# Referring Solar Cell Defect Segmentation in Electroluminescence Images

Shenghao Dong, Jinxia Zhang, *Member, IEEE*, Yu Shen, and Dehong Gao

***Abstract*—In the photovoltaic (PV) power generation field, accurately identifying solar cell defects based electroluminescence (EL) images is essential for maintaining high efficiency for PV power plants. Current solar cell defect segmentation methods typically segment all defects in the EL image uniformly, making it difficult to precisely identify specific defects according to maintenance needs. This limitation hinders personalized defect detection for smart operation and maintenance of PV power plants. To solve this problem, a novel task referred to as referring solar cell defect segmentation (RSCDS) is proposed in this article. The goal of the RSCDS task is to precisely segment the specified solar cell defects based on the referring text, tailored to the personalized maintenance requirements of actual PV power plants. Given the lack of relevant datasets, an RSCDS dataset is developed, abbreviated as Ref-EL-defect, comprising 60 000 pairs of defects and corresponding referring texts. The referring text can indicate single defect, multiple defects, or even no defects at all in the EL image, and such multigranularity correspondence enables accurate and personalized segmentation of defects. In addition, a multimodal multigranularity segment network is designed for the RSCDS task. By exploiting the characteristics of the solar cell defects, the multimodal fusion module and multigranularity perception grouping module are proposed to better adapt to the RSCDS task. State-of-the-art (SOTA) referring expression segmentation models designed for natural scene images are transferred to the RSCDS task, and experimental results demonstrate that the proposed method outperforms the SOTA models.**

***Index Terms*—Deep learning, defect segmentation, electroluminescence (EL) image, multimodal learning, photovoltaic (PV), referring expression segmentation (RES), solar cell.**

## I. Introduction

SOLAR energy, as a renewable resource, holds immense potential to address the global energy crisis. The efficiency and reliability of solar systems are predominantly determined by the performance of solar cells, which serve as their core component. However, solar cells are susceptible to various defects, such as cracks, finger interruptions, and corrosion during manufacturing, transportation, and operation. These defects can severely degrade the performance, safety, and longevity of solar systems. Consequently, the timely and accurate detection and segmentation of such defects are crucial for maintaining the optimal functionality and sustainability of solar energy systems.

In the early stages, defect detection in solar cells primarily relied on manual visual inspection. However, this method was not only time-consuming and labor-intensive but also highly susceptible to human error. For large-scale photovoltaic (PV) power plants, such manual approaches are impractical and economically unviable. Recent advancements in computer vision and deep learning have made automatic defect segmentation feasible, and deep learning-based models have shown great potential in this field [1], [2], [3], [4], [5], [6].

Defects in solar cells, which vary by type, size, and location, have distinct impacts on power loss and overall system performance. For instance, large-sized defects pose an immediate threat to structural integrity and require urgent intervention, while small-sized defects, though seemingly insignificant at first, may develop into more severe issues over time. Moreover, defects in contact with the busbar have a more pronounced impact on system performance compared to those located elsewhere. However, as illustrated in Fig. 1, traditional defect segmentation methods based on deep learning often rely on a unified semantic segmentation strategy. Since these methods only segment defects by type, they operate within a coarse-grained segmentation framework, making it difficult to capture finer attributes, such as defect size and location. When specific defect attributes need to be identified, previous traditional methods typically segment all defects of different types and then manually filter the target defects with the specific attributes from the

Received 2 February 2025; revised 28 March 2025; accepted 17 April 2025. This work was supported in part by the Research Fund for Advanced Ocean Institute of Southeast University, Nantong under Grant GP202411, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011435, in part by the Fundamental Research Funds for the Central Universities, the Natural Science Basic Research Program of Shaanxi under Grant 2024JC-YBMS-513, in part by the Key Research and Development Program of Zhejiang Province under Grant 2024C01025, and in part by the Key Research and Development Program of Hangzhou under Grant 2024SZD1A23. Paper no. TII-25-0708. *(Corresponding author: Jinxia Zhang.)*

Shenghao Dong and Yu Shen are with the Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China (e-mail: 220242228@seu.edu.cn; 230169411@seu.edu.cn).

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

Dehong Gao is with the School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Binjiang Institute of Artificial Intelligence, Hangzhou 310056, China (e-mail: dehong.gdh@nwpu.edu.cn).

Digital Object Identifier 10.1109/TII.2025.3567396

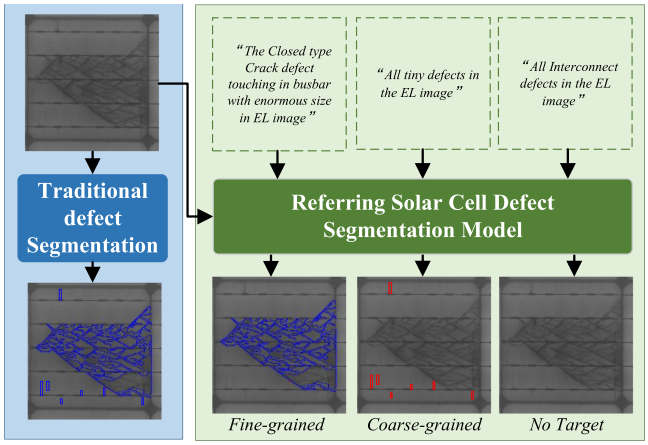


Fig. 1. Visual comparison between traditional defect segmentation tasks and our RSCDS task: For solar cell EL images containing multiple defects, traditional defect segmentation tasks apply a uniform segmentation strategy to all defects. In contrast, RSCDS task can accurately segment the specified defect based on the referring text. By combining image and text information, it enables more accurate segmentation of defects in solar cells.

segmentation results. The above-mentioned process is resource-intensive and susceptible to subjective bias, resulting in unreliable outcomes. This limitation significantly hinders the ability of existing segmentation approaches to meet the personalized demands of PV power plants. Consequently, there is an urgent need to develop a framework capable of accurately segmenting defects based on detailed and diverse requirements.

To address the above-mentioned limitation, a new task termed "referring solar cell defect segmentation (RSCDS)" is proposed. This task introduces a more refined defect segmentation framework. As illustrated in Fig. 1, RSCDS leverages referring text to specify the defects of interest in EL images, enabling precise segmentation tailored to the unique maintenance needs of PV plants. For example, given the referring text "all small-sized defects in the EL image," RSCDS can accurately segment all minor defects, meeting preventive maintenance requirements in real-world scenarios and preventing them from evolving into more severe issues.

Moreover, a multimodal dataset for RSCDS comprising 60 000 EL image–text pairs was constructed. The RSCDS dataset combines image processing techniques with text templates to generate textual descriptions for various defects in the images. To achieve flexible and robust segmentation, fine-grained templates were designed for precise single-defect segmentation, and coarse-grained templates were created for multi target defect segmentation. In addition, no-target templates were constructed to address cases where the described defect is absent in the image, thereby enhancing the dataset's robustness.

Furthermore, a model called multimodal multigranularity segmentation network (M2SegNet) is designed for the RSCDS task. To better integrate the multimodal data in the RSCDS task, a multimodal fusion (MMF) module is developed, which utilizes attention mechanisms to associate the referring text with corresponding regions in the image. To address the multigranularity aspect of the referring defect segmentation task, a multigranularity perception grouping (MPG) module is proposed. Inspired by characteristics of human visual cognition, this module enables the model to perceive targets ranging from coarse to fine granularity based on perceptual grouping method. It classifies and groups pixels of the feature map through learnable grouping queries.

In summary, the contributions of this study are as follows.

