Renewable Energy 256 (2026) 124505

Contents lists available at ScienceDirect

Renewable Energy

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

Knowledge distillation based on fast normalization attention and auxiliary multi-label recognition for continual photovoltaic defect segmentation

Xinyue Lei [a], Jinxia Zhang [a,b,*], Ying He [a], Shixiong Fang [a], Kanjian Zhang [a], Haikun Wei [a]

[a] Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing, 210096, Jiangsu, China
[b] Advanced Ocean Institute of Southeast University, Nantong, 226010, Jiangsu, China

ARTICLE INFO

Keywords:
Photovoltaic
Defect segmentation
Continual learning
Distillation
Catastrophic forgetting

ABSTRACT

Defect detection in solar cells is essential to ensure the reliability and efficiency of solar energy systems. Existing electroluminescence-based defect detection methods have been widely applied as a non-destructive technique in automatic and intelligent photovoltaic (PV) defect inspection systems. However, existing electroluminescence-based defect detection methods fail to consider the actual dynamic operating conditions (e.g., temperature and humidity) that impact defect evolution during the lifecycle of a PV module. To address the gap, we propose SegAdapt, a novel knowledge distillation framework designed for PV defect segmentation in dynamic conditions. SegAdapt leverages knowledge distillation to effectively transfer historical defect knowledge from previous model to the new one, thereby preserving prior learning while adapting to newly emerging defects. To highlight slender defects during distillation, we propose an attention mechanism based on fast normalization. To further retain old defect segmentation capabilities while learning new types, we perform auxiliary multi-label recognition at all levels of the model encoder during distillation. This strategy ensures the model effectively distinguishes various defects without catastrophic forgetting. Experiments demonstrate that our method is superior to other general methods in the continual learning setting, achieving robust and accurate segmentation of PV defects under dynamic operational scenarios.

1. Introduction

Photovoltaic (PV) power generation is a clean, environmental friendly, and sustainable energy source, which has seen rapid development globally [1]. In 2023, the newly installed photovoltaic capacity increased from 236 GW in 2022 to 446 GW, representing a growth of approximately 89% [2]. As the core component of PV systems, the reliability of PV modules is closely tied to the overall performance and efficiency of the power generation system. However, during assembly and operation, PV modules are inevitably exposed to external forces or environmental stresses, leading to defects such as cracks, corrosion, and interconnect [3]. These defects can shorten the lifespan of PV modules and influence their efficiency, negatively impacting the entire PV system [4]. Therefore, developing robust and accurate defect detection methods is crucial, as it could enhance the long-term reliability and efficiency monitoring of photovoltaic modules.

Existing methods typically detect solar cell defects through visual inspection, current–voltage (I-V) curve analysis [5], infrared thermal (IRT) imaging [6], and electroluminescence (EL) imaging techniques [7]. Visual inspection mainly relies on the subjective judgment of the experts, which is time-consuming and labor-intensive. While the I-V curve analysis is simple in principle, it is not sensitive to tiny defects. IRT suffers from low resolution and poor defect inspection accuracy [6]. In contrast, as a non-destructive technique, EL imaging provides high-resolution images, which help detect defects within solar cells. Therefore, we focus on detecting defects in solar cells through the EL imaging technique.

Although numerous EL-based defect segmentation [8,9] methods have been proposed, they are typically trained on fixed datasets with a predefined set of defect classes, as illustrated in Fig. 1(a). Therefore, they suffer from several limitations in real world applications. Firstly, they are inherently unsuitable for scenarios where new defect classes continuously emerge. In practice, this limitation is critical because PV systems are deployed in geographically diverse and dynamically changing environments, where factors such as variations in temperature, humidity, and other localized operating conditions can induce previously unseen defect types at different sites or at various stages of a system's operational lifespan. Secondly, it is impractical to store the

* Corresponding author.
E-mail addresses: leixinyue@seu.edu.cn (X. Lei), jinxiazhang@seu.edu.cn (J. Zhang), heying719@seu.edu.cn (Y. He), sxfang@seu.edu.cn (S. Fang), kjzhang@seu.edu.cn (K. Zhang), hkwei@seu.edu.cn (H. Wei).

https://doi.org/10.1016/j.renene.2025.124505
Received 31 March 2025; Received in revised form 30 August 2025; Accepted 22 September 2025
Available online 27 September 2025

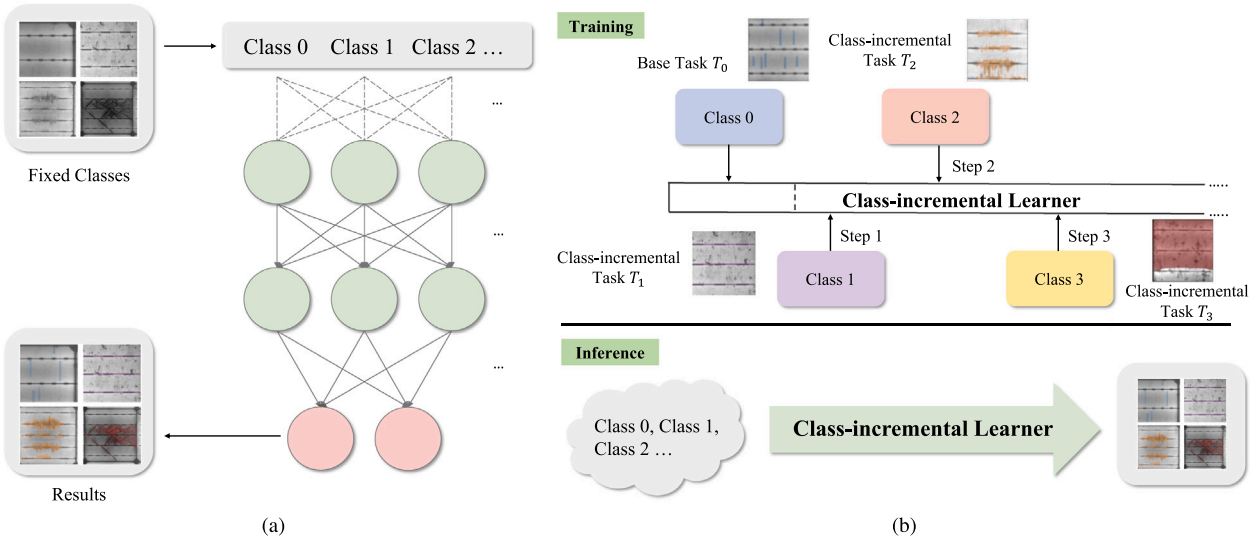

Fig. 1. Traditional one-off training paradigm and continual learning paradigm. (a) Traditional one-off training paradigm. (b) Continual learning paradigm.

entire historical dataset due to its scale and the evolving nature of defect distributions across locations and time. Thirdly, even if data storage were not a constraint, to accommodate these new data would require retraining the model with all the historical and newly collected data, which is both time-consuming and computationally expensive. These limitations significantly undermine the practicality and scalability of existing methods in long-term PV system monitoring. To address the limitations, this paper explores the PV defect segmentation in the context of continual learning[1] (CL) for the first time. Unlike previous methods, our approach could incrementally acquire and accumulate defect knowledge over time, closely resembling human lifelong learning mechanisms, as shown in Fig. 1(b).

