhancing its capability to reproduce specific features of EL images. Finally, the fine-tuned attention-enhanced diffusion is implemented, combined with different normal images as backgrounds to achieve controllable generation of various types of defects through adjustable masks and name. This approach enables the controllable generation of images across various types, expanding and balancing the dataset.

A. Initial Sample Augmentation

In real-world photovoltaic cells, certain types of defects, such as interconnect and corrosion, are extremely rare. To better fine-tune Stable Diffusion and enable it to learn the characteristics of these scarce defects, these defect categories

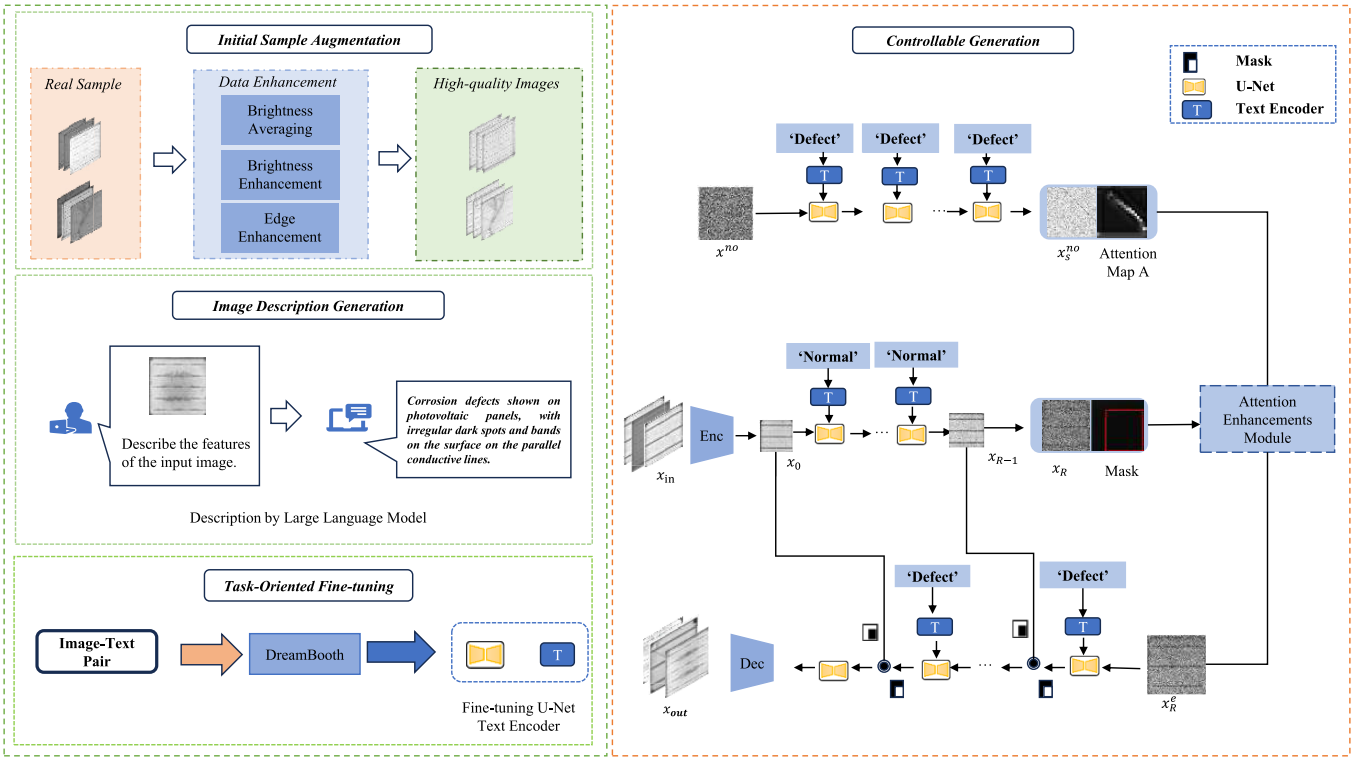


Fig. 1. The proposed framework integrates data processing, LLM-driven prompts generation, DreamBooth fine-tuning, and controllable defect generation with mask-guided editing and attention enhancement. The overview of our proposed method is as follows: First, images are processed through data processing (Section: III-A) and then input into the large language model to generate corresponding text prompts (Section: III-B). These prompts form high-quality image-caption pairs, which are input into DreamBooth for fine-tuning the text encoder and U-Net of Stable Diffusion (Section: III-C). This fine-tuned Stable Diffusion, which includes a mask adjustable in both size and position, is then used for controllable defect generation (Section: III-D). Finally, to enhance the editing capability, an attention enhancement module is proposed (Section: III-D3).

are augmented through simple replication to achieve the training data.

Moreover, to address the issue of inconsistent image brightness due to varying lighting conditions in the original sample set, systematic brightness normalization is conducted. By adjusting brightness and enhancing image edge features, the visual clarity of the images is improved and the identifiability of defect features is strengthened. The processed images are resized to a uniform resolution with 512×512 pixels to fit the requirements of subsequent Stable Diffusion fine-tuning.

B. Image Description Generation

Considering images in natural scenes, the definitions of PV cell defects (except for cracks) are relatively abstract, so it is necessary to supplement the abstract concepts with specific descriptions.

Thus, the images are input into a large language model [36], and in this study, ChatGPT-4 is chosen to automatically generate textual descriptions of the images. The prompt "Describe the features of the input image." is used to guide the model in generating appropriate textual descriptions for images of different defect categories. Accurate description of defects can provide more defect information for the model, which is necessary for the model to understand and generate more accurate defects.

C. Task-Oriented Fine-Tuning

Stable Diffusion is a powerful pre-trained model that contains vast knowledge of natural scenes. However, it lacks knowledge of photovoltaic scenarios. As a result, the model without fine-tuning is unable to generate proper EL images.

The text prompts obtained in III-B are combined with the corresponding images to form image-caption pairs, which are used for the DreamBooth fine-tuning of the U-Net and text encoder in Stable Diffusion. This fine-tuning step aims to make the model focus on specific features of photovoltaic cells when generating images.

During the fine-tuning process, the default settings of DreamBooth are adopted, with the objective of optimizing the model through the image-text pair reconstruction loss. This ensures that the generated images better align with the textual descriptions.

D. Controllable Generation

To generate diverse defect images and precisely control the image content, key attributes such as the type, location, and size of defects are adjusted based on input text and masks. Our framework builds on the diffusion model, which utilizes the fine-tuned Stable Diffusion as its backbone.

As shown in Fig. 1, our pipeline includes noising of normal image (section: III-D1), acquisition of the attention

map (section: III-D2), attention enhancement (section: III-D3), as well as text and mask guided editing (section: III-D4).

1) Noising of Normal Image: Firstly, different normal images x_{in} are used as input, which are then passed through the encoder to get the latent x_0. Subsequently, DDIM inversion process is applied to x_0. At each step, the latent x_r is deterministically transformed into the noisier latent x_{r+1} according to the DDIM inversion via Eq. (1). This deterministic noising process is conditioned on the normal text prompt n and the current timestep r. After R steps, ultimately the noisy latent x_R is obtained.

$$x_{r+1} = \sqrt{1-\alpha_{r+1}}\epsilon_\theta(x_r, r, n) + \sqrt{\alpha_{r+1}} f_\theta(x_r, r, n) \tag{1}$$

where α_r, known as the variance decay parameter, is a monotonically decreasing function that controls the scaling of noise levels across different time steps in the diffusion process. $\epsilon_\theta(x_r, r, n)$ is the U-Net-based noise prediction function that predicts the noise component at each time step r for the image x_r. Specifically, the U-Net takes the noisy latent x_r together with a timestep embedding r and the conditioning vector n, and outputs a noise map $\epsilon_\theta(x_r, r, n)$. The deterministic component function, $f_\theta(\cdot)$, is then calculated by $\frac{x_r - \sqrt{1-\alpha_r}\epsilon_\theta(x_r, r, n)}{\sqrt{\alpha_r}}$ where x_r is the noisy image at time step r, and $\epsilon_\theta(x_r, r, n)$ represents the noise predicted by the U-Net model at that step.