1) The RSCDS task is introduced in this study, designed to enable precise segmentation of specified defects in electroluminescence (EL) images for personalized maintenance requirements.
2) A large-scale dataset named Ref-EL-Defect, containing 60 000 EL image–text pairs, is presented in this study. To the best of our knowledge, this is the first large-scale multimodal dataset specifically developed for solar cell defect segmentation.
3) A novel framework called M2SegNet, which integrates MMF and MPG modules, is proposed in this study. Through experiments, state-of-the-art (SOTA) performance on the RSCDS task is demonstrated by the proposed method, surpassing existing referring image segmentation models tailored for natural scenes.

## II. Related Work

### A. Referring Expression Segmentation (RES)

In the current field of multimodal deep learning, RES is an important task. By deeply understanding and processing image–text information, RES enables the precise segmentation of relevant objects in an image based on referring text. This task has made significant progress in various visual domains, including natural scenes and camouflaged targets [7], [8], [9], [10], [11].

Hu et al. were the first to propose the RES task. RES originated from a similar task called referring expression comprehension (REC) [7], [12], [13], which produces the bounding box of a target in an image based on textual descriptions. The first datasets for RES and REC were ReferIt [8]. Yu et al. [9] later used the Microsoft COCO dataset as the image data source to construct the RefCOCO natural scene RES dataset, which has become widely used in research. Zhang et al. [10] developed the R2C7K dataset based on the task of camouflaged object detection, creating a large-scale referring camouflaged object segmentation dataset from real-world camouflaged object images. Liu et al. [11] further extended the concept by proposing the generalized RES task, which allows referring text to point to any number of target objects. They constructed the gRefCOCO dataset and proposed a baseline model called ReLA to better handle multitarget matching scenarios. To the best of our knowledge, existing RES datasets mainly focused on natural scenes.

However, currently there is no multimodal image–text dataset available for solar cell defect detection or segmentation, making it difficult to detect specific defect based on the personalized maintenance needs of each power plant.

### B. Solar Cell Defect Inspection

Noncontact, high-efficiency defect inspection using computer vision have become mainstream methods in the solar cell field [14], [15], [16], [17], [18].

Some studies have explored defect inspection using thermal infrared imaging. Shen et al. proposed a modified U-Net incorporating batch normalization and RMSprop to achieve accurate segmentation of PV arrays in complex thermal images, thereby facilitating subsequent defect detection within the PV array. Since eddy current thermography (ECT) exhibits advantages in detection sensitivity and spatial resolution compared to conventional thermography, Du et al. [18] integrated ECT with a convolutional neural network, demonstrating superior defect inspection performance through experimental comparisons. In addition, the study shows that ECT is more sensitive to external defects, such as surface scratches and impurities [18].

Due to its effectiveness in diagnosing internal defects, EL imaging has been widely adopted for solar cell defect inspection [19]. Dhimish and Mather [1] proposed a microcrack detection system for manufacturing execution. Otamendi et al. [2] created a scalable framework for automatic defect labeling in EL images. Zhang et al. [3] introduced a novel lightweight and high-performance defect detection model for solar cell EL images, utilizing neural architecture search and knowledge distillation. Fioresi et al. [4] developed the large-scale "UCF-EL" dataset, which includes 17 064 high-quality EL images, and built a corresponding defect segmentation method based on the DeepLab v3+ network. Wang et al. [5] introduced the RERN deep learning network, enabling more accurate segmentation of defect areas by the extraction and refinement of edge features. Zhang et al. [6] enhanced convolutional features using global similarity and saliency modules, outperforming five baseline models. To enhance defect representation, Yang et al. [20] proposed a fusion method that combines electrothermography and EL images, leveraging an L1-norm-based sparse representation algorithm to effectively integrate information from both modalities.

Despite their impressive performance, existing methods inspect all defects in a uniform manner, neglecting the specific maintenance requirements of PV systems. For example, coastal power plants must prioritize corrosion detection, whereas those in arid regions focus on thermal stress cracks. This limitation prevents existing approaches from effectively addressing the personalized inspection of specific defects.

## III. Dataset Construction and Statistics

Due to the lack of multimodal datasets containing text descriptions of PV defects, a multigranularity and multimodal RSCDS dataset in EL images, abbreviated as Ref-EL-Defect, is constructed to achieve personalized defect segmentation in solar cells.

### A. Dataset Construction

This section constructs an RSCDS dataset, i.e., Ref-EL-Defect, based on the publicly available solar cell defect segmentation dataset UCF-EL [4]. To automatically and accurately construct referring texts that describe personalized maintenance needs, four basic attributes of defects were summarized based on prior knowledge of solar cells. Table I details these key attributes: defect category, defect type, defect size, and defect location. For the images in the UCF-EL, image processing techniques are employed to extract the four attributes of each defect in the images. Next, text templates are designed using these basic defect attributes. With these templates, referring texts describing maintenance needs can be automatically generated. By utilizing various text templates, a wide range of referring texts can be produced.

TABLE I
Attributes and Characteristics of Solar Cell Defects

| Attribute | Description |
|---|---|
| Category-type | **Crack-closed:** A closed seam that does not significantly affect current flow.<br>**Crack-resistive:** Impairs current flow to the ribbon while maintaining battery connectivity.<br>**Crack-isolated:** Completely isolates a section of the cell from the ribbon (e.g., lobe cracks).<br>**Interconnect-disconnected:** Broken interconnections, often due to soldering failures.<br>**Interconnect-highly resistive:** Weak interconnections with high resistance, often caused by over-soldering.<br>**Contact-corrosion:** Corrosion caused by moisture entering cell gaps.<br>**Contact-finger interruption:** High strain at solder joints, visible as dark rectangles near busbars. |
| Location | **Intersected with busbar:** Defects that touch the busbar.<br>**Not intersected with busbar:** Defects that do not touch the busbar. |
| Size | **Small size:** Minor defects covering fewer pixels.<br>**Large size:** Significant defects covering more pixels. |

Based on relevant research in solar cell defects [14], the defects are divided into three major categories, further subdivided into seven defect types. The characteristics of each defect type are shown in the upper part of Table I.

The middle part of Table I lists the location attributes of defects. Since the busbar is the conductive strip that captures and converges current, defects touching with the busbar directly affect the core path of current convergence, significantly impacting the power output of the module. Therefore, the defects in solar cell are classified into two major categories based on whether the defect touches with the busbar.

As shown in the lower part of Table I, solar cell defects can be categorized by size into small size and large size defects. By drawing on the definition of small objects in the field of natural scenes [21], if the ratio of the defect area to the image area is less than 0.6%, it is defined as a defect with small size. Otherwise, it is identified as a defect with large size. The size of a defect directly affects the performance of the solar cell.

Table II presents some of the text templates designed in this study, which are used to construct various referring texts. As shown in the upper part of Table II, Each text template includes descriptions of the four different attributes of a solar cell defect, ultimately forming referring texts such as *"The* **[Type]** **[Category]** *defect(s)* **[Location]** *with* **[Size]** *in EL image."* This

TABLE II
GRANULARITY AND REFERRING TEXT TEMPLATES

| Granularity | Referring text templates |
|---|---|
| Fine granularity | The [Type] [Category] defect(s) [Location] with [Size] in EL image.<br>The [Category] defect(s) is [Type] and [Location] with [Size] in EL image. |
| Coarse granularity | All [Category/Size/Location] defects in EL image.<br>The [Category] failures with [Size] in EL image.<br>Various types of defects in EL image. |

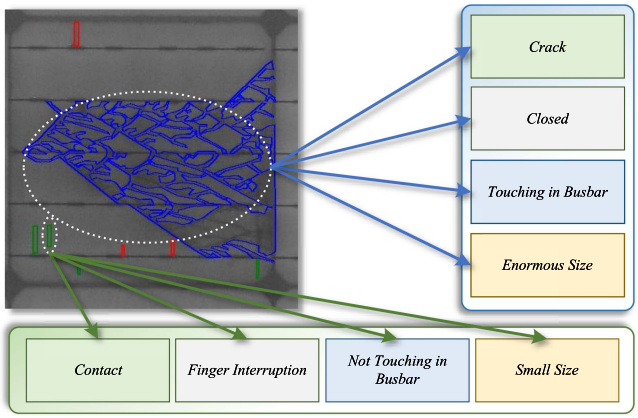


Fig. 2. Schematic diagram of the four attributes of solar cell defects. These four attributes are defect category, defect type, defect size, and defect location.