The main obstacle in CL is catastrophic forgetting, a phenomenon characterized by significant performance deterioration on previously learned tasks when models adapt sequentially to new tasks. This issue is particularly pronounced when identifying defects in EL solar cell images, where target defects are often indistinct and backgrounds are visually complex. Consequently, defects can be easily mistaken for background structures such as busbars. Additionally, PV defects exhibit substantial variations in spatial scale, further complicating their accurate segmentation. While existing CL methods perform well on general images with distinct subjects and clear backgrounds, they struggle with EL solar cell images due to the intricate textures and subtle defect patterns. Irrelevant knowledge about background rather than crucial defect-related information may be transferred to the new model, thereby exacerbating catastrophic forgetting.

To tackle the issues above, this paper proposes a novel knowledge distillation framework SegAdapt tailored for PV module defect segmentation in dynamic conditions. In our method, the proposed fast-normalization feature fusion mechanism explicitly emphasizes slender or thin defects during the distillation process, effectively guiding the model's attention toward these easily overlooked defect regions. Considering the subtle defect patterns inherent in EL images, we further implement auxiliary multi-label recognition, which comprises two parts: auxiliary multi-label classification using pseudo-labels to transfer defect-type knowledge and auxiliary multi-label logit distillation based on the old model to transfer soft-target knowledge. Through the auxiliary multi-label recognition, our model could retain global awareness of all learned defect categories and learn better representations of each defect class, thus reducing the forgetting of old defect classes.

The main contributions in our paper are as follows:

(1) To the best of our knowledge, we are the first to study PV defect segmentation in the context of continual learning. Considering the prior defects' characteristics, we propose SegAdapt, a novel knowledge distillation framework tailored for PV defect segmentation in dynamic conditions. SegAdapt utilizes accumulated defect knowledge and can effectively handle scale variations and noise sensitivity.
(2) To highlight slender or thin defects, we propose an attention mechanism based on fast-normalization feature fusion. This enhances the attention to slender defects and guides the knowledge distillation of relevant defect information.
(3) To capture defect knowledge across various levels and types, mitigate forgetting of old classes and facilitate adaption for new ones, we implement auxiliary multi-label recognition at all levels of the model encoder. This helps the model learn information about different defects across multiple levels effectively.
(4) Experimental results demonstrate that our SegAdapt is superior to other general CL methods — originally designed for natural scenes — in both new and existing defect segmentation.

2. Related work

2.1. EL-based automatic PV defect detection

Early EL-based research [10,11] primarily employed traditional pattern recognition methods. Most of them detect defects by analyzing the relationships in the frequency and spatial domains. However, these methods rely on manual parameter tuning and often lack robustness.

Since the advent of Convolutional Neural Network (CNN), scholars have increasingly adopted deep learning methods to detect defects. Some researchers focused on distinguishing normal and defective solar cells. Zhang et al. [12] designed a lightweight defect detection network using neural architecture search to optimize network structure. Zhang et al. [7] proposed a network guided by global pairwise similarity and serial saliency for automatic defect detection based on their visual characteristics. However, these methods fail to localize defects precisely.

Other researchers used object detection techniques to localize defect regions. Su et al. [13] proposed a complementary attention network that suppresses background noise and focuses on defect regions. Wang et al. [14] applied CL techniques for solar cell defect detection, enabling the model to adapt and detect new defect types quickly. This work demonstrates that CL aids in detecting newly emerging defects.

[1] Also known as 'incremental learning' or 'lifelong learning'.

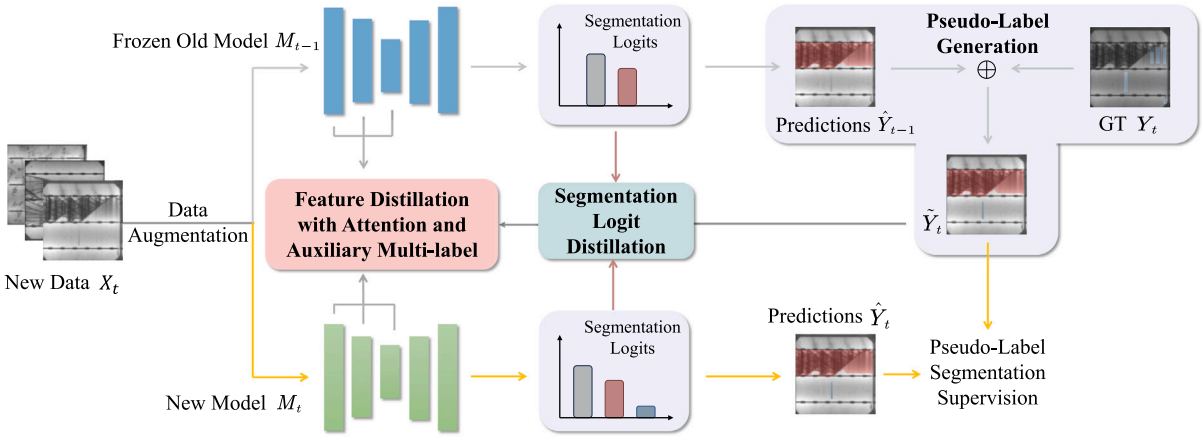

Fig. 2. Pipeline of the proposed SegAdapt. At training step t, the old model M_{t-1} is frozen, while the new model M_t is trainable. The input images X_t are fed into M_{t-1} and M_t. The new model M_t learns old defect types through distillation and pseudo-labels generated by the frozen old model.

Nonetheless, these methods still cannot accurately segment defects in solar cells.

Some studies aim to locate and segment defects in solar cells precisely. Jiang et al. [15] introduced squeeze-and-excitation (SE) attention mechanism [16] in M-net to prevent segmentation failure during inference despite success in training. To enhance the interaction with multi-level information, Zhou et al. [8] proposed a multi-level pyramid strategy and a neighbor attention fusion module to refine the defect segmentation. However, these methods fail to consider the actual dynamic operating conditions that impact defect evolution.

To address this issue, this paper explores PV defect segmentation in the context of continual learning. We propose a novel knowledge distillation framework SegAdapt, tailored for PV module defect segmentation in dynamic conditions.

2.2. Continual semantic segmentation

Class-incremental semantic segmentation has been a cutting-edge hotspot. Two main challenges in this field are catastrophic forgetting [17] and background semantic shift [18]. Catastrophic forgetting refers to the drastic performance degradation of a model on previously learned tasks when facing new tasks. Background semantic shift occurs when old types are classified as background in image masks that include new types, preventing the model from effectively learning the knowledge related to old types.

To alleviate these problems above, some researchers [19,20] proposed to retrospect known knowledge through sample replay. However, replay-based methods generally require exemplar memory, incurring additional costs. A more efficient approach is knowledge distillation, which transfers knowledge from the old model to the new one to alleviate forgetting. PLOP [21] retained long- and short-range spatial relationships between features through a multi-scale pooling feature distillation technique. ALIFE [22] addressed catastrophic forgetting through adaptive output distillation combined with feature replay. IDEC [23] employed feature knowledge distillation to prevent forgetting and apply region-based contrastive learning to tackle background semantic shift. LSKD [24] proposed output distillation based on class similarity, forcing the new model to mimic the old model's outputs for knowledge transfer.