Starting from the original EL image x_{in}, an image noise addition operation is performed to obtain the noisy image x_R. Since the noisy image x_R does not contain the defect information that needs to be added, directly inputting x_R into the diffusion model may lead to insufficient editing capability. To address this issue, defect information needs to be provided, and an attention enhancement module is designed to obtain the improved noisy image.

2) Acquisition of the Attention Map: To provide defect information, the process starts with a random initial noise image x^{no} and uses the defect name as the input text c to perform the image denoising operation. Guided by the target text, the noisy input x^{no} is progressively denoised to obtain the denoised image x_S^{no}, with a total of S denoising steps. The Denoise process is based on the equation below:

$$x_{s-1}^{no} = \sqrt{\frac{\alpha_{s-1}}{\alpha_s}}\left(x_s^{no} - \frac{\alpha_{s-1} - \alpha_s}{\alpha_{s-1}\sqrt{1-\alpha_s}}\epsilon_\theta(x_s^{no}, s, c)\right) \tag{2}$$

where x_s^{no} is the noisy image at the s step, ϵ_θ is the noise prediction from the U-Net in diffusion model: At each step s, the U-Net accepts the current noisy latent x_s^{no} and the timestep s as its model-side inputs. Under the conditioning of text c, it generates and outputs the predicted noise $\epsilon_\theta(x_s^{no}, s, c)$. The denoising process iterates over a total of S steps and the final denoised latent x_S^{no} is obtained. At each DDIM step s, given the noisy latent x_s^{no} and the text condition c, cross-attention tensors can be obtained. By averaging over the text dimension, the attention map can be derived which indicates the position of the defect described in the text within the image. Averaging these maps across different steps yields the spatial attention map A.

3) Attention Enhancement Module: Experiments demonstrate that when editing photovoltaic cell images using a diffusion model, if the attention map corresponding to the input noisy image has higher values, the success rate of image editing is higher; conversely, the success rate is lower when the attention map values are lower. Based on this observation, an attention enhancement module is proposed. This module aims to enhance the editing capability of diffusion model on EL images by increasing the values of the attention maps corresponding to the input noisy images.

a) Attention-guided pixel selection: The attention mechanism focuses on identifying areas within an image that are most relevant to the task at hand. In this process, the binary mask M_{topk} is generated by selecting the top-k pixel positions P that exhibit the highest attention values in the attention map A. This selection prioritizes pixels that are most likely to contain crucial information about the defects:

$$P = \{(i, j) \mid A_{i,j} \in \text{TopK}(A, k)\} \tag{3}$$

$$M_{\text{topk}}(i, j) = \begin{cases} 1, & \text{if } (i, j) \in P, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This method ensures that the generation process is focused on areas of the image most likely to benefit from enhanced detail, thereby improving the accuracy and quality of the defect detection.

b) User-Constrained Masking: While the attention-guided mask autonomously identifies key regions, users can further refine the process by specifying a region of interest through coordinates. This step allows users to focus on specific areas they consider significant, resulting in a controllable mask M. The final mask M_f combines these user inputs with the attention-driven selections, providing a tailored approach to image editing:

$$M_f = M \odot M_{\text{topk}} \tag{5}$$

This combined masking ensures that both user preferences and algorithmic insights are utilized to enhance the targeted image regions effectively.

c) Pixel Replacement: The final stage in the image enhancement process involves replacing the pixels within the boundaries of the final mask M_f. Pixels in the noisy image x_R that fall within this mask are replaced with values from the denoised image x_S^{no}, which contains the refined information:

$$x_R^e = x_R \odot (1 - M_f) + x_S^{no} \odot M_f \tag{6}$$

This pixel replacement strategy is crucial for preserving the integrity of the background structures while selectively enhancing the areas identified for defect-related improvements. The approach ensures that enhancements are seamlessly integrated into the image, maintaining natural appearance while focusing on defect details.

4) Text and Mask-Guided Editing: During the image editing phase, the optimized noisy image x_R^e is edited based on the target text. The noisy original image and the intermediate results during the editing process are fused based on the mask, enabling the model to perform multi-defect editing within the mask while avoiding unnecessary edits outside the mask.

Finally, the refined noise map x_R^e guided by the target prompt, is used for DDIM sampling, where the intermediate results are integrated into the DDIM trajectory. In each denoising step, the region inside the mask is generated by the target

defect, while the region outside the mask retains the original image content.

$$x^e_{r-1} = \text{Denoise}(x^e_r) \odot M + x_{r-1} \odot (1 - M) \tag{7}$$

where x^e_r is the refined noise map at time step r. Denoise process is based on (2). Similarly, after R steps of denoising, the refined noise map x^e_R is further processed to x^e_0. The final output image x_{out} is then generated by the decoder.

In addition, during the defect image editing process, issues of poor seamlessness may arise, where the edits inside and outside the mask are not coordinated. Poor seamlessness arises from excessive noise introduced during editing, which destroys the original image information. Based on this analysis, an editing regulator is proposed to adjust the step:

$$R = p \cdot T \tag{8a}$$

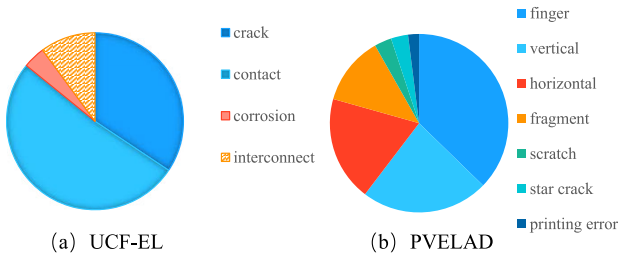
$$S = (1 - p) \cdot T \tag{8b}$$

where p is the ratio that controls the number of noising steps relative to the total number of steps, and T is the total number of steps. This approach addresses the issue of seamless editing when editing the original image into a specific defect image. The editing regulator can control the noise addition to the original EL image to stop early and input the less noisy image into the attention enhancement module. This way, the attention-enhanced noisy image retains more of the original information, avoiding the incoherence in image content caused by excessive editing.