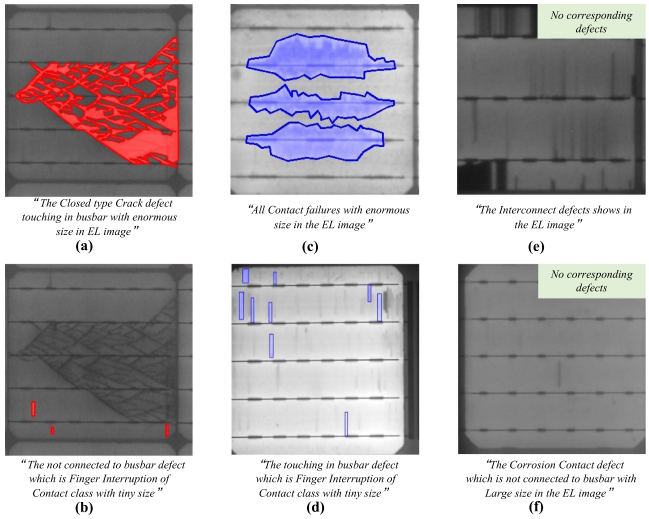


Fig. 3. Samples in the Ref-EL-Defect dataset, where (a) and (b) are the fine-grained samples, (c) and (d) are the coarse-grained samples, and (e) and (f) are the no-target samples.

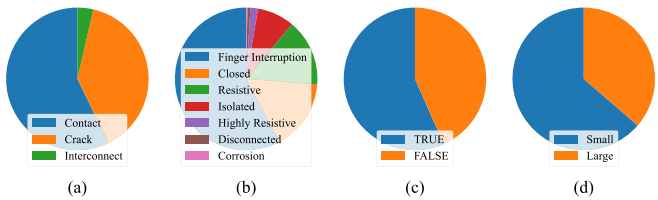


Fig. 4. Statistical Analysis of the Four Attributes from the Referring text Templates in the Datase.(a) depicts the quantity distribution of different defect attributes (e.g., contact, crack, interconnect); (b) details the types of defects, including finger interruption, closed, resistive, etc.; (c) shows the spatial distribution relative to the busbar ("TRUE" indicates defects in contact with the busbar, while "FALSE" denotes those not in contact); and (d) categorizes defects by size as either "Small" or Large.

method allows for the creation of fine-grained referring text-specified defect relationship. Fine-grained relationship refers to a single referring text corresponding to a specific defect in the EL image.

In order to enhance the robustness of the RSCDS dataset, it is also necessary to consider coarse-grained and no-target scenarios. Coarse-grained text represents a single referring text corresponding to multiple solar cell defects. As shown in the lower part of Table II, coarse-grained text templates are designed to automatically generate the corresponding referring texts. For the no-target scenario, referring texts are designed with attributes unrelated to all defects in the EL image.

For a specific defect, the four basic attributes of the defect are identified in Fig. 2. Based on the text templates provided in Table II, referring texts describing the specified defects are automatically generated. By designing diverse text templates, multiple referring texts can be generated for a single defect, thereby enhancing the diversity of the dataset.

### B. Data Statistics

Examples from the Ref-EL-Defect dataset are illustrated in Fig. 3, where (a)–(b) showcase fine-grained samples, (c)–(d) coarse-grained samples, and (e)–(f) no-target samples. The dataset comprises 60 000 referring texts, each corresponding to specific defects in EL images. These samples are categorized into three groups: fine-grained (40%), coarse-grained (40%), and no-target (20%). Fig. 4(a) and (b) focuses on the category and type analysis. The dataset contains the highest number of defects in the contact category, with finger interruption being the most frequent type. These defects also have the highest occurrence rate in real-world PV systems, providing the model with a realistic and practical learning environment. In Fig. 4(c), defects are classified as either "True" (touching the busbar) or "False" (not touching). The balanced distribution between these two types ensures comprehensive data representation. Fig. 4(d) shows that, small defects make up the majority of the dataset, accounting for 70%, while large defects represent the remaining 30%. Although larger size defects are less numerous, their impact on the actual conditions is equally significant due to the greater number of pixels cover.

## IV. PROPOSED FRAMEWORK FOR RSCDS TASK

Compared to traditional defect segmentation tasks, the RSCDS task is required to segment the corresponding defects based on a referring text. As a result, the RSCDS task involves processing multimodal information from both images and text. In addition, a referring text may correspond to one defect, multiple defects, or no defect at all. Therefore, the image–text

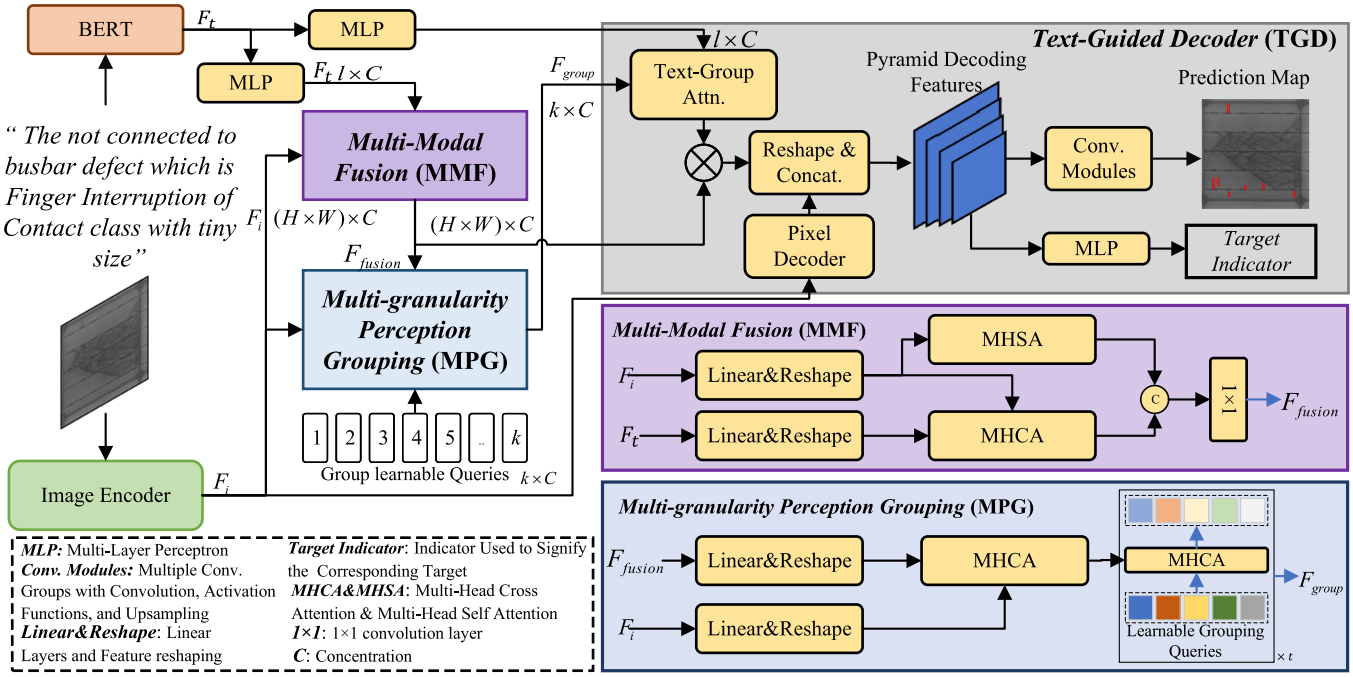


Fig. 5. Framework of the proposed M2SegNet. It consists of three main components: the MMF module, the MPG module, and the TGD. In the MMF module, text and image features are deeply integrated to generate fusion features. The MPG module progressively identifies solar cell defects from coarse to fine using learnable grouping queries, producing grouping features. These features are then combined and decoded in the TGD to generate the segmentation mask.

correspondence in this task exhibits multigranularity characteristics.