Although these methods perform well on images with distinct subjects and clear backgrounds, they struggle with EL solar cell images, where textures are often blurry and defects can easily be confused with background noise. Thus, their performance in defect segmentation requires further improvement.

3. Methodology

To the best of our knowledge, this is the first to study PV defect segmentation in the context of continual learning. Fully considering the large spatial scale variation of defects and susceptibility to noise, we propose SegAdapt, allowing continual segmentation of defects.

As shown in Fig. 1(b), the class-incremental continual segmentation process consists of a base task T_0 and class-incremental tasks $T_t\,(t = 1, 2, \ldots, P)$. In T_0, the base model M_0 is trained in a traditional one-off paradigm on the static base data $D_0 = \{X_0, Y_0\}$ to study the base class C_0. $X_0 \in \mathbb{R}^{C\times H\times W}$ denotes the input images and $Y_0 \in \mathbb{R}^{H\times W}$ denotes the corresponding ground truth. In the subsequent steps t, the model M_{t-1} updates by data D_t, which introduces new classes C_t. And the ground truth Y_t only covers C_t.

3.1. Overall framework

In the class-incremental task T_t, the model follows the framework illustrated in Fig. 2 to eliminate the issue of forgetting caused by the unavailability of historical data. To explicitly manage incremental knowledge updating, we construct a "teacher-student" architecture, where historical defect knowledge is distilled from a frozen teacher model M_{t-1} into the student model M_t. Specifically, this knowledge transfer is performed through attention-based feature distillation with auxiliary multi-label recognition. Additionally, adaptive segmentation logit distillation is utilized to further mitigate forgetting. During training, the input images X_t are fed to both M_{t-1} and M_t to generate multi-level feature maps F_{t-1} and F_t. Through the layer-wise feature distillation, low-level features that capture defect edges and shapes, along with high-level semantic features are transferred from the old model M_{t-1} to the new model M_t, ensuring continual defect segmentation accuracy across incremental tasks.

Since ground truth Y_t only covers new defect category C_t. If Y_t is the only output supervision, the model will classify the defect C_t as foreground while incorrectly recognizing the old defects $C_{0:t-1}$ as background, exacerbating catastrophic forgetting. Therefore, we utilize the prediction results $\hat{Y}_{t-1}$ from M_{t-1} combined with ground truth Y_t to generate pseudo-labels $\tilde{Y}_t$. This $\tilde{Y}_t$ is then used for auxiliary multi-label recognition and pseudo-label segmentation supervision at step t. Thus, our model can learn new classes through multi-category knowledge from distillation while reviewing the learned old classes via pseudo-labels.

Moreover, due to the imbalance defect distribution, we augment the minority type data before training to ensure successful training and assign higher weights to them during training. This helps the model learn representative features, improving generalization and robustness.

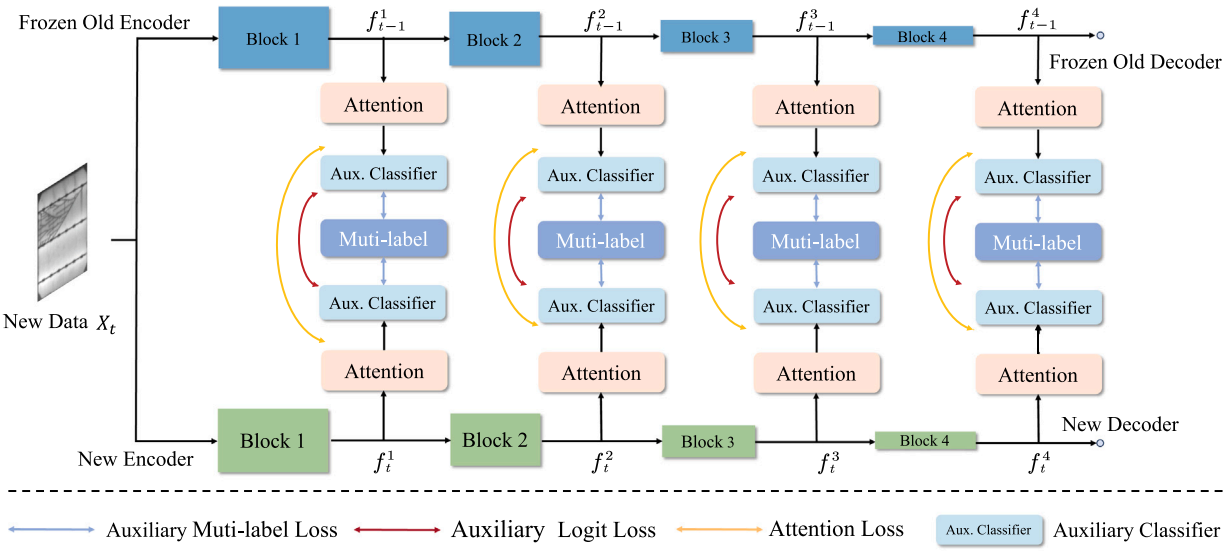

Fig. 3. Feature knowledge distillation with attention mechanism and auxiliary multi-label recognition. The old model and new model are attached with auxiliary classifiers at different levels.

The total loss function $\mathcal{L}_{\text{total}}$ during training consists of three components: the pseudo-label segmentation supervision $\mathcal{L}_{\text{CE}}$ [25], the adaptive segmentation logit distillation loss $\mathcal{L}_o$ [22], the feature distillation loss $\mathcal{L}_{\text{FD}}$:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{FD} + \mathcal{L}_o. \tag{1}$$

$\mathcal{L}_{\text{FD}}$ will be detailed in Section 3.2. Particularly, in the base learning task, the loss function only comprises $\mathcal{L}_{\text{CE}}$ to measure the loss between the ground truth and the predictions.

The remainder of this chapter will provide a detailed explanation of SegAdapt discussed in this section. Section 3.2 explains the proposed knowledge distillation based on fast normalization attention and auxiliary multi-label recognition, while Section 3.3 describes the generation and fusion process of pseudo-labels.

3.2. Feature distillation with attention and auxiliary multi-label recognition

Defects in solar cells exhibit large spatial scale variations and are easily confused with background noise, which increases the likelihood of the model confusing new and old defect types and even forgetting old defects during CL. To address catastrophic forgetting, we propose a multi-level feature distillation strategy based on attention and auxiliary multi-label recognition. This strategy fully utilizes the prior knowledge in the old model, performing feature alignment at all layers of the network. This helps the new model learn both the shape and semantic information of old defects from the old model. As shown in Fig. 3, to obtain multi-level features F_t and F_{t-1}, the training image X_t is input into both the new and old models. Only the new model is trainable, while the old model is frozen to prevent the backpropagation of gradients through the old model.

To highlight slender or thin defects which often suffer from insufficient representation in defect detection, we propose an attention mechanism based on fast-normalization feature fusion. For feature maps x at different levels, we explicitly model the inter-dependencies between pixels within an image patch R and obtain the attention map. The attention map then guides information transmission as part of the attention distillation process, which ensures slender defects that are easily overlooked could be robustly captured and maintained across learning steps. As shown in Eq. (2), our attention mechanism employs a computationally efficient weighted-average fusion operation within each image patch. This design not only reduces computational complexity but also maintains smooth differentiability, allowing stable gradient backpropagation and substantially improving the stability and effectiveness of the feature distillation process.

$$A_r = \sum_{i \in R_r} \frac{x_i}{\epsilon + \sum_{j \in R_r} x_j} \cdot x_i \tag{2}$$

x denotes the feature at each block of the feature extraction network, and i or j denotes the index of each point within the image patch R. x_i or x_j represents the corresponding feature value. A_r is the attention value obtained for the rth image patch. The size of the image patch R is $k \times k$, where $k = 2$. Additionally, to avoid numerical instability, ϵ is set to 0.001.