Algorithm 1 Controllable Defect Generation Pipeline

Require: Original defect-free image x_{in}
Defect name c
Optional mask M
Fine-tuned diffusion model
Ensure: Generated defect image x_{out}
for each image x_{in} **do**
$x_0 \leftarrow \text{Encoder}(x_{\text{in}})$
$x_R \leftarrow \text{AddNoise}(x_0)$ *via* Eq. (1)
$x_s^{\text{no}} \leftarrow \text{Denoise}(x_R, c)$ *via* Eq. (2)
$A \leftarrow \text{CrossAttentionMap}(x_s^{\text{no}}, c)$
$x_R^e \leftarrow \text{AttentionEnhance}(x_R, M, A, x_s^{\text{no}})$
$M_{\text{topk}} \leftarrow \mathbf{1}\{\text{TopK}(A, k)\}$ *via* Eq. (3), (4)
$M_f \leftarrow M_{\text{topk}} \odot M$ *via* Eq. (5)
$x_R^e \leftarrow M_f \odot x_s^{\text{no}} + (1 - M_f) \odot x_R$ *via* Eq. (6)
for $r = R$**down to**1 **do**
$x^e_{r-1} \leftarrow \text{Denoise}(x^e_r) \odot M + x_{r-1} \odot (1 - M)$ *via* Eq. (7)
end for
$x_{\text{out}} \leftarrow \text{Decoder}(x_0^e)$
end for

To improve readability and highlight the relationships among all components, we summarize our controllable defect generation pipeline in Algorithm 1. Given an input EL image x_{in}, a defect name c, and an optional mask M, we first encode x_{in} to obtain latent x_0, then employ DDIM inversion to add noise to x_0, yielding x_R via Eq. (1). Then, guided by the text c, the noisy input x^{no} is progressively denoised to obtain the denoised image x_S^{no}, with a total of S denoising steps via Eq. (2). At each DDIM step s, given the noisy latent x_s^{no} and the text condition c, cross-attention tensors can be obtained. By averaging over the text dimension, the attention map can be derived which indicates the position of the defect described in the text within the image. Averaging these maps across different steps yields the spatial attention map A. Based on the mask M and the attention map A of the latent x_s^{no}, we enhance the values of the attention map corresponding to the input noisy image x_R to get the enhanced noisy map x_R^e via Eq. (6). Then, based on the enhanced noisy map x_R^e and the mask M, the mask–guided reverse diffusion is performed via Eq. (7) to get the latent x_0^e after R steps. Finally, the latent x_0^e is decoded to produce the controllable defect image x_{out}.



(a) UCF-EL (b) PVELAD

Fig. 2. Defect category quantity distribution on UCF-EL and PVELAD. (a) UCF-EL, (b) PVELAD.

IV. Experiment

This section first introduces the experimental settings and the dataset used, then validates the quality of the generated images and presents the results. Furthermore, since photovoltaic cell images contain busbars, an evaluation metric is proposed to assess the quality of generated defect images from the perspective of busbar generation quality.

The method is based on Stable Diffusion v1.5, utilizing the Adam optimizer with a batch size set to 1. A total of 280 images from each category are selected for fine-tuning. During fine-tuning, the learning rate is set to 1e-6, and for the text encoder, it is set to 5e-7. All training and generation processes are conducted using a single NVIDIA 3090 GPU.

The UCF-EL dataset [9] is used to evaluate our method, which contains 17,064 annotated images across four defect categories: *crack*, *corrosion*, *contact*, and *interconnect*. The crack defect appears as narrow black lines with clear boundaries or as intensity-gradient cracks [37]. Corrosion shows as darkened regions starting at the busbar and extending vertically [38]. Contact defects appear as small dark rectangles along the finger lines [39]. The interconnect defect manifests as bright spots along the busbar [40]. The distribution of the dataset is shown in Fig. 2.

The PVELAD dataset [41] is also used to evaluate our method; we evaluated 7 categories: *finger*, *vertical dislocation*, *horizontal dislocation*, *fragment*, *scratch*, *star crack*, and *printing error*. Finger appears as local breaks along metallization fingers; horizontal and vertical dislocation as grid or busbar misalignment along the x or y axis; fragment as irregular dark flakes; scratch appearing in EL images as thin, elongated low-luminance streaks along processing directions or near ribbons; star crack as star-shaped dark fissures; and printing error as widened or uneven metallization.

A. Evaluation Metrics

1) FID: Fréchet Inception Distance (FID) is a metric commonly used to evaluate generative models, particularly in image generation [42]. FID compares the similarity between generated and real images by extracting features from both sets using a pretrained deep learning model. A lower FID indicates higher resemblance to real images in visual content, making it a key metric for evaluating generative model realism. The mean vectors and covariance matrices of the two sets are used to compute the Fréchet inception distance, which quantifies the difference between their Gaussian distributions. It is calculated using the following formula:

$$FID = \|\mu_x - \mu_y\|^2 + \mathrm{Tr}(\Sigma_x + \Sigma_y - 2\sqrt{\Sigma_x \Sigma_y}) \tag{9}$$

where μ_x and μ_y represent the mean vectors of the real and generated images, respectively, while Σ_x and Σ_y denote their corresponding covariance matrices. The term Tr refers to the trace operation, which computes the sum of the diagonal elements of a matrix.

2) SSIM: Structural Similarity Index Measure (SSIM) [43] is a metric used to measure the visual similarity of two images. SSIM aims to provide a measurement metric that matches human visual perception, reflecting changes in image quality. Unlike traditional pixel-based difference metrics such as Mean Squared Error (MSE) or Peak Signal-to-Noise Ratio (PSNR), SSIM considers the structural information of images, which more closely aligns with how the human visual system operates. SSIM evaluates image quality by comparing luminance, contrast, and structure. These components are combined into a single value between -1 and 1, where 1 indicates perfect similarity, offering a more comprehensive assessment than pixel-based methods.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{10}$$

where μ_x and μ_y represent the mean intensities of reference and generated images, σ_x^2 and σ_y^2 represent their variances, σ_{xy} represents for their covariance capturing structural alignment, c_1 and c_2 are constants to stabilize the division with a weak denominator.

3) Proposed Busbar Spacing Assessment (BSA): In comparison to other industrial defect image generation techniques, a unique characteristic of photovoltaic cell defect generation is the presence of horizontally spaced busbars in the background. In real photovoltaic cell images, these busbars are evenly distributed. During the image generation process, issues such as well-rendered defects accompanied by uneven spacing between busbars may arise, as illustrated in Fig. 3. This can result in images where the busbar spacing appears irregular.

To address this, an evaluation metric called Busbar Spacing Assessment (BSA) is proposed to assess the quality of generated defect images specifically from the perspective of busbar spacing regularity. Given the labor-intensive nature of manual screening after generating a large number of images, the BSA index offers a way to automatically screen and evaluate the final generated images.

The process begins by detecting the horizontal busbars in generated images using the Hough transform. This method efficiently excludes edge artifacts and accurately extracts the positions of valid busbars. The spacing between adjacent busbars is then calculated using the following equations:

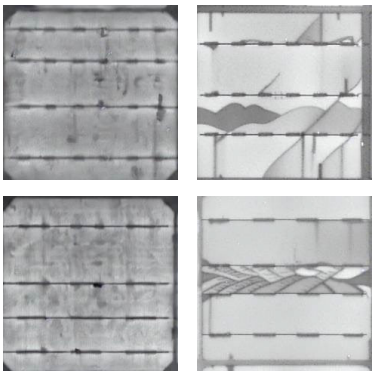


Fig. 3. Illustration of uneven busbar spacing in generated defect images.

$$d_i = |x_{i+1} - x_i|, \tag{11}$$

$$\mathrm{BSA} = \frac{1}{n-1}\sum_{i=1}^{n-1} d_i \tag{12}$$

where $i = 1, 2, \ldots, n-1$, x_i refers to the position of the i-th busbar, and d_i represents the distances between consecutive busbars. This metric is specifically designed to automatically filter out images from a large dataset where the busbar spacing does not meet the established standards to better align the generation results with real-world data distributions.