Section IV-A introduces the task definition of RSCDS task. Section IV-B presents the overall architecture of the proposed M2SegNet model. Sections IV-C and IV-D provide detailed descriptions of our MMF module and MPG module, which are used to handle the alignment between modalities and the processing of image–text information at different granularities. Section IV-E elaborates on the text-guided decoder (TGD), which further refines and segments the specified defects under the guidance of the text.

## A. Task Description

The goal of the RSCDS task is to segment specified defects in EL images of solar cells, based on the input referring text, thereby achieving personalized segmentation of solar cell defects, which facilitates subsequent operation and maintenance of PV power plants. The input for the RSCDS task consists of two parts: the first part is the EL image of the solar cell, denoted as $I^{\mathrm{EL}} \in \mathbb{R}^{H\times W\times 3}$, where $H$ and $W$ represent the height and width of the EL image, respectively; the second part is the referring text of the target defect to be detected, denoted as $T^{\mathrm{Ref}} \in \mathbb{R}^{l}$, where $l$ represents the length of the input referring text. The final output is $I^{\mathrm{mask}} \in \mathbb{R}^{H\times W\times 1}$, which represents the segmentation mask corresponding to the referring text in the image.

## B. Overall Architecture

The overall framework of the proposed model is depicted in Fig. 5. In this study, the Swin Transformer is employed as the image encoder to extract image features $F_i \in \mathbb{R}^{H\times W\times C_1}$, where $H$ and $W$ denote the spatial dimensions of the features, and $C_1$ denotes the channel dimension. The input referring text is encoded using BERT, producing text features $F_t \in \mathbb{R}^{l\times C_2}$, where $l$ is the length of the referring text and $C_2$ represents the embedding dimension of the text features.

The text features $F_t$ and image features $F_i$ are first fed into the MMF module. This module fuses and aligns the image and text features based on the attention mechanism, outputting the fusion features $F_{\mathrm{fusion}} \in \mathbb{R}^{H\times W\times C}$, which are image features guided by the referring text.

To better handle multigranularity information, the fusion features are further input into the MPG module. In this module, the learnable grouping queries $Q_{\mathrm{group}} \in \mathbb{R}^{k\times C}$ are introduced, where $k$ denotes the number of perception groups. The proposed MPG module can aggregate pixels of the same category within the feature map. In this study, different numbers of queries are defined at different scales, accompanied by the execution of the cross-attention mechanism $t$ times. It can gradually perceive the target defect from coarse-grained to fine-grained, with fewer categories set at deeper features and more categories at shallower features. The final output of the MPG module at each scale constitutes the grouping features $F_{\mathrm{group}} \in \mathbb{R}^{k\times C}$.

Subsequently, the text features, image features, grouping features, and fusion features are all sent to the TGD, which outputs the segmentation mask $I_{\mathrm{mask}} \in \mathbb{R}^{H\times W\times 1}$, achieving precise segmentation.

## C. Multimodality Fusion

To achieve efficient and precise MMF, a multimodal feature fusion module is designed. This module effectively fuses the

text features with the image features. At each feature scale, a submodule is employed to fuse the image features and text features at that specific scale, as shown in Fig. 5. Multiple submodules are used to perform the fusion at different scales, generating a MMF feature set that contains features at various scales.

The MMF module is based on the attention mechanism. Specifically, the input text features $F_t$ are first aligned the feature scale by a linear layer and reshape operation. The input image features $F_i$ are aligned the feature scales in the same way, resulting in reshaped features of size $HW \times C$. Then, cross-attention between text and image features and self-attention for the image features are performed parallel. The cross-attention helps to capture the correspondence between the referring text and image regions, while the self-attention focuses on the locations of defects within the image.

The results of the cross-attention between text and image features and the self-attention for the image features are concatenated and then fused through a $1 \times 1$ convolution to produce the image features guided by the referring text. The specific process can be defined as follows:

$$F_{\text{fusion}} = \text{Conv}_{1\times1}(\text{Concat}(\text{MHCA}(\gamma^t, \gamma^i), \text{MHSA}(\gamma^i))) \quad (1)$$

$$\gamma^i = \text{Linear}(\text{Reshape}(F_i)), \quad \gamma^t = \text{Linear}(\text{Reshape}(F_t)) \quad (2)$$

where $\text{Conv}_{1\times1}$ represents the $1 \times 1$ convolution, Concat denotes feature concatenation, MHCA and MHSA stand for multihead cross-attention (MHCA) and multihead self-attention mechanism, respectively, and $\gamma^t$ and $\gamma^i$ represent the dimension-aligned text features and image features, respectively.

In this model, feature fusion is conducted at different scales of the image features, ultimately generating multiscale fusion features as the output.

### D. Multigranularity Perception Grouping

To accurately identify multigranularity targets, the MPG module is proposed. Based on perceptual grouping characteristics, it enables a progressive identification from coarse to fine granularity.

Inspired by learnable query-based framework [22], learnable grouping queries are designed to analyze components of varying granularity within the feature map. Unlike single-modal query-based segmentation framework, MPG is tailored for multigranularity segmentation and integrates multimodal information, enhancing the robustness of defect identification.

As shown in Fig. 5, first, the feature map $F_i$ extracted by the image encoder at a certain scale and the fusion feature $F_{\text{fusion}}$ extracted in the MMF submodule are aligned in feature dimensions through linear layers. It can be defined as

$$\gamma^i = \text{Linear}(\text{Reshape}(F_i)), \gamma^{\text{fusion}} = \text{Linear}(\text{Reshape}(F_{\text{fusion}})) \quad (3)$$

where $\gamma^i$ and $\gamma^{\text{fusion}}$ represent the image features and fusion features after dimension alignment, respectively.

Subsequently, the MPG module needs to effectively adapt to multimodal features. The aligned features are then passed through MHCA, allowing the model to inject the image features into the fusion features, thereby enhancing robustness of the model. It can be defined as

$$f_i' = \text{MHCA}(\gamma^i, \gamma^{\text{fusion}}). \quad (4)$$

Instead of treating queries as instance-level representations [22], the MPG module introduces learnable grouping queries that dynamically adapt to the granularity of different feature levels. These queries allow the module to dynamically adjust its focus based on the granularity of the current module and the content of the image. Each query acts as a semantic cluster center, progressively refining the defect representation. Specifically, the grouping features $f_i^{\text{GP}}$ are generated through $t$ iterations of cross-attention

$$Q_{\text{learnable}}^{j+1} = \text{MHCA}_j(f_i', Q_{\text{learnable}}^j), \quad j \in [0, t-1] \quad (5)$$

$$F_{\text{group}} = \text{MHCA}_{t-1}(f_i', Q_{\text{learnable}}^{t-1}). \quad (6)$$

The core of the MPG module is the progressive perceptual grouping mechanism. To achieve progressive perceptual grouping, the MPG module adaptively adjusts the number of grouping queries based on feature depth. Different numbers of learnable queries are set at different feature scales. The MPG submodule sets fewer group queries for deep features (corresponding to coarse granularity) and gradually increases the number of group queries for shallow features (corresponding to fine granularity). This enables multigranularity solar cell defect perception grouping, from coarse to fine granularity. With $n$ different scales of image features, $n$ different granularity grouping features can be obtained. The numbers of learnable queries will be discussed in Table VII in the ablation study.