Next, inspired by feature distillation [26], we accomplish the attention distillation based on our proposed normalization-based attention and perform distillation via the cosine similarity:

$$\mathcal{L}_{attention} = \frac{1}{N} \sum_{l=1}^{N} L_{cos}\left(A\left(F^l{}_{t-1}\right), A\left(F^l{}_t\right)\right), \tag{3}$$

where N denotes the number of feature levels, which in this paper is 4. $A(\cdot)$ refers to the attention mechanism defined in Eq. (2), and l indicates which level the feature map is from. Unlike the pure feature distillation, we distill the attention maps which can transfer the highlighted defect regions.

Unlike natural scenes, PV defects exhibit much larger variations in spatial scale. Solely relying on attention distillation may cause the new model to focus excessively on larger defect types, neglecting the semantic information of smaller defects, thereby exacerbating catastrophic forgetting during CL. To learn defect-type knowledge and soft-target knowledge across different levels, we further implement auxiliary multi-label recognition, which includes the auxiliary multi-label classification using pseudo-labels and the auxiliary multi-label logit distillation based on the old model. Among them, the auxiliary multi-label classification is defined as the BCE (Binary Cross-entropy) [27]:

$$\mathcal{L}_{multi-label} = \frac{1}{N} \sum_{l=1}^{N} L_{BCE}\left(c_l\left(A\left(F_t^l\right)\right), T(\tilde{Y}_t)\right). \tag{4}$$

Here, $c_l(\cdot)$ is the auxiliary classifier attached at each level of the feature extraction network, and the function T transforms the pixel-level pseudo-label $\tilde{Y}_t$ to image-level labels. The weighted BCE loss calculates weights based on the pixel proportion of each defect category, i.e., assigning higher weights to tail classes with fewer pixels to ensure sufficient focus during training. Notably, $\tilde{Y}_t$ contains previously learned

old types and newly emerging types. This forms the basis of defect-type knowledge, which captures the presence of different defect types in each image. Through the auxiliary multi-label classification using $\tilde{Y}_t$, the model could incorporate defect-type knowledge that reinforces its memory of features from old defect types but also enhances its understanding of new defect features. This facilitates feature alignment across multiple network levels, ultimately leading to improved overall performance in defect segmentation.

Additionally, we perform auxiliary multi-label logit distillation based on the old model across multiple levels using the KL (Kullback–Leibler) divergence [26]:

$$\mathcal{L}_{logit} = \frac{1}{N}\sum_{l=1}^{N} L_{KL}\left(c_l\left(A\left(F_{t-1}^{l}\right)\right), c_l\left(A\left(F_{t}^{l}\right)\right)\right). \tag{5}$$

If the probability distributions of the two logits are very similar, their KL divergence will be small, indicating that the new model's feature representation at that level is highly similar to the old model's. Conversely, if the KL divergence is large, it suggests a significant difference in features between the new and old models. Through the auxiliary multi-label logit distillation, the model incorporates the soft-target knowledge which captures the soft decision boundaries of the old model.

In summary, the feature distillation loss $\mathcal{L}_{\mathrm{FD}}$ in Section 3.1 consists of attention loss, auxiliary multi-label logit loss, and auxiliary multi-label classification loss:

$$\mathcal{L}_{FD} = \lambda_a \mathcal{L}_{attention} + \lambda_l \mathcal{L}_{logit} + \lambda_m \mathcal{L}_{multi-label}, \tag{6}$$

where λ_a, λ_l, and λ_m are the hyper-parameters respectively.

3.3. Pseudo-labels

Since the ground truth Y_t in the training set D_t only marks the defect C_t being learned, previously learned old types $C_{0:t-1}$ are treated as background. This means that during CL, the model's understanding of the background dynamically changes with each task, leading to background semantic shift, which exacerbates the problem of catastrophic forgetting. To address this issue, we employ a pseudo-label strategy adopted by [21] to leverage the prediction results $\hat{Y}_{t-1}$ from the model at step $t-1$ and retain knowledge of previously learned types:

$$\tilde{Y}_t(i,j) = \begin{cases} Y_t(i,j), & \text{if } Y_t(i,j) \neq 0, \\ \hat{Y}_{t-1}(i,j), & \text{if } Y_t(i,j) = 0 \text{ and } \hat{Y}_{t-1}(i,j) \neq 0 \\ & \text{and } p > \tau, \\ 0, & \text{otherwise}. \end{cases} \tag{7}$$

During incremental learning ($t > 2$), if a pixel (i, j) belongs to the defect in the current ground truth Y_t, the true label is used directly. If the pixel (i, j) is background in the current ground truth Y_t but some defect in $\hat{Y}_{t-1}$, and the probability $p(i, j)$ of belonging to some old defect class $C_{0:t-1}$ exceeds the threshold τ, the prediction result is adopted. All remaining pixels are treated as background. The right half of Fig. 2 also illustrates the pseudo-label fusion process, where old types predicted by the old model are marked in red, and new types are marked in blue.

4. Results

4.1. Dataset

The dataset used is the public solar cell dataset UCF EL [3]. It comprises both monocrystalline and polycrystalline silicon solar cell images. The total image number is 11,851. There are four main types of defects, as illustrated in Fig. 4. Our training and testing set divisions are consistent with the original work.

To train smoothly, we perform data augmentation on the original 258 images of these two rare classes before training. The data augmentation includes random rotation from +25° to −25°, random scaling from 0.5 to 2.0, random flipping, and random cropping with a scale factor of 0.8 applied with a probability of 0.8. As a result, the number of images of these two classes is expanded to approximately 4000, which helps provide sufficient training samples for the CL process. It is worth noting that the purpose of performing image augmentation before training is to enable the successful training of the continual PV defect segmentation task, rather than eliminating the class imbalance problem in the dataset. After image augmentation, the class imbalance problem still persists, as illustrated in Fig. 5.

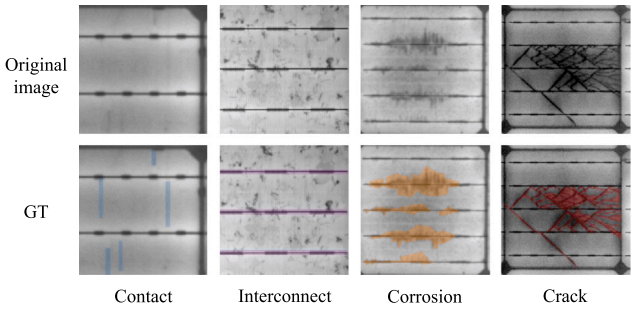

Fig. 4. Examples of each defect category are shown, with the corresponding ground truth (GT) masks overlaid on the original images. The GT annotations indicate the true defect regions, where contact is marked in blue, interconnect in purple, corrosion in yellow, and crack in red. Among them, contact and interconnect appear particularly thin and subtle, compared with the busbar and noise in the background.