4) Proposed Mean Spatial Offset (MSO): To evaluate the alignment between generated defects and specified positions, and thereby verify the controllability of defect positions, we propose the metric of the Mean Spatial Offset (MSO). Given a user-specified mask, the coordinates of the mask's center point are denoted as $\mathbf{g}_i$. The actual region of the generated defect is denoted as $\mathcal{R}_i$, and its center point is denoted as $\mathbf{p}_i$. The offset between generated defects and specified positions is measured as the Euclidean distance between their center points in pixels:

$$d_i = \|\mathbf{p}_i - \mathbf{g}_i\|_2 \quad . \tag{13}$$

Averaging over N samples yields the Mean Spatial Offset (MSO), which summarizes overall alignment:

$$\mathrm{MSO} = \frac{1}{N}\sum_{i=1}^{N} d_i. \tag{14}$$

To make results comparable across different mask sizes with width w_i and height h_i, we report a normalized MSO that divides each Euclidean distance by the geometric mean of the mask size:

$$\mathrm{MSO}_{\mathrm{norm}} = \frac{1}{N}\sum_{i=1}^{N} \frac{d_i}{\sqrt{w_i h_i}}, \tag{15}$$

Lower MSO (and $\mathrm{MSO}_{\mathrm{norm}}$) indicates better spatial alignment between the user-specified location and the generated defect.

TABLE I

COMPARISON OF REALISM ACROSS DATASETS. EACH CELL SHOWS **FID↓ / SSIM↑**

Dataset	Category	AnoDiff [30]	DualAnoDiff [32]	Ours
UCF-EL	crack	167.64 / 0.981	212.34 / 0.983	**119.67 / 0.999**
	contact	184.81 / 0.982	203.64 / 0.984	**132.40 / 0.998**
	corrosion	169.28 / 0.987	100.29 / 0.984	**65.19 / 0.995**
	interconnect	223.01 / 0.988	226.54 / 0.986	**178.37 / 0.996**
PVELAD	finger	126.85 / 0.994	127.43 / 0.990	**77.30 / 0.998**
	vertical	160.40 / 0.993	136.47 / 0.993	**118.54 / 0.996**
	horizontal	176.43 / 0.991	192.16 / 0.992	**133.84 / 0.995**
	fragment	193.73 / 0.990	173.24 / 0.992	**112.67 / 0.997**
	star crack	91.61 / 0.992	142.70 / 0.993	**70.45 / 0.996**
	scratch	103.79 / 0.987	193.68 / 0.994	**96.61 / 0.996**
	print error	197.27 / 0.989	221.80 / 0.994	**107.88 / 0.997**
Mean (all)		163.17 / 0.989	175.48 / 0.990	**110.27 / 0.997**

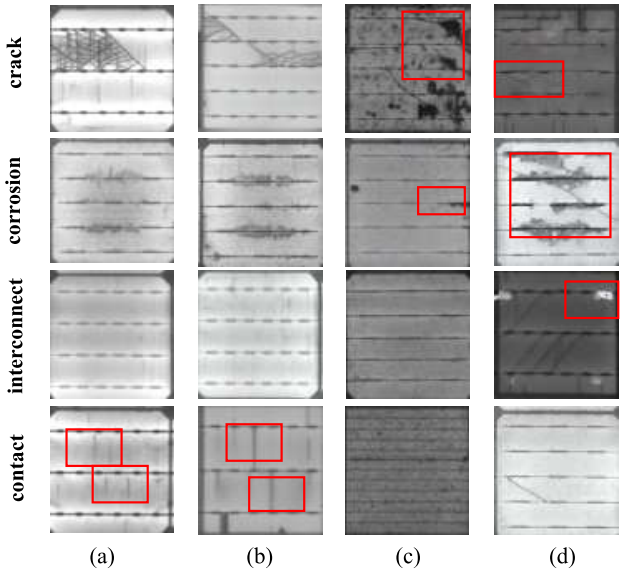


Fig. 4. Generated samples from the UCF-EL dataset. Columns: (a) Real samples, (b) Ours (c) DualAnoDiff, (d) AnoDiff.

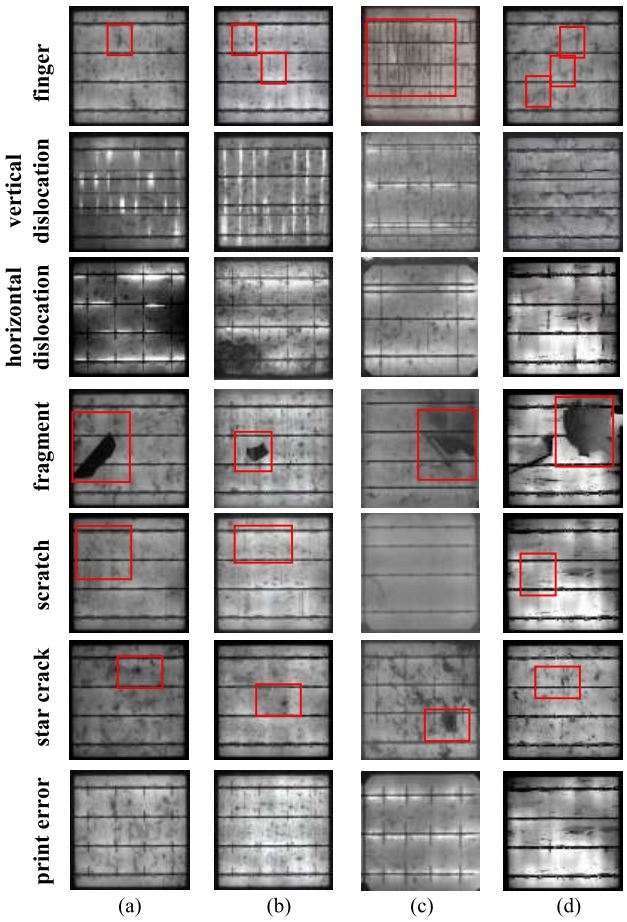


Fig. 5. Generated samples from the PVELAD dataset. Columns: (a) Real samples, (b) Ours (c) DualAnoDiff, (d) AnoDiff.

TABLE II

COMPUTATION AND MEMORY UNDER IDENTICAL SETTINGS

Method	FLOPs (TFLOPs)	Peak VRAM (MB)
AnoDiff	38.18	14066
DualAnoDiff	278.66	11402
Ours	84.04	12818

B. Results of the Generated Results

1) Comparison With Existing Methods: To evaluate the effectiveness of our method, we compare it with AnoDiff [29] and DualAnoDiff [31] across eleven defect categories using FID (↓) and SSIM (↑). The results are summarized in Table I. As shown, our method consistently achieves better performance across all defect types, with significantly lower FID and higher SSIM. Notably, improvements are particularly evident on challenging categories such as corrosion, fragment, and print error, where our approach produces more realistic and structurally consistent defect images. These results confirm that our method not only improves the overall fidelity of generated samples but also ensures robustness on UCF-EL dataset and PVELAD dataset across diverse defect categories.

In addition to the quantitative results in Table I, Fig. 4 and Fig. 5 provide side-by-side visual comparisons among AnoDiff [29], DualAnoDiff [31], and our method. As shown in Fig. 4, on UCF-EL, our method produces images that are highly similar to real EL images while preserving PV-specific structures. In contrast, AnoDiff [29] often breaks background consistency: for corrosion, there is inconsistent busbar thickness, and for interconnect the contrast with the background is excessively high, creating a pasted appearance. DualAnoDiff [31] shows weaker PV-structural fidelity, where the number of busbars is often incorrect and the generated defects are weak or missing, which indicates poor integration of defects into the EL geometry. Meanwhile, as shown in Fig. 5, on PVELAD, especially for small-footprint defects such as scratch and star crack, our generations remain closest to real images, preserving fine defect morphology and smooth tonal transitions. AnoDiff [29] and DualAnoDiff [31] tend to miss or attenuate the scratch signal, and for star crack they often over-amplify the pattern, which leads to unnaturally strong spokes and reduced realism. Overall, our results present clearer defect semantics with seamless background integration.