### E. Text-Guided Decoder

In the TGD, the perceptual grouping feature $F_{\text{group}}^j$ on scale $j$ first facilitates cross-attention mechanism with the text feature $F_t$, yielding text-enhanced grouping features $F_\gamma^j$. These enhanced features are then multiplied by the text-guided fusion features $F_{\text{fusion}}^j$ at scale $j$, resulting in the decoding features $F_{\text{decode}}^j$ for that scale. Through this process, multiscale decoding features $\{F_{\text{decode}}^j\}_{j=1}^n$ are obtained based on the multiscale image features. The workflow of scale $j$ is simplified as follows:

$$F_\gamma^j = \text{MHCA}(F_{\text{group}}^j, F_t^j), F_{\text{decode}}^j = F_\gamma^j \times F_{\text{fusion}}^j \quad (7)$$

where MHCA represents the MHCA mechanism, and the superscript $j$ indicates the current scale.

In addition, to extract and reconstruct pixel-level features from the input image to optimize segmentation results, the image features $F_i$ are passed through a pixel decoder for presegmentation [22]. The result of the operation are the mask features $F_{\text{mask}}$, which is concatenated with the multiscale decoding features $\{F_{\text{decode}}^j\}_{j=1}^n$, forming the pyramid decoding features $\{F_{\text{decode}}^j\}_{j=0}^n$.

The resulting pyramid decoding features are processed through convolutional modules, involving convolution, concatenation, and upsampling, with the Relu activation. Ultimately the output mask $I_{\text{mask}}$ is generated. The process is defined as follows:

$$I_{\text{decoder}}^j = \text{Conv}_{\text{blk}}(\text{Concat}(F_{\text{decode}}^j, I_{\text{decoder}}^{j+1})), 0 \le j \le n \quad (8)$$

$$I_{\text{mask}} = I_{\text{decoder}}^0. \quad (9)$$

As for no target samples, to determine whether the input image contains a target corresponding to the referring text, the pyramid decoding features $\{F^j_{\text{decode}}\}^n_{j=0}$ are dimensionally adjusted and passed through an MLP to compute the target indicator. The target indicator is a binary classification result that assesses the presence or absence of a target corresponding to the referring text.

The model's loss function comprises two components: the mask segmentation loss and the target presence loss. These losses are calculated based on the model's two outputs, $I_{\text{mask}}$ and the target indicator. The final model loss is a weighted sum of these two losses. It can be defined as

$$L = \alpha L_{\text{ce\_mask}} + \beta L_{\text{ce\_target}}. \tag{10}$$

The cross-entropy loss is chosen to compute the mask part and the target part of (10).

## V. Experiments and Discussion

### A. Experiment Setup

During training, the parameters of the M2SegNet backbone network are frozen and the batch size is set to 4. The optimizer uses Adam. The learning rate was initialized to 1e-5 and gradually decayed to 1e-7. All experiments are implemented by PyTorch and conducted on the Nvidia RTX 3090 GPUs. In the MPG module, the number of learnable grouping queries is set to $[2, 5, 9]$, and the iterations of MHCA $t$ are set to 3. The loss function weights $\alpha$ and $\beta$ are set to 1 and 0.1, respectively. For computational efficiency, the model was tested in a single GPU environment.

Its deployment on edge devices was explored. The model was evaluated in an Intel Xeon Gold 6148 CPU environment. In addition, it was assessed on the NVIDIA Jetson Xavier NX edge device.

### B. Evaluation Metrics

In the experiments, the cumulative Intersection over Union (cIoU) metric, commonly applied in RES tasks, was used to evaluate the model's performance [11]. However, cIoU is less accurate for small defects. Therefore, the generalized IoU (gIoU) metric was also employed, which averages IoU across all images and optimizes for no-target samples [11]. For no-target samples, the IoU is set to 1 if the target indicator is correctly predicted and 0 otherwise. Precision at different thresholds (Pr@X) measures the proportion of test samples meeting the IoU threshold, with values ranging from 0.1 to 0.9. The mean Pr@X (mPr) is adopted as an evaluation metric. Furthermore, recall, precision, and F1 Score are employed to evaluate segmentation performance. The GFLOPs, FPS, and Params are also employed to evaluate model efficiency.

### C. Comparison Results on RSCDS Task

To demonstrate the effectiveness of our approach, the proposed M2SegNet model is compared with six SOTA methods for RES in natural scenes: ReSTR [23], LAVT [24], PolyFormer [25], ReLA [11], CARIS [11], and ReMamber [27], as shown in Table III. For a fair comparison, all methods were fine-tuned on the RSCDS dataset. The experimental results demonstrate that the M2SegNet model outperforms other models across most metrics.

*1) Performance Analysis:* The proposed model achieves high cIoU and gIoU scores, benefiting from the MMF module, which enhances defect detail capture through effective visual–text integration. The gIoU metric further highlights its advantage in segmenting small defects and handling multigranularity samples. In addition, the MPG module aids in progressively localizing specified defects, improving recognition of small and no-target samples. While M2SegNet's recall metric is slightly lower than the ReMamber model, it still outperforms all other models, but M2SegNet has a clear advantage in Precision, which reflects segmentation accuracy and completeness, largely due to the MMF module's superior fusion performance. Furthermore, M2SegNet surpasses ReMamber and others in the F1 score, balancing both precision and recall. In addition, M2SegNet leads in the mPr metric, confirming its superior overall performance on RSCDS tasks.

*2) Computational Efficiency and Deployment Feasibility:* As shown in Table III, the M2SegNet has a computational cost of 246.31 GFLOPs and achieves an inference FPS of 10.96 on a single GPU. M2SegNet achieves superior segmentation metrics compared to existing methods while maintaining comparable computational efficiency.

Furthermore, the model's feasibility has been verified for deployment on both CPU and edge hardware platforms. In a CPU environment, M2SegNet demonstrated comparable efficiency to existing models while achieving superior performance. The ReMamber encountered failures in CPU environments because of its Vmamba architecture. On the NVIDIA Jetson Xavier NX platform, empirical evaluations indicated that M2SegNet achieved an inference FPS of 1.33, comparable to the 1.42 FPS of the ReLA model. With efficiency comparable to ReLA, M2SegNet achieves superior performance. Notably, low computational efficiency remains a common challenge in current methods. Although M2SegNet can be deployed on edge devices, achieving real-time performance still requires further optimization. A lightweight architecture can be a promising direction for future research to balance accuracy and inference speed on edge devices.

### D. Ablation Study

*1) Different Backbones:* As shown in Table IV, this section conducted an in-depth analysis of the M2SegNet framework originally using Bert and Swin-T as text and image encoders through three distinct encoder replacement strategies as follows.

1) Replace both text and image encoders with pretrained multimodal encoders of ResNet-based CLIP and ViT-based CLIP, i.e., CLIP (RN50) and CLIP (ViT16), since CLIP is a SOTA multimodal model renowned for its strong alignment between textual and visual features [28].
2) Only substitute the original Swin-T with alternative image encoders, such as ResNet and ViT.
3) Only substitute the original Bert with other text encoders, such as T5 [29] and Roberta [30].