Fig. 5. The distribution of each class in the training set before and after data augmentation.

To simulate real-world CL scenarios, certain defects are treated as base classes C_0 in the base task T_0, and the remaining types are sequentially introduced as new classes C_t in incremental tasks T_t. Accordingly, we design three scenarios that reflect different class-incremental dynamics in Table 1. "3-1" means contact, interconnect, and corrosion are the base classes C_0, with crack as the incremental class C_1, including two steps. Similarly, "2-2" means contact and interconnect are the base classes C_0, followed by corrosion and then crack, including three steps.

4.2. Evaluation metrics

The quantitative evaluation metric employed in this paper is the standard mean Intersection over Union (mIoU), a metric widely adopted in segmentation tasks [28]:

$$mIoU = \frac{1}{C_K}\sum_{i=1}^{C_K} \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{8}$$

Table 1
Settings for different CL scenarios. Bg, Ct, It, Co, Ck indicate background, contact, interconnect, corrosion, crack, respectively.

CL scenario	C_0	C_1	C_2	C_3
3-1	Bg, Ct, It, Co	Ck	–	–
2-2	Bg, Ct, It	Co	Ck	–
1-1	Bg, Ct	It	Co	Ck

Table 2
Quantitative experimental results (%), where the best performance is in **bold** and the second best performance is underlined. $Fine-tune$ and $Joint$ are the lower bound and upper bound respectively.

Method	3-1 (2 steps)			2-2 (3 steps)			1-1 (4 steps)		
	$mIoU_{old}$	$mIoU_{new}$	mIoU	$mIoU_{old}$	$mIoU_{new}$	mIoU	$mIoU_{old}$	$mIoU_{new}$	mIoU
$Fine-tune$	3.30	33.77	9.39	0.18	12.91	5.27	0.02	6.67	4.01
LwF [29]	30.74	38.80	32.35	27.41	19.28	24.16	38.68	11.90	22.61
PLOP [21]	34.58	57.00	39.06	27.62	22.23	25.46	48.57	12.33	26.83
ALIFE [22]	38.75	**61.31**	43.26	34.02	23.99	30.01	52.45	19.13	32.46
IDEC [23]	38.18	51.60	40.86	31.23	27.94	29.91	60.35	18.48	35.23
LSKD [24]	41.54	59.72	45.18	38.44	28.63	34.52	60.46	19.94	36.15
Ours	**43.03**	59.54	**46.33**	**41.38**	**48.61**	**44.27**	**61.99**	**24.26**	**39.35**
$Joint$	–	–	47.50	–	–	47.50	–	–	47.50

where TP_i, FP_i, and FN_i are the numbers of true positive, false positive, and false negative pixels for class i. And C_K is the total number of classes.

Based on Eq. (8), we evaluate the performance of CL methods through three derived metrics: $mIoU_{old}$, $mIoU_{new}$, and mIoU scores. Following prior work such as [22], these metrics are computed after completing all training steps in each CL scenario. Specifically, $mIoU_{old}$ reflects the model's anti-forgetting ability by measuring performance on the old classes C_0, while $mIoU_{new}$ captures plasticity by evaluating performance on newly introduced classes $C_{1:K}$. The overall mIoU on $C_{0:K}$ reflects the trade-off between plasticity and anti-forgetting capability.

4.3. Implementation details

For fairness, we use DeepLab v3 with a ResNet-101 backbone pre-trained on ImageNet as the semantic segmentation model in all the experiments. During the training process of each method, the model is updated using the SGD optimizer, with an initial learning rate of 0.0005. In every step, we train each model for 100 epochs with a batch size of 8. In the training process, data augmentation methods including random rotation and random cropping are applied. The resolution of input images is set to 512 × 512 pixels. For hyper-parameters, we set $\tau = 0.55$, $\lambda_a = 1$, $\lambda_l = 20$, $\lambda_m = 1$. Moreover, for ALIFE [22], we do not consider the replay module. All experiments are conducted on a server equipped with 2 NVIDIA GeForce RTX 3090 GPUs.

4.4. Comparison to competing methods

According to scenarios in Table 1, we compare our method with five advanced general CL methods in Table 2. In addition, we provide the performance of two non-continual learning methods, "Fine-tune" and "Joint", which can be seen as the results of traditional PV defect segmentation methods in the dynamic condition. "Fine-tune" refers to the model updates only through new data, which causes catastrophic forgetting about old classes, and the result can be seen as the lower bound. "Joint" refers to training on both the historical and new data at one time, under the assumption that all historical data are always available and training costs are not a concern, and can be seen as the upper bound.

Under the 3-1 scenario, our method achieves the optimal performance in anti-forgetting, where $mIoU_{old}$ is 1.49% higher than the suboptimal LSKD [24]. Although ALIFE [22] performs better on incremental learning classes, the performance of our method on all classes is optimal. Under 2-2 and 1-1 multi-step scenarios, our method outperforms other methods by at least 1.53% in $mIoU_{old}$ and 4.32% in $mIoU_{new}$. This suggests that we have good effects in reducing forgetting and maintaining model stability through the defect-type and soft-target knowledge.

Moreover, it could be observed from Table 2 that the 3-1 scenario yields higher $mIoU_{new}$ but lower $mIoU_{old}$, while the 1-1 scenario exhibits the opposite pattern. The observed discrepancy reflects the inherent characteristics of the PV defect dataset, i.e., its severe class imbalance and varying learning difficulty among defect classes. Some defect types (e.g., interconnect and corrosion) constitute less than 1% of the training samples, while certain defect types (e.g., interconnect and contact) exhibit subtle morphological features and thin structural patterns, making them more challenging to learn. In the 3-1 scenario, the challenging defect types — interconnect and contact — are treated as old classes, while the new classes consist of easier types. As a result, $mIoU_{new}$ is higher, whereas $mIoU_{old}$ is lower. In contrast, in the 1-1 scenario, these challenging types are treated as new classes, and the old classes consist of easier types. Consequently, the $mIoU_{old}$ is higher and $mIoU_{new}$ is lower.

Regarding the variation of mIoU across different scenarios, more incremental steps make it more difficult to balance the trade-off between plasticity and anti-forgetting capability, leading to lower overall performance. For example, the highest mIoU is achieved in the 3-1 scenario with the fewest incremental steps, while the lowest mIoU is achieved in the 1-1 scenario with the most incremental steps; the 2-2 scenario falls in between.