To evaluate each method in a more comprehensive manner, we also compare their computational costs by reporting FLOPs per inference and peak VRAM, as is shown in Table II.

TABLE III
COMPARISON OF RECOGNITION PERFORMANCE ACROSS METHODS (HIGHER IS BETTER)

Category	AnoDiff [30]	DualAnoDiff [32]	Ours
crack	76.47	31.37	**81.25**
contact	77.06	60.07	**77.17**
corrosion	76.00	67.27	**96.97**
interconnect	64.78	65.43	**76.39**
finger	58.85	50.69	**62.32**
vertical	34.63	64.06	**84.89**
horizontal	76.08	44.25	**88.58**
fragment	72.08	54.16	**73.54**
star crack	63.92	55.28	**66.98**
scratch	64.21	41.71	**77.03**
print error	47.65	52.86	**83.76**

TABLE IV
SPATIAL ALIGNMENT PER CATEGORY MEASURED BY MSO (PX) AND NORMALIZED MSO_{norm} (LOWER IS BETTER)

Dataset	Category	MSO ↓	MSO_{norm} ↓
UCF-EL	crack	42.49	0.188
	interconnect	40.48	0.157
	contact	53.99	0.229
	corrosion	20.92	0.052
PVELAD	finger	31.63	0.078
	scratch	24.92	0.153
	star crack	44.46	0.058
	fragment	17.36	0.052

Combined with the aforementioned computational cost and model performance results, our proposed method demonstrates superior controllability and performance compared with other diffusion-based methods, achieved under comparable computational resources and reasonable memory usage, ultimately striking a favorable balance between efficiency and quality.

To further quantitatively assess the performance of images generated by various methods on downstream tasks, we train the same recognition model (AlexNet) on datasets augmented via different methods and compare their accuracy in the defect recognition task. The training protocol, including the optimizer, learning rate schedule, batch size, as well as the training, validation, and test splits, remains identical for all methods; the only variable is the source of synthetic images.

As shown in Table III, our method achieves the highest recognition accuracy across all eleven defect categories. The gains are particularly evident on challenging classes such as vertical dislocation, horizontal dislocation, and corrosion, where our generated samples provide stronger task-relevant cues and preserve background context more effectively.

2) Region-Specific Defect Generation: In contrast, Fig. 6 demonstrates our framework's ability to apply dynamic masks for region-specific defect generation. This method involves the use of arbitrarily positioned and geometrically resizable masks, which are precisely applied to selected areas of the background. Despite the spatial restrictions imposed by these masks, the generated defects maintain contextual consistency with the surrounding areas. This targeted approach allows for detailed simulation of defects in specific regions.

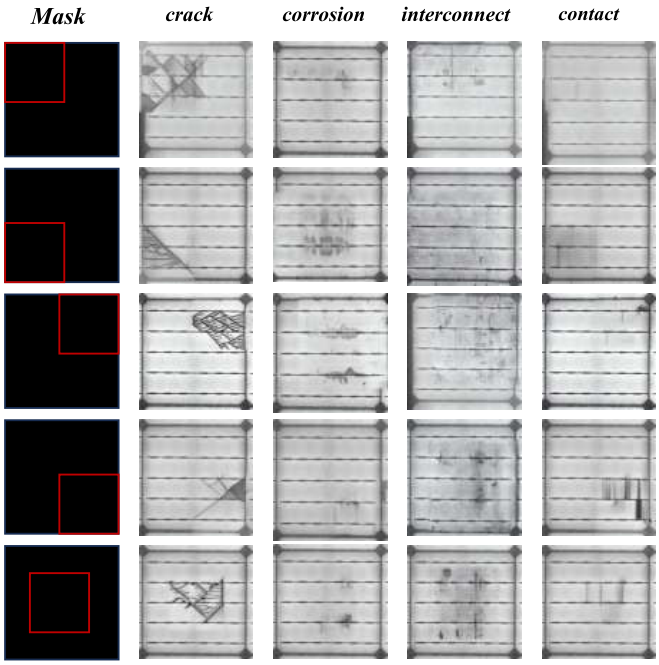


Fig. 6. Generated examples of different chosen masks.

To quantitatively evaluate the spatial precision of defect generation and verify the controllability of our method, we report the Mean Spatial Offset (MSO) and its normalized counterpart (MSO_{norm}) in Table IV. As shown in Table IV, on UCF-EL, corrosion shows the best localization (lowest MSO and MSO_{norm}), whereas contact is the most challenging. Averaged over the four categories, the error is 39.47 px and the normalized error is 0.157, indicating overall tight alignment with occasional drift on contact-like defects.

Considering practical constraints, only four categories in PVELAD support region-constrained generation (finger, scratch, star crack and fragment), the localization analysis is restricted to these classes. As shown in Table IV, PVELAD exhibits tighter localization overall, fragment attains the best alignment in both absolute and normalized terms, while star crack yields the largest absolute offset. These trends suggest that our spatial control remains robust across datasets, and that normalized MSO is complementary to absolute MSO in revealing scale-sensitive localization behavior.

3) User Study: To further evaluate the realism of the generated defect images, we conducted a user study. Specifically, we conducted a four-alternative forced-choice user study to quantify perceived realism. In each trial, a real EL image was shown as a reference, and participants were asked to select the candidate most similar to it from four anonymized options: the ground-truth real image, the image generated by our method, the image generated by AnoDiff, and the image generated by DualAnoDiff. The results show that across all trials, the real image received 56.5% of the votes, our method 29.8%, AnoDiff 8.6%, and DualAnoDiff 5.0%. As expected, the real image ranked first; importantly, our method ranked second and far surpassed the other two methods, indicating that our synthesized images are perceptually closer to real samples.

TABLE V
FOR REGION CONTROLLABLE GENERATION, FID AND SSIM SCORES ARE COMPARED USING DIFFERENT STRATEGIES

Data Processing	DreamBooth	LLM Aug	FID ↓				SSIM ↑			
			crack	corrosion	contact	interconnect	crack	corrosion	contact	interconnect
	✓		224.4331	75.7345	221.1510	265.2430	0.9902	0.9914	0.9735	0.9837
✓	✓		119.6673	121.8480	210.7875	211.4280	0.9987	0.9931	0.9668	0.9860
	✓	✓	240.5210	151.1030	273.4650	210.4955	0.9683	0.9925	0.9851	0.9845
✓	✓	✓	**119.6673**	**65.1891**	**132.4014**	**178.3675**	**0.9987**	**0.9945**	**0.9979**	**0.9956**

Note: ↑: Higher is better, ↓: Lower is better.

TABLE VI
FOR REGION CONTROLLABLE GENERATION, FID AND SSIM SCORES ARE COMPARED USING DIFFERENT STRATEGIES

Data Processing	DreamBooth	LLM Aug	FID ↓				SSIM ↑			
			crack	corrosion	contact	interconnect	crack	corrosion	contact	interconnect
	✓		170.8178	154.4403	188.3436	183.8118	0.9847	0.9915	0.9924	0.9953
✓	✓		199.2870	126.3882	231.5915	150.4041	0.9975	0.9898	0.9957	0.9927
	✓	✓	173.9832	124.8040	150.8214	152.8740	0.9938	0.9950	0.9903	0.9948
✓	✓	✓	**159.5300**	**117.2960**	**157.7958**	**147.1924**	**0.9980**	**0.9978**	**0.9969**	**0.9960**

Note: ↑: Higher is better, ↓: Lower is better.

