Solar Energy

Lightweight Vision Architecture with Mutual Distillation for Robust Photovoltaic Defect Detection in Complex Environments

--Manuscript Draft--

Manuscript Number:	SEJ-D-24-04166R1
Article Type:	Research paper
Section/Category:	Photovoltaic Cells and Cell Physics
Keywords:	Photovoltaic defect detection; Lightweight deep learning; Mutual distillation; Edge computing; Electroluminescence imaging
Abstract:	With the rapid growth of solar photovoltaic installations, defect detection in PV power stations has become crucial for ensuring operational safety and economic efficiency, as undetected defects can lead to significant performance degradation and potential hazards. Unmanned Aerial Vehicle (UAV)-based Electroluminescence (EL) imaging offers an efficient solution for large-scale inspection. However, the harsh environmental conditions and complex imaging scenarios pose significant challenges to detection models, while edge computing deployment demands strict resource constraints. This study introduces SCRViT, a lightweight deep learning model that substantially improves detection performance on low-quality EL images through a spatial-channel reconstruction mechanism and a peer network co-learning strategy. Experimental results demonstrate that the proposed method achieves 88.19% detection accuracy on simulated outdoor environment datasets, surpassing state-of-the-art approaches by 4.77% while reducing model parameters by 55.6%. Through multi-dimensional interpretability studies—including Shapley value feature attribution, GradCAM attention pattern analysis, and information-theoretic mechanism analysis—this research systematically elucidates the model's environmental adaptation mechanisms. This lightweight yet robust solution enables real-time defect detection on edge devices, improving inspection efficiency and reducing operational costs while providing reliable decision support for practical applications in complex outdoor environments.

Authors' Response to Reviews of

Solar Energy

Ms. Ref. No.: SEJ-D-24-04166

Lightweight Vision Architecture with Mutual Distillation for Robust Photovoltaic Defect Detection in Complex Environments,

---

RC: ***Reviewers' Comment***, AR: Authors' Response, ☐ Manuscript Text

Dear Editor and Reviewers,

We sincerely thank you for taking the time to review our manuscript "Lightweight Vision Architecture with Mutual Distillation for Robust Photovoltaic Defect Detection in Complex Environments" (Ms. Ref. No.: SEJ-D-24-04166). Your insightful comments and constructive suggestions have greatly helped us improve the quality of our work. We particularly appreciate your careful reading and thoughtful feedback, which have led to significant improvements in both the technical content and presentation clarity of our manuscript.

We have carefully addressed all the comments and made corresponding revisions to the manuscript. All changes are marked in red in the revised version. Below we provide our detailed point-by-point responses to each comment. We hope the revised manuscript better meets your expectations and standards for publication in Solar Energy.

1. Reviewer #1

1.1. Reply 1

RC: ***Is there any difference in detection accuracy based on the method for different types of defects, such as microcracks, broken grid lines, dark spots, etc.? Are there any related experiments?***

AR: Thank you for your valuable question about defect-specific detection performance. We want to clarify that our current study focuses on binary classification (defective/non-defective) for both mono-crystalline and poly-crystalline modules, which is consistent with the annotation granularity provided in the public ELPV dataset.

1. Public Dataset: The ELPV dataset, which serves as our primary benchmark, only provides binary annotations (defective/non-defective) for both mono-crystalline and poly-crystalline modules.

2. Industrial Practice Dataset: While we have collected EL images from actual production lines with specific defect type annotations (including microcracks, broken grid lines, dark spots, etc.), we encountered several practical challenges in building a balanced dataset. The real-world defect occurrence exhibits significant class imbalance, with certain defect types being particularly rare in the production environment. This natural imbalance, combined with the high costs of EL imaging equipment and the time-intensive nature of expert annotation in production environments, made it impractical to obtain sufficient samples across all defect categories for meaningful statistical analysis.