TABLE III
COMPARISON OF SEGMENTATION METRICS (%) ON RSCDS TASK BETWEEN SOTA MODELS AND M2SEGNET IN NATURAL SCENES

| Method | Publication year | gIoU | cIoU | Recall | Precision | F1_Score | mPr | GFLOPs | FPS | Params | $FPS_{CPU}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ReSTR [23] | 2022 | 8.73 | 42.36 | 11.1 | 18.2 | 13.79 | 9.32 | 248.56 | 21.86 | 129.7M | 0.74 |
| LAVT [24] | 2022 | 39.88 | 65.38 | 50.48 | 56.27 | 52.99 | 43.63 | 160.11 | 18.85 | 225.6M | 1.11 |
| PolyFormer [25] | 2023 | 11.99 | 29.58 | 17.35 | 18.26 | 17.79 | 8.83 | 175.34 | 15.32 | 309.7M | 1.34 |
| ReLA [11] | 2023 | 51.37 | 71.45 | 67.06 | 71.31 | 71.26 | 55.53 | 225.05 | 12.04 | 223.9M | 0.69 |
| CARIS [26] | 2023 | 55.22 | 72.39 | 62.22 | 67.77 | 64.89 | 54.41 | 249.63 | 11.97 | 226.5M | 0.66 |
| ReMamber [27] | 2024 | 57.15 | 74.87 | 69.13 | 67.9 | 66.69 | 57.8 | 501.31 | 10.04 | 254.5M | - |
| M2SegNet | - | 59.21 | 75.57 | 67.51 | 72.81 | 69.46 | 58.55 | 246.31 | 10.96 | 260.0M | 0.61 |

TABLE IV
RESULTS OF DIFFERENT BACKBONES

| Method | gIoU | cIoU | Recall | Precision | F1_Score | mPr |
|---|---|---|---|---|---|---|
| M2SegNet | 59.21 | 75.57 | 67.51 | 72.81 | 69.46 | 58.55 |
| CLIP (RN50) | 45.12 | 67.43 | 56.13 | 61.01 | 58.47 | 46.45 |
| CLIP (ViT16) | 50.92 | 72.06 | 62.83 | 64.11 | 63.46 | 52.40 |
| ResNet | 37.83 | 62.46 | 50.59 | 53.87 | 52.46 | 39.46 |
| ViT | 38.26 | 62.67 | 51.44 | 54.27 | 52.82 | 39.52 |
| T5 | 54.51 | 75.39 | 67.41 | 65.27 | 66.32 | 56.09 |
| Roberta | 57.51 | 77.23 | 70.08 | 67.76 | 68.90 | 59.15 |
| Bert(large) | 60.26 | 75.27 | 69.87 | 71.61 | 70.45 | 59.59 |

TABLE V
RESULTS OF SEGMENTATION WITH VARYING TEXT COMPLEXITY

| Dataset | gIoU | cIoU | Recall | Precision | F1_Score | mPr |
|---|---|---|---|---|---|---|
| Template | 59.21 | 75.57 | 67.51 | 72.81 | 69.46 | 58.55 |
| LLM | 57.53 | 81.28 | 71.54 | 67.68 | 67.84 | 59.31 |
| Template* | 60.33 | 76.03 | 69.38 | 73.63 | 71.08 | 59.37 |
| LLM* | 62.57 | 76.41 | 73.39 | 75.90 | 74.62 | 63.57 |

"*" indicates that the datasets constructed with extended attributes.

TABLE VI
RESULTS OF MMF AND MPG ABLATION EXPERIMENTS

| Method | gIoU | cIoU | Recall | Precision | F1_Score | mPr |
|---|---|---|---|---|---|---|
| Baseline | 54.06 | 71.86 | 64.58 | 66.80 | 64.79 | 53.37 |
| +MMF | 57.86 | 74.52 | 67.49 | 70.09 | 68.39 | 57.09 |
| +MPG | 58.74 | 74.85 | 67.32 | 69.79 | 68.33 | 58.07 |
| +MMF+MPG | 59.21 | 75.57 | 67.51 | 72.81 | 69.46 | 58.55 |

Experimental results show that while the CLIP-based model is competitive, it underperforms proposed M2SegNet across all evaluation metrics. The poor performance of CLIP is likely due to its lack of multiscale features, which conflicts with the model's multiscale architectural components, causing performance degradation. While pretrained multimodal models typically exhibit strong performance, misalignment in design would significantly compromise their effectiveness.

Only replacing the image encoder with ResNet or ViT led to performance degradation. By comparison, replacing both text and image encoders with pretrained multimodal encoders of ResNet-based CLIP and ViT-based CLIP can achieve higher performances. This observation demonstrates that the pretrained CLIP model improves modality alignment, thereby enhancing segmentation performance.

Specifically, only replacing the text encoder with T5 resulted in a minor performance decline. Larger scale text encoders, such as Bert(large) and Roberta, have demonstrated enhancements in specific metrics. This improvement implies their latent capacity for optimizing overall performance.

Notably, the model exhibits greater sensitivity to image encoder than text encoder alterations, with weaker image encoders drastically reducing performance. In addition, advanced text encoders help elevate the model's performance ceiling.

*2) Text Complexity:* This section explores the impact of textual complexity on segmentation performance. Given the inherent semantic singularity limitation of template-based text construction, textual complexity is enhanced along two dimensions: 1) incorporating extended multidimensional defect attributes besides the original four core attributes, including geometric topology (angular relationships between defects and busbars) and the position relative to the edge (defect located at the edge or center regions) and 2) using the large language model (LLM) to generate sophisticated referring texts based on diverse defect attributes, expanding the linguistic diversity.

Table V compares the template-based and LLM-based methods using both core and extended attributes. Comparing the datasets constructed with core attributes, the LLM-based method outperforms the template-based method on some metrics, i.e., cIoU, recall, and mPr. When applied to datasets incorporating extended attributes, which are marked with "*," the LLM-based approach achieves superior performance across all metrics. These results indicate that complex and diverse texts fully exploit the potential of the text encoder, with enriched linguistic diversity contributing to improved performance. During training, a broader range of rich and varied referring texts significantly enhances the text encoder's ability to process diverse inputs. Notably, despite its advantage in linguistic diversity, the LLM-based approach requires significantly more computational resources to generate texts compared to the template-based method.

*3) MMF and MPG:* The proposed model primarily consists of an image encoder, a text encoder, an MMF module, a multi-granularity perceptual grouping (MPG) module, and a TGD. This section evaluates a baseline model consisting of the image encoder, text encoder, and TGD, while further validating the contributions of the MMF and MPG modules to the final defect detection performance. The experimental results are shown in Table VI, with the image and text encoders implemented using the same network structure as LAVT.

As shown in Table VI, the model's segmentation performance improves significantly when either the MMF module or the MPG module is introduced individually, demonstrating the effectiveness of both modules. When both MMF and MPG are introduced