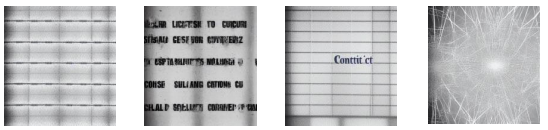


Fig. 7. Images generated by Stable Diffusion without fine-tuning.

The study results demonstrate that our synthetic data is perceived as highly realistic and comparable to real-world samples, providing validation beyond standard image similarity metrics.

C. Ablation Study

1) Impact of Model Fine-Tuning: As shown in Fig. 7, in contrast to the DreamBooth-fine-tuned model, the non-fine-tuned model fails to generate images relevant to EL images, producing outputs that lack any meaningful connection to the target domain.

2) Impact of Data Processing: Data normalization and augmentation during processing have enhanced the baseline for subsequent generation tasks. A comparison between rows 1 and 2 of Table V reveals that applying only data processing and DreamBooth significantly reduced the FID score and optimized the SSIM for crack and interconnect defects during full-image generation, indicating improved alignment with real data distributions. For region-controllable generation, the application of a controllable mask to the normal image as a background has enabled precise generation of specific defects within masked areas. As evidenced in row 2 of Table VI, the FID score for the corrosion class decreased from 154.4403 to 126.3882 after adopting data processing techniques.

3) Impact of LLM-Enhanced Prompts: The primary objective of LLM-enhanced prompts is to mitigate semantic ambiguity in abstract defect categories by enriching their textual descriptors. As shown in the comparison between

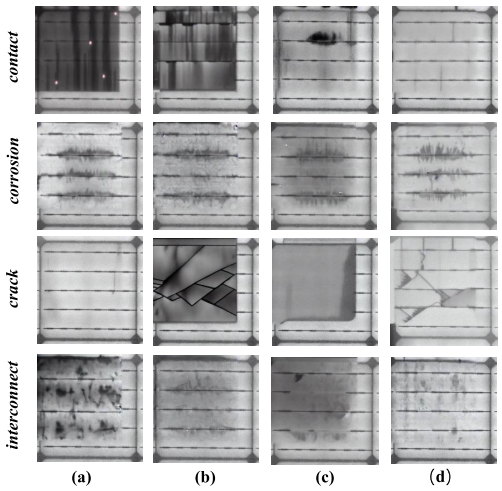


Fig. 8. Qualitative comparison of region-controllable generation under different experimental settings. (a) Without LLM-enhanced prompts. (b) Without data processing and LLM-enhanced prompts. (c) Without data processing. (d) Ours (Proposed method with all strategies).

Rows 1 and 3 of Table V, this strategy leads to substantial quality improvements. Specifically, for interconnect defects—previously the most difficult category to model—the FID score drops significantly from 265.2430 to 210.4955. This improvement extends to region controllable generation (Table VI), where our method achieves notable gains across most categories, demonstrating precise control over morphological features. These results underscore the critical role of LLM-driven prompt optimization in acting as strong semantic supervision to balance specificity with generalizability.

4) Effect of Our Pipeline: The integrated pipeline, combining data processing, DreamBooth adaptation, and LLM-enhanced prompts, achieved state-of-the-art performance. This success can be attributed to the effective coordination among the various phases: data processing ensures high-quality input; DreamBooth adaptation tailors the model for specific tasks;

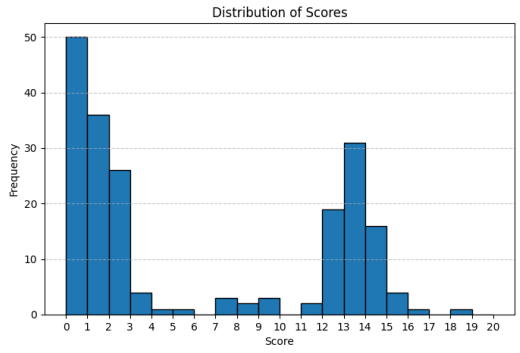


Fig. 9. The original real images scores distribution of BSA.

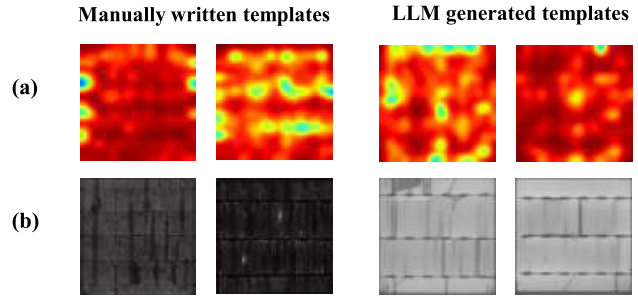


Fig. 10. Qualitative comparison of global generation samples and cross-attention maps for the contact category. The figure contrasts the Baseline (left block) and Ours (right block). (a) Cross-attention maps for global generation. (b) The resulting globally generated images.

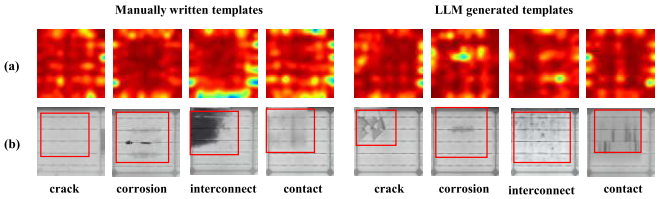


Fig. 11. Qualitative comparison of region controllable generation samples and cross-attention maps between the Baseline (left block) and Ours (right block). (a) Cross-attention maps. (b) The generated images.

and prompt fine-tuning optimizes text prompts to accurately align with the target outputs. As detailed in row 4 of Table V, the multi-stage approach significantly reduced the contact FID from 221.1510 to 132.40, while SSIM improved to 0.9979. Additionally, region controllable generation demonstrated promising results. As shown in rows 1 and 4 of Table VI, the FID score for corrosion defects decreased to a lower value of 117.2960, with an SSIM of 0.9978. As illustrated in Fig. 8, the model employing all strategies achieves the best editing results.

Beyond the quantitative metrics, models lacking comprehensive strategies or only employing partial solutions often encountered category confusion during image generation. In contrast, models implementing all strategies can independently generate images for each category based on specific textual prompts, effectively preventing confusion between different categories.

5) Impact of BSA: Firstly, the original images are evaluated using the BSA metric the distribution of the scores is analyzed. As shown in Fig. 9, the scores of most images are below 18. Meanwhile, experimental validation on 10,000 generated images demonstrates the effectiveness of this metric: 5% of the samples showed uneven spacing, where 95% of these cases showed moderate variance ($S < 18$) and only 5% showed severe misalignment ($S > 18$). As a result, in subsequent work using generated images, BSA score can be used to automatically filter the generated images. A threshold $S_{th} = 18$ is established to automatically filter out images with significant spacing irregularities $S > 18$, ensuring only physically plausible samples are retained.

D. Case Study: Why LLM Prompts Work Better

To comprehensively evaluate the interpretability and controllability of our method, we conduct a detailed case study comparing the proposed LLM-enhanced framework against simple, manually written prompt fine-tuning. For each case, we visualize both the generated images and the cross-attention maps of the object token.