Due to these limitations, particularly the severely imbalanced sample distribution across different defect types, we were unable to conduct statistically meaningful experiments on type-specific detection performance. We acknowledge that such analysis would provide valuable insights for practical applications, and

have identified this as an important direction for future work when a more balanced and comprehensive dataset becomes available.

1.2. Reply 2 page 4

RC: ***In section 3.2, there is a point that "This adaptive feature reconstruction mechanism aligns precisely with the characteristics of EL image defect detection tasks, effectively addressing image quality issues cause by environmental factors". Could you please elaborate on the "align precisely" and how to "effectively addressing"?***

AR: Thank you for this insightful comment. We agree that the original description of the mechanism's alignment with EL image characteristics and its effectiveness needed more precise elaboration. We have revised this part to provide a more detailed and specific explanation of how our module's architecture addresses EL image detection challenges.

> We leverage the spatial and channel reconstruction module to systematically~~adaptively~~[R1] enhance RepViT's Stage structure. By incorporating SCConv into each Stage's downsampling module and employing structural re-parameterization, we optimize inference efficiency. The module's dual-unit architecture specifically addresses EL image detection challenges: the Spatial Reconstruction Unit handles non-uniform illumination and motion blur through adaptive feature separation and reconstruction, while the Channel Reconstruction Unit captures diverse defect patterns through efficient group transformation. This mechanism provides a robust feature foundation for subsequent defect identification by maintaining discriminative features under environmental interference while reducing computational redundancy.~~This adaptive feature reconstruction mechanism aligns precisely with the characteristics of EL image defect detection tasks, effectively addressing image quality issues caused by environmental factors while providing a robust feature foundation for subsequent defect identification.~~[R1]

In this revision, we have:

1. Replaced the vague phrase "aligns precisely" with a concrete explanation of how each unit in our dual-unit architecture specifically addresses different aspects of EL image detection challenges.

2. Elaborated on the "effectively addressing" by detailing how the Spatial Reconstruction Unit handles environmental factors (illumination variation and motion blur) and how the Channel Reconstruction Unit processes various defect patterns.

3. Added specific mechanisms (feature separation, reconstruction, and group transformation) to explain how the module maintains feature quality while reducing computational overhead.

1.3. Reply 3 page 6

RC: ***In the simulated UAV dataset, what is the reference basis of the image degradation operations for UAV-based inspection scenarios? What is the basis for determining the parameters? For instance, how to reflect the influence of the UAV's flight altitude, as well as the impact of pollution in the usage environment on photovoltaic panels.***

AR: We appreciate the reviewer's insightful questions about our dataset simulation. Let us address these questions one by one:

1. Regarding the reference basis for image degradation operations: The selection of degradation factors

(illumination variation, motion blur, and image quality degradation) is based on the comprehensive analysis of air-to-air visual detection challenges presented in [36]. This work systematically analyzed the key challenges in UAV-based detection, including complex backgrounds, varying viewing angles, and environmental interference factors.

2. Regarding the basis for determining specific parameters: The parameter values were determined through a practical calibration process. We obtained reference EL images from professional PV inspection institutions (due to commercial confidentiality, only a limited number of sample images were shared). Then, we conducted extensive parameter tuning experiments using image processing libraries to match our simulated degradation effects with these real-world reference images. Specifically, we iteratively adjusted the parameters (brightness, contrast, blur kernel size, etc.) until the degraded images exhibited similar visual characteristics to the reference images. While this approach is empirical in nature, it was a practical solution given that UAV-based EL inspection is a relatively new field without standardized public datasets for reference.

3. Regarding flight altitude and pollution effects:

- Flight Altitude: We acknowledge this limitation in our current work. The impact of flight altitude on image quality is an important factor that we plan to incorporate in future iterations of our dataset simulation framework.
- Surface Pollution: Based on practical inspection experience, the probability of significant surface pollution affecting EL imaging is relatively low, as regular maintenance procedures typically prevent such issues. Therefore, this factor was not prioritized in our current simulation framework.