Fig. 6 illustrates the qualitative evaluation results in 3-1 scenario. The first and second rows display monocrystalline cell images, while the third row shows polycrystalline cells with complex background noise. The second column shows the results of PLOP [21]. It exhibits poor performance for all defects. The third column shows the results of ALIFE [22], which only relies on output distillation. While it can roughly segment the crack defects, there are errors and omissions in the judgment of defect types. It misidentifies contacts as cracks in the polycrystalline cell. The fourth and fifth columns display results from the IDEC [23] and LSKD [24], both employing output and feature distillation. IDEC performs well in segmenting crack and contact but shows errors in corrosion and interconnect defects, particularly in the polycrystalline cells. LSKD yields similar results. Although it excels in segmenting cracks and contacts, it completely fails to identify interconnects in the image. In contrast, our method performs better in both monocrystalline and polycrystalline cells. Our method provides more precise and reliable accuracy in defect classification and segmentation, which is attributed to the proposed SegAdapt. Notably, it is the only

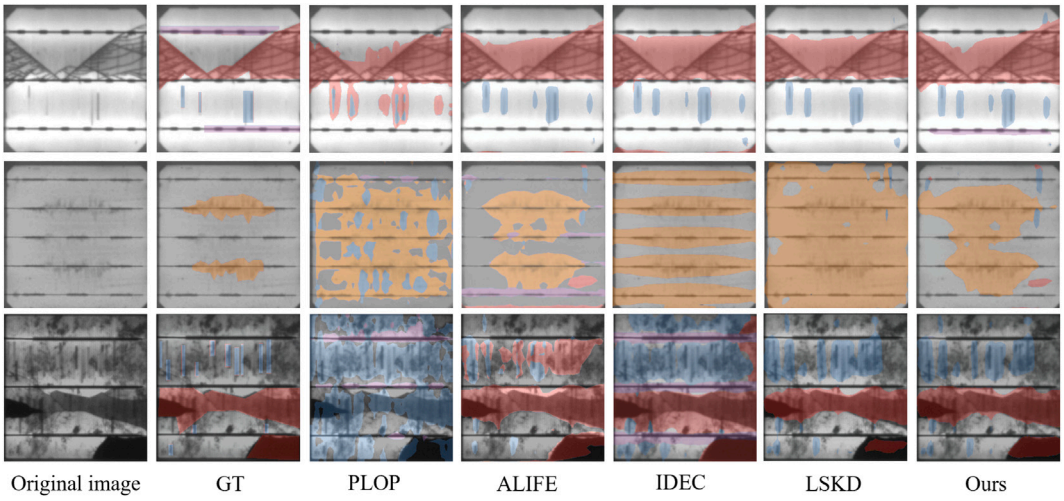

Fig. 6. Qualitative results in 3-1 scenario on the test set, with contact marked in blue, interconnection in purple, corrosion in yellow, and crack in red. GT is referred to as ground truth, which is the true defect mask for the original image. Our method is better than others (e.g., more accurate classification and shape of defects).

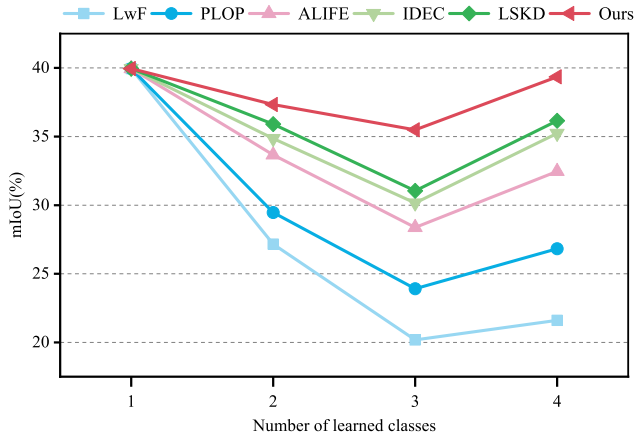

Fig. 7. Further evaluation results in 1-1 scenario on the test set. Our method maintains the highest mIoU after all incremental steps, indicating its superior ability to resist forgetting and adapt well to new types.

approach capable of correctly segmenting thin interconnects. This further reinforces the efficacy of the proposed attention mechanism in emphasizing feature information for slender or thin defects, facilitating better information transmission during feature distillation.

In Fig. 7, we evaluate mIoU as it learns an increasing number of classes in the 1-1 scenario. Overall, the model performance exhibits a trend of initial decline followed by recovery. Specifically, there is a noticeable degradation for all methods when the second and third class is learned, which only rebounds in the fourth class. The drop in performance can be attributed to the introduction of interconnect and corrosion. Interconnects are particularly challenging to learn due to their small quantity and slender nature. Additionally, corrosions rarely co-occur with other defects and, like interconnects, appear in limited numbers. The model struggles to learn interconnects, resulting in poor-quality pseudo-labels. It negatively impacts the model's ability to learn corrosions, further degrading overall performance. Conversely, cracks introduced in the fourth step have a larger quantity and spatial scale, making them easier for the model to learn, thereby facilitating recovery in performance.

Through in-depth analysis, the performance of LwF [29] rapidly declines at the latter steps. Although more advanced methods like IDEC [23] and LSKD [24] perform somewhat better, they still exhibit performance degradation. In contrast, our method shows minimal decline in performance, indicating its superior ability to resist forgetting and adapt well to new types. Overall, our method achieves a commendable balance between plasticity and anti-forgetting capability.

4.5. Ablation results

In this subsection, we discuss three ablation experiments to demonstrate the effectiveness of our method.

4.5.1. Visual comparison w/ and w/o proposed SegAdapt

We visualize the focus regions of the baseline and our method using Gradient-weighted Class Activation Mapping (Grad-CAM) [30]. Fig. 8 demonstrates that our method enhances the model's attention towards slender or thin defects, allowing it to identify areas overlooked by the baseline. Additionally, the results in the second and third columns show that our method can focus on the shape and location of defects more accurately. This further confirms that the proposed feature knowledge distillation method, based on attention mechanisms and auxiliary multi-label recognition effectively highlights defect features, helping the model learn accurate representations of defects and thereby achieving more precise results.

4.5.2. Module contribution

We evaluate the impact of various feature distillation modules in Table 3. As shown in Table 3, each of the proposed modules independently contributes to improvements in both $mIoU_{old}$ and overall mIoU. Among the three modules, the proposed attention distillation contributes the most to enhancing the model's resistance to forgetting. For instance, in 3-1 scenario, incorporating the attention distillation increases $mIoU_{old}$ by 3.57%. Regarding the overall mIoU value, the auxiliary multi-label classification is essential to balance the plasticity and anti-forgetting capability when applied independently. It leads to an increase of 1.62% in the 3-1 scenario and an increase of 5.98% in the 2-2 scenario. Nevertheless, the best performance is achieved when all three modules are combined.

4.5.3. Contribution for the attention mechanism

This study compares the proposed attention mechanisms with other classical attention mechanisms. The comparison methods include the channel attention mechanism SE [16], the channel and spatial attention

Table 3
Ablation results showing the contribution of each module, using ALIFE [22] as the baseline. The abbreviations are as follows: "A. D." stands for attention distillation, "A. L. D." for auxiliary multi-label logit distillation, and "A. M. C." for auxiliary multi-label classification.

Modules			3-1			2-2			1-1		
A. D.	A. L. D.	A. M. C.	$mIoU_{old}$	$mIoU_{new}$	mIoU	$mIoU_{old}$	$mIoU_{new}$	mIoU	$mIoU_{old}$	$mIoU_{new}$	mIoU
			38.73	61.31	43.25	34.02	23.99	30.01	52.45	19.31	32.46
✓			42.30	51.94	44.23	39.16	30.58	35.73	55.35	21.37	34.96
	✓		38.76	63.96	43.80	36.61	28.37	33.31	54.41	19.54	33.49
		✓	42.12	55.86	44.87	39.43	30.83	35.99	55.16	21.74	35.11
	✓	✓	41.50	59.08	45.02	38.20	30.59	35.16	54.87	22.61	35.51
✓	✓		42.81	53.58	44.96	39.60	30.76	36.06	56.77	21.43	35.57
✓		✓	41.78	56.78	44.77	39.33	37.24	38.49	58.17	22.74	36.91
✓	✓	✓	43.03	59.54	46.33	41.38	48.61	44.27	61.99	24.26	39.35

Table 4
Quantitative results (%) of the attention mechanism in scenario 3-1. Time denotes the whole training time. The proposed attention mechanism is efficient in training time and helpful in performance.