1) Analysis of Global Generation: In the global generation task, as shown in Fig. 10, the baseline model exhibits attention maps with a rigid pattern, where high-activation regions align with the background busbars rather than the defect semantics. This indicates that the baseline model implicitly entangles defect features with the structural priors of the photovoltaic panel, failing to isolate the defect concept. This issue is particularly detrimental for the contact category. Due to the inconspicuous nature of contact defects, the baseline's inability to distinguish features leads to a complete failure in comprehension, resulting in failed generation. In contrast, our method guided by LLM prompts successfully disentangles the defect semantics from the background, manifesting as distinct, short vertical linear segments in the attention maps. This demonstrates a superior understanding of the contact class and achieves high-fidelity generation.

2) Region Controllable Generation: As shown in Fig. 11, the advantages of our method are further pronounced in the mask-guided region controllable generation task. For the baseline model, the attention maps in Row (a) are scattered and ambiguous, often incorrectly focusing on the background busbars rather than the defect itself. This distraction leads to distinct generation failures. First, for the crack category, the baseline's excessive fixation on rigid background structures results in a failure to synthesize the defect entirely. As seen in Row (b), the masked region remains virtually unchanged, appearing as a pure background image without any visible crack. Second, for categories like interconnect, the baseline erroneously focuses on irrelevant background areas instead of the target defect. This spatial misalignment directly leads to a failure in generating the specific defect features. Conversely, our method exhibits strong, well-defined attention clusters in Row (a). The attention is concentrated on the defect-relevant areas, and this precise attention guidance ensures that the generated defects in Row (b) are not only visually distinct but also seamlessly integrated into the designated regions.

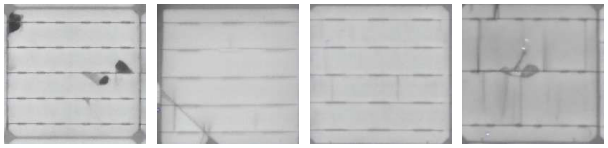


Fig. 12. Images generated using manually written templates with the prompt "crack".

Furthermore, real-world training images inevitably contain multiple defect types; for example, images labeled as crack frequently coexist with contact defects. This ambiguity causes the model to learn incorrect text-image mappings. Consequently, as shown in Fig. 12, the model trained without LLM prompts suffers from semantic confusion: when generating images with a specific category prompt, the model often fails to follow the text guidance, resulting in generated images that do not match the prompt.

Overall, this case study reveals the underlying reason behind the superior performance of LLM-enhanced prompts: instead of serving merely as a coarse label, they provide stronger semantic supervision that enhances attention alignment, improves editing stability, and ensures more reliable and controllable defect generation.

E. Discussion

The proposed method synthesizes natural, diverse defects across eleven categories while preserving PV-specific background structure. Remaining limitations include occasional failures or instability for categories such as unreported black core and corner, likely due to semantic–visual mismatch for abstract labels and limited priors for high-contrast, localized shapes that can induce attention drift. In addition, diffusion models still have room for improvement in inference time and resource consumption.

To rigorously enhance the generalization performance and stability of our controllable defect generation pipeline, future work could expand the testing scope to external, diverse real-world datasets. Additionally, the defect taxonomy could be broadened to encompass a larger variety of less common and more abstract defect types, such as potential induced degradation (PID) or subtle texture anomalies, providing detailed per-defect category reporting to track stability across all classes. This targeted expansion will ensure the synthetic data is robust and representative of the complexities found in varied industrial settings.

Furthermore, beyond FID and SSIM, more perceptually and physically grounded evaluation is needed. We propose strengthening assessment with structured human perception studies (building on our 4AFC user study) and incorporating expert reviews from PV module inspectors to verify the realism and diagnostic utility of generated defects.

Moreover, future research could focus on accelerating the inference process of generative diffusion models. Techniques such as model distillation, early-stopping during denoising, adaptive timestep scheduling, and lightweight backbone architectures could significantly reduce computational latency while maintaining generation fidelity. These optimizations would make the controllable defect synthesis pipeline more practical for real-time inspection and on-site deployment scenarios.

Another promising direction lies in minimizing the overall resource footprint of the model. Exploring quantization, low-rank approximation, and mixed-precision computation could lower GPU memory usage and power consumption during both training and inference. Besides, distributed training is also an effective approach to reducing resource consumption and shortening training time. Integrating these techniques can help achieve an efficient balance between model capacity and operational cost, supporting scalable deployment in industrial environments.

V. Conclusion

This paper addresses the challenges of limited real-world samples and class imbalance in photovoltaic cell defect recognition by proposing a novel framework that integrates an attention-enhanced diffusion model with LLM-driven prompts for controllable defect generation. The large language model is utilized to generate accurate descriptions for model fine-tuning. The introduction of the attention enhancement module significantly improves the precision of defect editing, enabling the generation of diverse types of defects independently using text prompts. Adjustable masks and text prompts facilitate precise control over the size, position, and types of defects. Additionally, the Busbar Spacing Assessment (BSA) metric has been introduced to aid in the selection of generated images, and we further propose the Mean Spatial Offset (MSO) (with a normalized variant) to quantitatively assess spatial alignment between user-specified regions and the generated defects.

The experiments on UCF-EL and PVELAD show that our method achieves superior generation metrics (lower FID and higher SSIM) and delivers higher recognition accuracy than competing approaches when trained on the generated images. The method not only enables the generation of realistic defect images, reducing the cost of acquiring real samples for industrial quality inspection, but also effectively balances the quantity of different categories within the dataset, ensuring a more uniform distribution. While promising, realism gaps may persist under rare morphologies or unseen capture conditions. Future work could expand cross-site evaluations, incorporate expert-in-the-loop assessments, and explore lighter backbones and training strategies to further improve efficiency and robustness in factory-scale deployments.

Acknowledgment

The authors thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this article.

References

[1] L. Hernández-Callejo, S. Gallardo-Saavedra, and V. Alonso-Gómez, "A review of photovoltaic systems: Design, operation and maintenance," *Sol. Energy*, vol. 188, pp. 426–440, Aug. 2019.