These limitations point to important directions for future research, particularly in developing more comprehensive simulation frameworks that incorporate a wider range of real-world factors.

Based on the reviewer's suggestions, we have revised Section 4.1.2 to clarify the basis of our parameter selection and acknowledge the current limitations in simulating certain environmental factors. The specific modifications are as follows:

> To systematically evaluate model performance in real-world UAV inspection scenarios [36][R1], we construct a dataset with simulated environmental interference. Built upon the ELPV dataset, this dataset incorporates three typical environmental disturbance factors (illumination variation, motion blur, and image quality degradation) to simulate actual imaging conditions during UAV inspection. The specific degradation parameters were determined through iterative calibration experiments based on reference images obtained from professional PV inspection institutions.[R1]

1.4. Reply 4 page 7

RC: ***The descriptions for Figure 6 does not match the actual contents.***

AR: We appreciate the reviewer pointing out this inconsistency. After careful review, we found that the description of image layout in Figure 6 was indeed inaccurate. We have revised the figure caption to correctly reflect the content arrangement.

> Visualization of environmental factors' progressive impact on EL image quality. Top row shows the original image and three single environmental effects: brightness/contrast (B), motion blur (M),

and posterize effect (P); Bottom row shows combined effects of multiple factors: B+M (brightness + motion blur), B+P (brightness + posterize), M+P (motion blur + posterize), and B+M+P (all three factors combined).~~Visualization of environmental factors' progressive impact on EL image quality. First row shows the original image and single effects of brightness/contrast (B) and motion blur (M); Second row shows posterize effect (P) and two-factor combinations (B+M, B+P); Third row shows the remaining two-factor combination (M+P) and the final combined effect (B+M+P). This progression demonstrates how multiple environmental factors jointly degrade image quality in UAV-based inspection scenarios.~~[R1]

The revised caption now accurately describes the layout and content of the figure, providing clearer guidance for readers to understand the progression of environmental effects on image quality.

2. Reviewer #2

2.1. Reply 1 page 1

RC: ***Please add references for the paragraph starting with "However, UAV-based EL detection....." from row number 55-59 of the First page.***

AR: Thank you for this valuable suggestion. We fully agree that the technical challenges of UAV-based detection should be properly supported by citations. We have added a highly relevant reference that specifically addresses the impact of environmental factors on UAV-based inspection quality. The revised text now reads:

However, UAV-based EL detection in outdoor environments faces severe technical challenges that significantly impact image quality[11][R2]. Ambient light interference leads to substantial reduction in image signal-to-noise ratio and contrast, making defect features less distinguishable. UAV vibration and attitude variations cause image blur, compromising the fine details crucial for defect identification. Additionally, environmental factors such as wind speed and temperature variations affect image acquisition quality, introducing noise and distortions. These factors collectively degrade field-acquired EL images compared to laboratory standards, creating a significant gap between ideal and practical conditions. Current deep learning approaches encounter a critical dilemma in addressing these challenges. While existing models achieve high detection accuracy under ideal darkroom conditions, they are constrained by their computational complexity and parameter scale for edge device deployment. Conversely, lightweight models that meet computational constraints show significantly reduced accuracy when processing low-quality outdoor EL images. This limitation becomes particularly critical as the rapid expansion of PV installations demands more reliable and efficient inspection solutions.

The added reference [11] provides a systematic study of environmental impacts on UAV-based visual inspection quality, particularly focusing on how wind conditions and vehicle motion affect image quality. This work strongly supports our discussion of the technical challenges faced in outdoor EL detection.

2.2. Reply 2 page 2

RC: ***Please provide references for the paragraph starting with "Current deep learning approaches..." from rows 25 to 35 of the Second page.***

AR: Thank you for your suggestion. We have added a relevant reference to support our discussion about the limitations of lightweight models. The revised text now reads:

Current deep learning approaches encounter a critical dilemma in addressing these challenges. While existing models achieve high detection accuracy under ideal darkroom conditions, they are constrained by their computational complexity and parameter scale for edge device deployment[12][R2]. Conversely, lightweight models that meet computational constraints show significantly reduced accuracy when processing low-quality outdoor EL images. This limitation becomes particularly critical as the rapid expansion of PV installations demands more reliable and efficient inspection solutions.