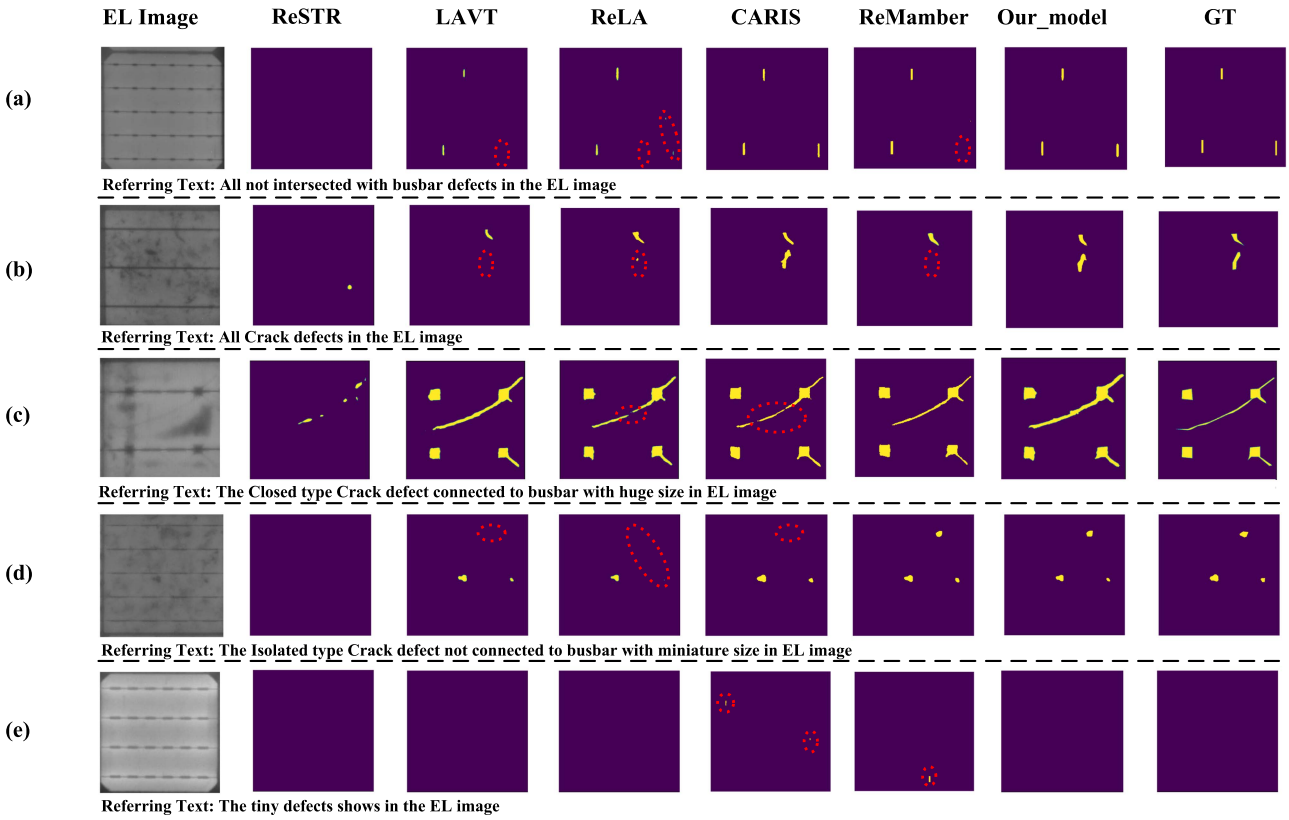


Fig. 6. Visualization of the segmentation effect of each model on RSCDS task.

TABLE VII
RESULTS OF QUANTITATIVE ABLATION EXPERIMENTS IN MPG WITH LEARNABLE CATEGORY QUERIES

| Categories | gIoU | cIoU | Recall | Precision | F1_Score | mPr |
|---|---|---|---|---|---|---|
| (9, 5, 2) | 56.20 | 74.93 | 63.49 | 70.42 | 66.19 | 55.26 |
| (2, 2, 2) | 56.63 | 74.56 | 63.98 | 70.51 | 66.63 | 55.71 |
| (5, 5, 5) | 58.29 | 75.34 | 66.82 | 70.92 | 68.48 | 57.60 |
| (9, 9, 9) | 58.78 | 76.44 | 66.91 | 71.31 | 69.03 | 58.03 |
| (2, 5, 9) | 59.21 | 75.57 | 67.51 | 72.81 | 69.46 | 58.55 |

simultaneously, all metrics reach the highest values. This result confirms that the proposed MMF and MPG modules work synergistically to enhance the model's segmentation accuracy and robustness.

*4) Number of Learnable Grouping Queries in MPG:* In the MPG module, inspired by positioning characteristics of human vision, the number of learnable grouping queries is designed to go from fewer to more, progressing from coarse to fine. For deep features that capture global information, only two grouping classes are used to distinguish foreground and background. As features become more detailed, mid-level features are divided into five classes, adding three main defect categories. At the shallowest level, the number of classes increases to nine, covering seven specific defect types. Thus, the final category settings across three feature scales are (2, 5, 9).

To investigate the impact of this parameter on performance, ablation experiments were performed by varying the number of grouping queries in the MPG module. Tests were carried out on configurations with a fine-to-coarse manner, such as (9, 5, 2), as well as configurations with uniform category numbers, such as (2, 2, 2), (5, 5, 5), and (9, 9, 9). As demonstrated in Table VII, the (2, 5, 9) configuration yielded the highest performance across all metrics.

## E. Visualization and Analysis

In this study, the segmentation results on the proposed dataset were visualized to enable a more intuitive comparison across different models, including the proposed M2SegNet, ReSTR [23], LAVT [24], PolyFormer [25], ReLA [11], CARIS [11], and ReMamber [27], as illustrated in Fig. 6. In Fig. 6, (a) and (b) depict coarse-grained targets, (c) and (d) depict fine-grained targets, and (e) corresponds to no-target samples. It can be observed that ReSTR struggles to effectively detect defects. LAVT performs poorly on coarse-grained targets and exhibits missed detections, particularly in Fig. 6(a), (b), and (d). CARIS and ReMamber perform well in detecting defects under fine-grained semantics and can effectively detect corresponding defects based on the referring text. However, their segmentation performance on no-target and coarse-grained targets is suboptimal, such as missed detections in Fig. 6(a), (b), and (d), and redundant outputs in Fig. 6(e). Although ReLA is optimized for no-target samples, it still fails to segment all targets effectively for coarse-grained scenarios [as seen in Fig. 6(a), (b), and (d)]. From the visualization results, it is evident that the M2SegNet model achieves superior defect segmentation performance across coarse-grained,

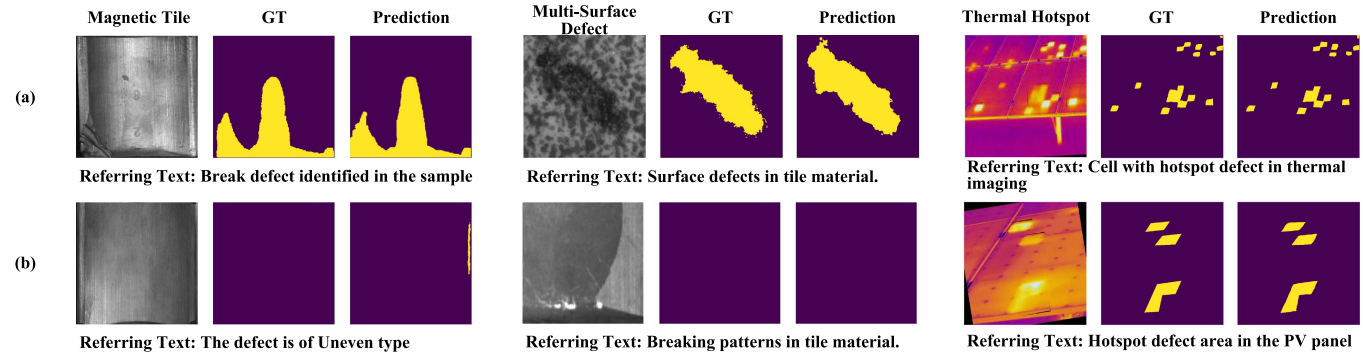


Fig. 7. Visualization of the segmentation effect of defects in other domains and imaging modalities.

TABLE VIII
QUANTITATIVE RESULTS ON OTHER DOMAINS

| Dataset | gIoU | cIoU | Recall | Precision | F1_Score | mPr |
|---|---|---|---|---|---|---|
| MT defect | 79.28 | 84.36 | 83.33 | 86.28 | 85.92 | 80.92 |
| MS defect | 80.11 | 78.73 | 72.75 | 86.41 | 86.19 | 81.44 |
| TI defect | 88.13 | 89.33 | 93.09 | 93.23 | 93.18 | 91.99 |

fine-grained, and no-target scenarios, delivering more accurate defect segmentation.

### F. Generalization to Defects on Other Domains

To evaluate the generalization capability of M2SegNet, this section conducts experiments on industrial surface defect and thermography-based PV hot spot segmentation tasks. Based on the magnetic tile defect (MT defect, 457 images) [31], multisurface defect (MS defect, 2672 images) [32], and PV thermography imaging defect (TI defect, 2063 images) datasets, referring texts were generated by the multimodal LLM to construct three multimodal datasets. The training and test sets were split at an 8:2 ratio.

As shown in Table VIII, M2SegNet consistently achieves high performance across diverse datasets. These results highlight M2SegNet's reliable generalization capabilities across various defect types and imaging modalities, ensuring robust segmentation in real-world scenarios with diverse data.