	Pub./Year	$mIoU_{old}$	$mIoU_{new}$	mIoU	Time (h)
SE [16]	$CVPR_{18}$	42.70	56.12	45.38	30.79
CBAM [31]	$ECCV_{18}$	40.33	59.30	44.12	35.23
CA [32]	$CVPR_{21}$	41.83	56.07	44.68	35.14
CCA [33]	$CVPR_{24}$	42.41	55.94	45.12	90.92
Ours	–	43.03	59.54	46.33	9.53

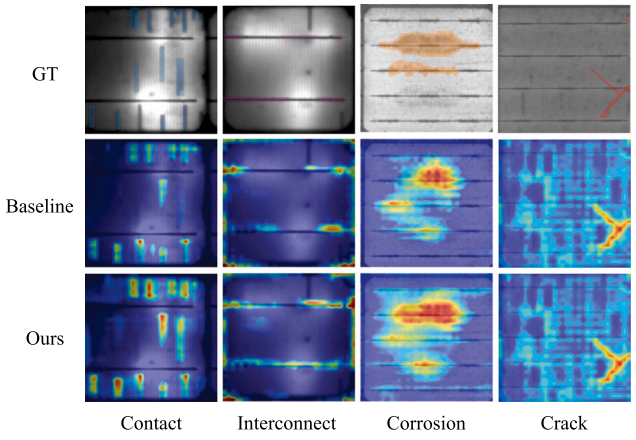

Fig. 8. The performance on the test set of the UCF EL dataset overlaid by Grad-CAM [30]. Various typical defective solar cells are presented. GT, which is the true defect mask is listed as the reference. Baseline is ALIFE, which does not incorporate the modules we proposed. The heat map highlights the area that the model focuses on. Our method significantly enhances the model's attention toward slender or thin defects, allowing it to identify regions overlooked by the baseline.

block CBAM [31], channel attention CA [32], and spatial attention CCA [33].

From Table 4, it can be observed that our attention achieves the best performance in previously learned classes. The $mIoU_{old}$ is 0.33% higher than the second-best SE attention mechanism and 2.70% higher than the weakest CBAM attention mechanism. In terms of overall performance, our method is also the best, outperforming the second-best by 0.95% and the lowest-performing method by 2.21%. Regarding the performance on new types, CBAM shows the best performance. However, it also exhibits the most significant degree of forgetting for learned types. This suggests that plasticity and stability are, to some extent, conflicting goals. Therefore, an ideal continual learning method should strive to achieve a relative balance between these two aspects.

In terms of training time, the proposed attention mechanism is the most efficient, requiring only 9.53 h to complete training with the same number of epochs. In contrast, models that incorporate other attention mechanisms require more than 30 h of training time, with CCA taking as long as 90 h.

Overall, the proposed attention mechanism is not only efficient in training time but also helpful in performance. This indicates that our attention mechanism improves defect segmentation performance while greatly reducing training costs, which is highly beneficial for practical model deployment.

4.5.4. *Effect of class introduction order*

To study the impact of class order, we conducted an additional set of experiments in which the class introduction order followed an easy-to-hard order for 3-1 and 1-1 scenarios. The learning difficulty is estimated based on the per-class IoU of the Joint approach. Accordingly, the defect classes are introduced in the following order: crack, contact, corrosion, and interconnect.

As shown in Table 5, our method consistently achieves the best performance among all CL approaches under 3-1 and 1-1 scenarios. Since the most challenging class is always treated as the new class, it consistently leads to lower $mIoU_{new}$ across all CL methods. These results further confirm that even under the alternative class introduction order, our method can outperform other CL approaches in continual PV defect segmentation tasks.

5. Conclusion

This paper first studies PV defect segmentation under continual learning setting. Fully considering the prior characteristics of defects, we propose a novel knowledge distillation framework SegAdapt tailored for PV defect segmentation in dynamic conditions. SegAdapt facilitates efficient fine-tuning with new data while effectively identifying both old and new defect types. To highlight slender defects, we introduce a fast-normalization attention mechanism to guide relevant feature distillation. To mitigate forgetting and adapt to new types, auxiliary multi-label recognition ensures feature alignment across all encoder levels. Experiments across diverse incremental settings demonstrate that SegAdapt consistently outperforms state-of-the-art general CL methods in photovoltaic defect segmentation, effectively balancing incremental knowledge representation and retention. The consistent performance across various scenarios and various orders validates that our approach — featuring the proposed fast normalization attention and auxiliary multi-label recognition — can significantly enhance the robustness of the knowledge distillation process. In summary, SegAdapt provides a practical knowledge-updating framework for PV defect segmentation, which is highly valuable for real-world intelligent operation and maintenance systems.

Table 5
Ablation results (%) under 3-1 and 1-1 scenarios with the easy-to-hard class introduction order. *Joint* serves as the upper bound.

Method	3-1			1-1		
	$mIoU_{old}$	$mIoU_{new}$	mIoU	$mIoU_{old}$	$mIoU_{new}$	mIoU
LwF [29]	33.68	1.07	27.16	44.55	2.37	19.24
PLOP [21]	38.49	5.93	31.98	45.39	3.11	20.02
ALIFE [22]	41.48	6.33	34.45	49.17	5.84	23.17
IDEC [23]	42.53	5.31	35.09	55.92	3.60	24.53
LSKD [24]	42.79	8.24	35.88	59.51	4.18	26.31
Ours	43.55	10.48	36.94	64.76	7.45	30.37
Joint	–	–	47.50	–	–	47.50

CRediT authorship contribution statement

Xinyue Lei: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Jinxia Zhang:** Writing – review & editing, Supervision, Conceptualization. **Ying He:** Investigation. **Shixiong Fang:** Supervision. **Kanjian Zhang:** Supervision. **Haikun Wei:** Supervision.

Declaration of competing interest

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

Acknowledgments

This work was supported by National Natural Science Fund of China (Grant number 62573123), Research Fund for Advanced Ocean Institute of Southeast University, Nantong (GP202411), and the Fundamental Research Funds for the Central Universities. We also thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

[1] Z. Jiang, G. Du, Z. Song, S. Cao, G. Wang, L. Ma, Electromagnetic method for detecting black piece on monocrystalline silicon photovoltaic panels, IEEE Trans. Instrum. Meas. 73 (2024) 1–8, http://dx.doi.org/10.1109/TIM.2024.3351250.

[2] Abu Dhabi, Renewable energy statistics 2024, 2024, The International Renewable Energy Agency, Abu Dhabi, Jul. 2024. [Online]. Available: https://www.irena.org/Publications/2024/Jul/Renewable-energy-statistics-2024.

[3] J. Fioresi, D.J. Colvin, R. Frota, R. Gupta, M. Li, H.P. Seigneur, S. Vyas, S. Oliveira, M. Shah, K.O. Davis, Automated defect detection and localization in photovoltaic cells using semantic segmentation of electroluminescence images, IEEE J. Photovoltaics 12 (1) (2022) 53–61, http://dx.doi.org/10.1109/JPHOTOV.2021.3131059.