The added reference [12] provides a comprehensive overview of lightweight deep learning models, particularly addressing the trade-off between model efficiency and accuracy, which directly supports our discussion of the challenges in deploying deep learning models on edge devices.

2.3. Reply 3 page 2

RC: ***Please add an elaboration of CNN and other keywords which are extensively used.***

AR: We sincerely thank the reviewer for this constructive suggestion to improve the clarity of our manuscript. We have added detailed explanations of key technical terms when they first appear in the paper. The specific modifications are as follows:

Photovoltaic (PV) module electroluminescence (EL) image defect detection has evolved from traditional methods to lightweight deep neural networks. Deitsch et al. [13] systematically demonstrated Convolutional Neural Network (CNN), a deep learning architecture that uses convolution operations to automatically extract hierarchical visual features,[R2] advantages over traditional SVM methods in automated defect detection, providing a foundation for deep learning approaches in this field.

Lightweight neural network architecture design has emerged as a core research direction in deep learning. In the CNN domain, MobileNetV3 [18] achieved efficient feature extraction through hardware-aware architecture search and innovative structural design, while EfficientNetV2 [19] optimized network architecture through training-aware architecture search and compound scaling strategies.Recently, Transformer architectures, which originated from natural language processing and employ self-attention mechanisms to model long-range dependencies in input data,[R2] have demonstrated unique advantages in lightweight visual model design.

Among detection methods, current-voltage (IV) analysis struggles with minor defects [3], infrared (IR) imaging shows inconsistencies between hotspots and actual defects [4], while electroluminescence (EL) imaging, a non-destructive inspection technique that captures light emission from solar cells under forward bias conditions to reveal various types of defects including microcracks and inactive areas,[R2] offers superior reliability in capturing microscopic module features [5].

These elaborations help readers better understand the technical concepts and their applications in the context of our research. We appreciate the reviewer's suggestion which has improved the accessibility of our manuscript.

2.4. Reply 4 page 9

RC: ***The article discusses environmental challenges. Could it provide more details on how these challenges were specifically addressed during the experiments? Including examples of the model's performance under different conditions would make the research more practically relevant.***

AR: We appreciate the reviewer's suggestion for providing more detailed analysis of our model's performance under different environmental conditions. We have added comprehensive experimental results and analysis to address this concern:

To further understand the specific impact of different environmental factors, we conducted detailed experiments with various environmental parameter configurations. The results are presented in Table 9. [R2]

Table 1: Impact of Different Environmental Factors on Model Performance

Environmental Factors	Parameters	Acc(%)	Prec(%)	Rec(%)	F1(%)
Baseline	B(0.0), M(0), P(8)	88.67	87.78	83.03	85.33
Brightness (B)	B(0.5), M(0), P(8)	88.05	87.36	83.35	85.31
	B(-0.5), M(0), P(8)	85.00	83.90	79.07	81.41
Motion Blur (M)	B(0.0), M(7), P(8)	88.24	91.59	80.89	85.89
Bit Depth (P)	B(0.0), M(0), P(7)	88.05	87.90	82.83	85.28
B+M	B(-0.5), M(7), P(8)	87.68	86.53	82.76	84.60
B+P	B(-0.5), M(0), P(7)	87.88	87.41	81.89	84.56
M+P	B(0.0), M(7), P(7)	87.82	86.50	82.38	84.38
B+M+P	B(-0.5), M(7), P(7)	87.50	85.16	83.53	84.34