[2] E. Kabir, P. Kumar, S. Kumar, A. A. Adelodun, and K.-H. Kim, "Solar energy: Potential and future prospects," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 894–900, Feb. 2018.
[3] M. Abdelhamid, R. Singh, and M. Omar, "Review of microcrack detection techniques for silicon solar cells," *IEEE J. Photovolt.*, vol. 4, no. 1, pp. 514–524, Jan. 2014.
[4] L. Yu, W. Yu, Y. Jia, and T. Chai, "Real-time caustic ratio prediction in alumina digestion process for closed-loop operation: A cloud-edge deep learning approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 10787–10800, 2025.
[5] W. Tang, Q. Yang, K. Xiong, and W. Yan, "Deep learning based automatic defect identification of photovoltaic module using electroluminescence images," *Sol. Energy*, vol. 201, pp. 453–460, May 2020.
[6] B. Su, H. Chen, and Z. Zhou, "BAF-detector: An efficient CNN-based detector for photovoltaic cell defect detection," *IEEE Trans. Ind. Electron.*, vol. 69, no. 3, pp. 3161–3171, Mar. 2022.
[7] B. Su, H. Chen, P. Chen, G. Bian, K. Liu, and W. Liu, "Deep learning-based solar-cell manufacturing defect detection with complementary attention network," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 4084–4095, Jun. 2021.
[8] J. Balzategui, L. Eciolaza, and N. Arana-Arexolaleiba, "Defect detection on polycrystalline solar cells using electroluminescence and fully convolutional neural networks," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2020, pp. 949–953.
[9] J. Fioresi et al., "Automated defect detection and localization in photovoltaic cells using semantic segmentation of electroluminescence images," *IEEE J. Photovolt.*, vol. 12, no. 1, pp. 53–61, Jan. 2022.
[10] L. Pratt, D. Govender, and R. Klein, "Defect detection and quantification in electroluminescence images of solar PV modules using U-net semantic segmentation," *Renew. Energy*, vol. 178, pp. 1211–1222, Nov. 2021.
[11] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, Dec. 2019.
[12] L. Pérez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*.
[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.
[15] V. W. de Vargas, J. A. S. Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. V. Barbosa, "Imbalanced data preprocessing techniques for machine learning: A systematic mapping study," *Knowl. Inf. Syst.*, vol. 65, no. 1, pp. 31–57, Jan. 2023.
[16] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 6390–6404, Sep. 2023.
[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[18] T. Chakraborty and A. K. Chakraborty, "Hellinger net: A hybrid imbalance learning model to improve software defect prediction," *IEEE Trans. Rel.*, vol. 70, no. 2, pp. 481–494, Jun. 2021.
[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[20] C. Mantel et al., "Machine learning prediction of defect types for electroluminescence images of photovoltaic panels," *Proc. SPIE*, vol. 11139, May 2019, Art. no. 1113904.
[21] S. Ding, W. Jing, H. Chen, and C. Chen, "Yolo based defects detection algorithm for EL in PV modules with focal and efficient IoU loss," *Appl. Sci.*, vol. 14, no. 17, p. 7493, Aug. 2024.
[22] L. Pratt, J. Mattheus, and R. Klein, "A benchmark dataset for defect detection and classification in electroluminescence images of PV modules using semantic segmentation," *Syst. Soft Comput.*, vol. 5, Dec. 2023, Art. no. 200048.
[23] J. Wang et al., "Deep-learning-based automatic detection of photovoltaic cell defects in electroluminescence images," *Sensors*, vol. 23, no. 1, p. 297, Dec. 2022.
[24] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[25] C. Shou et al., "Defect detection with generative adversarial networks for electroluminescence images of solar cells," in *Proc. 35th Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Oct. 2020, pp. 312–317.
[26] H. F. M. Romero et al., "Synthetic dataset of electroluminescence images of photovoltaic cells by deep convolutional generative adversarial networks," *Sustainability*, vol. 15, no. 9, p. 7175, Apr. 2023.
[27] N. Drir, A. Mellit, M. Bettayeb, and M. Dhimish, "Enhanced photovoltaic defect detection using perceptual loss in DCGAN and VGG16-integrated models on electroluminescence images," *IEEE J. Photovolt.*, vol. 15, no. 6, pp. 759–769, Nov. 2025.
[28] X. He, Z. Luo, Q. Li, H. Chen, and F. Li, "DG-GAN: A high quality defect image generation method for defect detection," *Sensors*, vol. 23, no. 13, p. 5922, Jun. 2023.
[29] T. Hu et al., "AnomalyDiffusion: Few-shot anomaly image generation with diffusion model," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 8, pp. 8526–8534.
[30] Y. Tai, K. Yang, T. Peng, Z. Huang, and Z. Zhang, "Defect image sample generation with diffusion prior for steel surface defect recognition," 2024, *arXiv:2405.01872*.
[31] Y. Jin et al., "Dual-interrelated diffusion model for few-shot anomaly image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 30420–30429.
[32] H. Sun, Y. Cao, H. Dong, and O. Fink, "Unseen visual anomaly generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 25508–25517.
[33] H. Ding, N. Huang, Y. Wu, and X. Cui, "LEGAN: Addressing intraclass imbalance in GAN-based medical image augmentation for improved imbalanced data classification," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.
[34] J. Wang et al., "Self-improving generative foundation model for synthetic medical image generation and clinical applications," *Nature Med.*, vol. 31, no. 2, pp. 609–617, Feb. 2025.
[35] N. Konz, Y. Chen, H. Dong, and M. A. Mazurowski, "Anatomically-controllable medical image generation with segmentation-guided diffusion models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2024, pp. 88–98.
[36] F. Bie et al., "RenAIssance: A survey into AI text-to-image generation in the era of large model," 2023, *arXiv:2309.00810*.
[37] L. Papargyri et al., "Modelling and experimental investigations of microcracks in crystalline silicon photovoltaics: A review," *Renew. Energy*, vol. 145, pp. 2387–2408, Jan. 2020, doi: 10.1016/j.renene.2019.07.138.
[38] N. Iqbal et al., "Characterization of front contact degradation in monocrystalline and multicrystalline silicon photovoltaic modules following damp heat exposure," *Sol. Energy Mater. Sol. Cells*, vol. 235, Jan. 2022, Art. no. 111468, doi: 10.1016/j.solmat.2021.111468.
[39] M. Köntges et al., "Review of failures of photovoltaic modules," Int. Energy AgencyPhotovoltaic Power Systems Programme (IEA PVPS), Paris, France, Tech. Rep. IEA-PVPS T13-01:2014, Mar. 2014.
[40] D. J. Colvin, E. J. Schneller, and K. O. Davis, "Impact of interconnection failure on photovoltaic module performance," *Prog. Photovolt., Res. Appl.*, vol. 29, no. 5, pp. 524–532, May 2021, doi: 10.1002/pip.3401.
[41] B. Su, Z. Zhou, and H. Chen, "PVEL-AD: A large-scale open-world dataset for photovoltaic cell anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 404–413, Jan. 2023.
[42] Y. Yu, W. Zhang, and Y. Deng, "Frechet inception distance (FID) for evaluating GANS," China Univ. Mining Technol. Beijing Graduate School, Beijing, China, Tech. Rep., 2021, vol. 3.
[43] D. M. Rouse and S. S. Hemami, "Understanding and simplifying the structural similarity metric," in *Proc. 15th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 1188–1191.

Ying He received the B.S. degree in building electrical and intelligentization from Nanjing Tech University in 2023. She is currently pursuing the master's degree with the School of Automation, Southeast University. Her current research focuses on the generation and segmentation of industrial defect images. Her recent work is centered on balancing and supplementing the existing photovoltaic panel defect segmentation dataset using diffusion models.

Zechao Zhan received the B.E. degree in automation from Harbin Institute of Technology (HIT), China, in 2023. He is currently pursuing the M.E. degree in electronic information with the School of Automation, Southeast University, China. His research interests include AIGC, multi-modal image editing, multi-modal vision-language understanding, image generation, and computer vision.

Jinxia Zhang (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China, in 2009 and 2015, respectively. She was a Visiting Scholar with the Visual Attention Laboratory, Brigham, Women's Hospital, and Harvard Medical School, from 2012 to 2014. She is currently an Associate Professor with the School of Automation, Southeast University. Her research interests include saliency detection, knowledge transfer, computer vision, and machine learning.

Liping Chen received the Master of Optical Engineering degree, the M.B.A. degree from Tongji University, and the master's degree from the School of Science, Jiangnan University. She was a Senior Engineer and a Vice Principal of Jiangsu (Suntech) Institute for Photovoltaic Technology, and off-campus cooperative advisor. She joined Suntech (Wuxi) in August 2005. She has been engaged in research and development and technical management of highefficiency solar cells for more than 16 years. During her work, she participated in eight nationallevel, provincial-level, and international cooperative scientific research projects as the core technical backbone, applied for 24 patents, and applied for 11 patents as the first inventor, which had four invention patents and two utility model patents authorized by China National Intellectual Property Administration. She was awarded the title of "Young and Middle-aged Experts with Outstanding Contributions in Wuxi City" in 2020, and won the First Prize of Jiangsu Photovoltaic Science and Technology Award in 2020 and the Second Prize in the Science and Technology Progress Award of Chinese Renewable Energy Society in 2020.