[4] C. Li, Y. Yang, S. Spataru, K. Zhang, H. Wei, A robust parametrization method of photovoltaic modules for enhancing one-diode model accuracy under varying operating conditions, Renew. Energy 168 (2021) 764–778, http://dx.doi.org/10.1016/j.renene.2020.12.097.

[5] M. De Riso, S. Hassan, P. Guerriero, M. Dhimish, S. Daliento, Enhanced photovoltaic panel diagnostics: Advancing a high-precision and low-cost I – V curve tracer, IEEE Trans. Instrum. Meas. 73 (2024) 1–10, http://dx.doi.org/10.1109/TIM.2024.3484517.

[6] A.K.V. de Oliveira, M. Aghaei, R. Rüther, Automatic inspection of photovoltaic power plants using aerial infrared thermography: A review, Energies 15 (6) (2022) http://dx.doi.org/10.3390/en15062055.

[7] J. Zhang, Y. Shen, J. Jiang, S. Fang, L. Chen, T. Yan, Z. Li, K. Zhang, H. Wei, W. Guo, Automatic detection of defective solar cells in electroluminescence images via global similarity and concatenated saliency guided network, IEEE Trans. Ind. Inform. 19 (6) (2023) 7335–7345, http://dx.doi.org/10.1109/TII.2022.3211088.

[8] P. Zhou, R. Wang, C. Wang, H. Chen, K. Liu, SIIF: Semantic information interactive fusion network for photovoltaic defect segmentation, Appl. Energy 371 (2024) 123643, http://dx.doi.org/10.1016/j.apenergy.2024.123643.

[9] X. Chen, T. Karin, C. Libby, M. Deceglie, P. Hacke, T.J. Silverman, A. Jain, Automatic crack segmentation and feature extraction in electroluminescence images of solar modules, IEEE J. Photovoltaics 13 (3) (2023) 334–342, http://dx.doi.org/10.1109/JPHOTOV.2023.3249970.

[10] H. Chen, H. Zhao, D. Han, W. Liu, P. Chen, K. Liu, Structure-aware-based crack defect detection for multicrystalline solar cells, Measurement 151 (2020) 107170, http://dx.doi.org/10.1016/j.measurement.2019.107170.

[11] S.A. Anwar, M.Z. Abdullah, Micro-crack detection of multicrystalline solar cells featuring shape analysis and support vector machines, in: Proceeding of the IEEE International Conference on Control System, Computing and Engineering, 2012, pp. 143–148, http://dx.doi.org/10.1109/ICCSCE.2012.6487131.

[12] J. Zhang, X. Chen, H. Wei, K. Zhang, A lightweight network for photovoltaic cell defect detection in electroluminescence images based on neural architecture search and knowledge distillation, Appl. Energy 355 (2024) http://dx.doi.org/10.1016/j.apenergy.2023.122184.

[13] B. Su, H. Chen, P. Chen, G. Bian, K. Liu, W. Liu, Deep learning-based solar-cell manufacturing defect detection with complementary attention network, IEEE Trans. Ind. Inform. 17 (6) (2021) 4084–4095, http://dx.doi.org/10.1109/TII.2020.3008021.

[14] S. Wang, H. Chen, Z. Zhang, B. Su, Multi-scale feature decoupling and similarity distillation for class-incremental defect detection of photovoltaic cells, Measurement 225 (2024) 113997, http://dx.doi.org/10.1016/j.measurement.2023.113997.

[15] Y. Jiang, C.H. Zhao, Attention classification-and-segmentation network for micro-crack anomaly detection of photovoltaic module cells, Sol. Energy 238 (2022) 291–304, http://dx.doi.org/10.1016/j.solener.2022.04.012.

[16] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 7132–7141, http://dx.doi.org/10.1109/CVPR.2018.00745.

[17] M. McCloskey, N.J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of Learning and Motivation, vol. 24, Academic Press, 1989, pp. 109–165, http://dx.doi.org/10.1016/S0079-7421(08)60536-8.

[18] F. Cermelli, M. Mancini, S. Rota Bulo, E. Ricci, B. Caputo, Modeling the background for incremental learning in semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 9230–9239, http://dx.doi.org/10.1109/cvpr42600.2020.00925.

[19] T. Kalb, B. Mauthe, J. Beyerer, Improving replay-based continual semantic segmentation with smart data selection, in: Proceedings of the IEEE 25th International Conference on Intelligent Transportation Systems, ITSC, 2022, pp. 1114–1121, http://dx.doi.org/10.1109/ITSC55140.2022.9922284.

[20] L. Zhu, T. Chen, J. Yin, S. See, J. Liu, Continual semantic segmentation with automatic memory sample selection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 3082–3092, http://dx.doi.org/10.1109/CVPR52729.2023.00301.

[21] A. Douillard, Y. Chen, A. Dapogny, M. Cord, PLOP: Learning without forgetting for continual semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 4040–4050, http://dx.doi.org/10.1109/CVPR46437.2021.00403.

[22] Y. Oh, D. Baek, B. Ham, ALIFE: Adaptive logit regularizer and feature replay for incremental semantic segmentation, in: Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 14516–14528, http://dx.doi.org/10.5555/3600270.3601325.

[23] D. Zhao, B. Yuan, Z. Shi, Inherit with distillation and evolve with contrast: Exploring class incremental semantic segmentation without exemplar memory, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 11932–11947, http://dx.doi.org/10.1109/TPAMI.2023.3273574.

[24] Q. Wang, Y. Wu, L. Yang, W. Zuo, Q. Hu, Layer-specific knowledge distillation for class incremental semantic segmentation, IEEE Trans. Image Process. 33 (2024) 1977–1989, http://dx.doi.org/10.1109/TIP.2024.3372448.

[25] A. Mao, M. Mohri, Y. Zhong, Cross-entropy loss functions: Theoretical analysis and applications, in: Proceedings of the 40th International Conference on Machine Learning, ICML, 2023, pp. 23803–23828, http://dx.doi.org/10.48550/arXiv.2304.07288.

[26] J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: A survey, 2020, http://dx.doi.org/10.48550/arXiv.2006.05525, ArXiv E-Prints.

[27] G. Tsoumakas, I.M. Katakis, Multi-label classification: An overview, Int. J. Data Warehous. Min. 3 (2007) 1–13, http://dx.doi.org/10.4018/jdwm.2007070101.

[28] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136, http://dx.doi.org/10.1007/s11263-014-0733-5.

[29] Z. Li, D. Hoiem, Learning without forgetting, IEEE Trans. Pattern Anal. Mach. Intell. 40 (12) (2018) 2935–2947, http://dx.doi.org/10.1109/tpami.2017.2773081.

[30] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2017, pp. 618–626, http://dx.doi.org/10.1109/ICCV.2017.74.

[31] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19, http://dx.doi.org/10.1007/978-3-030-01234-2_1.

[32] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 13708–13717, http://dx.doi.org/10.1109/CVPR46437.2021.01350.

[33] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, Y. Yao, Poly kernel inception network for remote sensing detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 27706–27716, http://dx.doi.org/10.1109/CVPR52733.2024.02617.