Fig. 7 further provides visualization of segmentation outcomes. For the MT dataset, the model accurately segments defects when the referring text aligns with image content and maintains stability on defect-free samples. However, when misalignment occurs [e.g., sample (b) in the "MT" column where the text describes a nonexistent "uneven" defect], the model predicts the actual "break" defect instead of generating blank masks. This phenomenon likely stems from the limited size of the dataset, which restricts the transformer-based M2SegNet's ability to effectively align multimodal information. On the larger MS defect dataset, the issue of misalignment observed in the MT dataset did not occur, further supporting the hypothesis that dataset size plays a critical role in model performance.

## VI. CONCLUSION

In response to the personalized maintenance needs of current PV power plants, the RSCDS task is proposed and a multimodal, multigranularity dataset Ref-EL-Defect is constructed for this task. A novel referring defect segmentation model M2SegNet is also designed for this task. Through the carefully crafted MMF and MPG modules, M2SegNet effectively processes multimodal and multigranularity information, significantly improving the accuracy and efficiency of RSCDS. Compared to current RES models in natural scenes, our model achieves superior performance in the RSCDS task. The proposed RSCDS task has the potential to greatly expand the application of deep learning in the maintenance processes of PV power plants, effectively enabling the personalized segmentation of defects in solar cells.

Future research could focus on developing lightweight architectures to improve efficiency on edge devices. Furthermore, employing advanced fusion strategies to integrate multiple imaging modalities, such as EL and thermal imaging, could further enhance defect segmentation performance.

## ACKNOWLEDGMENT

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

## REFERENCES

[1] M. Dhimish and P. Mather, "Ultrafast high-resolution solar cell cracks detection process," *IEEE Trans. Ind. Inf.*, vol. 16, no. 7, pp. 4769–4777, Jul. 2020.
[2] U. Otamendi, I. Martinez, I. G. Olaizola, and M. Quartulli, "A scalable framework for annotating photovoltaic cell defects in electroluminescence images," *IEEE Trans. Ind. Inf.*, vol. 19, no. 9, pp. 9361–9369, Sep. 2023.
[3] J. Zhang, X. Chen, H. Wei, and K. Zhang, "A lightweight network for photovoltaic cell defect detection in electroluminescence images based on neural architecture search and knowledge distillation," *Appl. Energ.*, vol. 355, 2024, Art. no. 122184.
[4] 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.
[5] C. Wang, H. Chen, and S. Zhao, "RERN: Rich edge features refinement detection network for polycrystalline solar cell defect segmentation," *IEEE Trans. Ind. Inf.*, vol. 20, no. 2, pp. 1408–1419, Feb. 2024.

[6] J. Zhang et al., "Automatic detection of defective solar cells in electroluminescence images via global similarity and concatenated saliency guided network," *IEEE Trans. Ind. Inf.*, vol. 19, no. 6, pp. 7335–7345, Jun. 2023.
[7] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Comput. Vis.–ECCV: 14th Eur. Conf.*, Springer Int. Publishing, 2016, pp. 108–124.
[8] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 787–798.
[9] L. Yu et al., "Modeling context in referring expressions," in *Proc. Comput. Vis.–ECCV 14th Eur. Conf.*, Springer International Publishing, 2016, pp. 69–85.
[10] X. Zhang et al., "Referring camouflaged object detection," 2023. [Online]. Available: https://arxiv.org/abs/2306.07532
[11] C. Liu, H. Ding, and X. Jiang, "GRES: Generalized referring expression segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23592–23601.
[12] Y. Liao et al., "A real-time cross-modality correlation filtering method for referring expression comprehension," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10877–10886.
[13] D. Liu, H. Zhang, Z. J. Zha, and F. Wu, "Learning to assemble neural module tree networks for visual grounding," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4672–4681.
[14] U. Hijjawi et al., "A review of automated solar photovoltaic defect detection systems: Approaches, challenges, and future orientations," *Sol. Energy*, vol. 266, 2023, Art. no. 112186. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X23008204
[15] D. Mery and C. Arteta, "Automatic defect recognition in X-ray testing using computer vision," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 1026–1035.
[16] S. Mukherjee and S. T. Acton, "Oriented filters for vessel contrast enhancement with local directional evidence," in *Proc. IEEE 12th Int. Symp. Biomed. Imag.*, 2015, pp. 503–506.
[17] Y. Wang et al., "A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network," *ISA Trans.*, vol. 96, pp. 457–467, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019057819302903
[18] B. Du, Y. He, Y. He, J. Duan, and Y. Zhang, "Intelligent classification of silicon photovoltaic cell defects based on eddy current thermography and convolution neural network," *IEEE Trans. Ind. Inform.*, vol. 16, no. 10, pp. 6242–6251, Oct. 2020.
[19] M. Waqar Akram, G. Li, Y. Jin, and X. Chen, "Failures of photovoltaic modules and their detection: A review," *Appl. Energ.*, vol. 313, 2022, Art. no. 118822. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261922002677
[20] R. Yang et al., "Electromagnetic induction heating and image fusion of silicon photovoltaic cell electrothermography and electroluminescence," *IEEE Trans. Ind. Inform.*, vol. 16, no. 7, pp. 4413–4422, Jul. 2020.
[21] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-CNN for small object detection," in *Proc. Comput. Vis.–ACCV* 2017, pp. 214–230.
[22] B. Cheng et al., "Masked-attention mask transformer for universal image segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2112.01527
[23] N. Kim et al., "ReSTR: Convolution-free referring image segmentation using transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18124–18133.
[24] Z. Yang et al., "LAVT: Language-aware vision transformer for referring image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18134–18144.
[25] J. Liu et al., "PolyFormer: Referring image segmentation as sequential polygon generation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18653–18663.
[26] S.-A. Liu et al., "CARIS: Context-aware referring image segmentation," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 779–788.
[27] Y. Yang et al., "Remamber: Referring image segmentation with mamba twister," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 108–126.
[28] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn*. PMLR, Jul. 2021, vol. 139, pp. 8748–8763.
[29] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
[30] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692
[31] Y. Huang et al., "Surface defect saliency of magnetic tile," in *Proc. IEEE 14th Int. Conf. Automat. Sci. Eng.*, 2018, pp. 612–617.
[32] Y. Bao et al., "Triplet-graph reasoning network for few-shot metal generic surface defect segmentation," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5011111.

**Shenghao Dong** received the B.S. degree in automation, in 2024, from the School of Automation, Southeast University, Nanjing, China, where he is currently working toward the Master's degree in electronic and information engineering.

His research interests include multimodal learning, machine learning, and photovoltaics defect inspection.

**Jinxia Zhang** (Member, IEEE) received the B.S. degree in computer science and technology and the Ph.D. degree in control science and engineering from the Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2009 and 2015, respectively.

From 2012 to 2014, she was a Visiting Scholar with Visual Attention Lab, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. She is currently an Associate Professor with the School of Automation, Southeast University, Nanjing, China. Her research interests include saliency detection, knowledge transfer, computer vision, and machine learning.

**Yu Shen** received the B.S. degree in microelectronics from the School of Internet of Things, Jiangnan University, Wuxi, China, the master's degree in control engineering, and the Ph.D. degree in control science and engineering from the School of Automation, Southeast University, Nanjing, China, in 2013 and 2016, respectively.

She is currently the Postdoctoral in control science and engineering with the School of Automation, Southeast University. Her research interests include solar energy, photovoltaics, computer vision, and machine learning.

**Dehong Gao** received the Ph.D. degree in natural language processing from The Hong Kong Polytechnic University, Hong Kong, China, in 2014.

From 2014 to 2022, he was a Senior Algorithm Expert with Alibaba Group, Hangzhou, China. He is currently an Associate Professor with the School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China. His research interests include information retrieval, recommendation, natural language processing, and machine learning.