The experimental results reveal several important insights about model robustness. First, brightness reduction (B=-0.5) shows the most significant impact with a 3.67% accuracy decrease (from 88.67% to 85.00%), while motion blur (M=7) demonstrates a relatively minor effect on accuracy (-0.43%) but notably improves precision by 3.81 percentage points (from 87.78% to 91.59%). Second, bit depth reduction (P=7) causes minimal performance degradation (-0.62% in accuracy) while maintaining similar precision and recall levels. Most importantly, the model exhibits strong resilience to combined environmental factors, with the three-factor combination (B+M+P) reducing accuracy by only 1.17 percentage points (from 88.67% to 87.50%) while maintaining high recall (83.53%) and F1-score (84.34%). Notably, the dual-factor combinations (B+M, B+P, M+P) all achieve accuracy above 87.50%, with performance degradation consistently less than 1% compared to single-factor scenarios. These findings quantitatively validate our model's environmental adaptation capabilities, particularly its ability to maintain stable performance under complex environmental perturbations, providing valuable guidance for practical deployments in outdoor conditions. [R2]

2.5. Reply 5 page 2

RC: ***Please provide references for equations wherever necessary.***

AR: We sincerely thank the reviewer for this constructive suggestion. We have carefully reviewed all equations in the manuscript and added appropriate references to their original sources. The modifications include adding citations for the key equations in the following sections:

1. For the SCConv module equations in Section 3.2, we have added reference to Li et al. [**?**] who originally proposed this module;

2. For the ESE module equations in Section 3.3, we have cited Lee and Park [**?**] as these equations are based on their Enhanced Squeeze-and-Excitation module;

3. For the Deep Mutual Learning equations in Section 3.4, we have added reference to Zhang et al. [**?**], who introduced this learning framework;

4. For the Shapley value analysis equations in Section 4.5, we have cited Lundberg and Lee [**?**], who proposed the SHAP framework.

The specific modifications are as follows:

The SCConv module [33][R2] consists of two core components: a Spatial Reconstruction Unit (SRU) and a Channel Reconstruction Unit (CRU). The SRU addresses feature distribution heterogeneity through a separate-and-reconstruct mechanism. Given an input feature map $X \in \mathbb{R}^{N\times C\times H\times W}$, group normalization is first applied:

$$X_{norm} = GN(X) = \gamma \frac{X-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{1}$$

The ESE module [34][R2] functions as an intelligent filter, dynamically adjusting channel weights based on feature statistics. Similar to multi-channel signal processing, this adaptive weighting mechanism emphasizes channels containing critical defect information while suppressing those degraded by environmental factors, enabling more precise feature extraction. For an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the module first applies global average pooling:

$$s_c = F_{gap}(X_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_c(i,j) \tag{2}$$

Given N samples $X = \{x_i\}_{i=1}^{N}$ from M classes with corresponding label set $Y = \{y_i\}_{i=1}^{N}$, where $y_i \in \{1, 2, \ldots, M\}$, the probability of sample x_i belonging to class m processed by neural network Θ_1 is calculated as [32][R2]:

$$p_1^m(x_i) = \frac{\exp(z_1^m)}{\sum_{m=1}^{M}\exp(z_1^m)} \tag{3}$$

To gain deeper insights into how environmental factors influence SCRViT's decision-making mechanism, we conducted feature attribution analysis based on Shapley values [40][R2]. Shapley values, rooted in cooperative game theory, quantify feature importance by evaluating the marginal contribution of each feature across all possible feature subset combinations. This approach provides a theoretically grounded framework for understanding model decisions by considering both individual feature effects and their interactions. For an input image x and predicted class c, the Shapley value of feature

i is defined as:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{4}$$

These additions provide proper attribution to the original works and help readers access relevant background information. We believe these modifications have improved the academic rigor of our manuscript.

2.6. Reply 6

RC: ***Visualization of this research could be improved with a flowchart which would help to clarify complex concepts and make it more engaging for readers.***

AR: Thank you for this constructive suggestion. We would like to point out that Figure 2 in our manuscript already serves this purpose effectively:

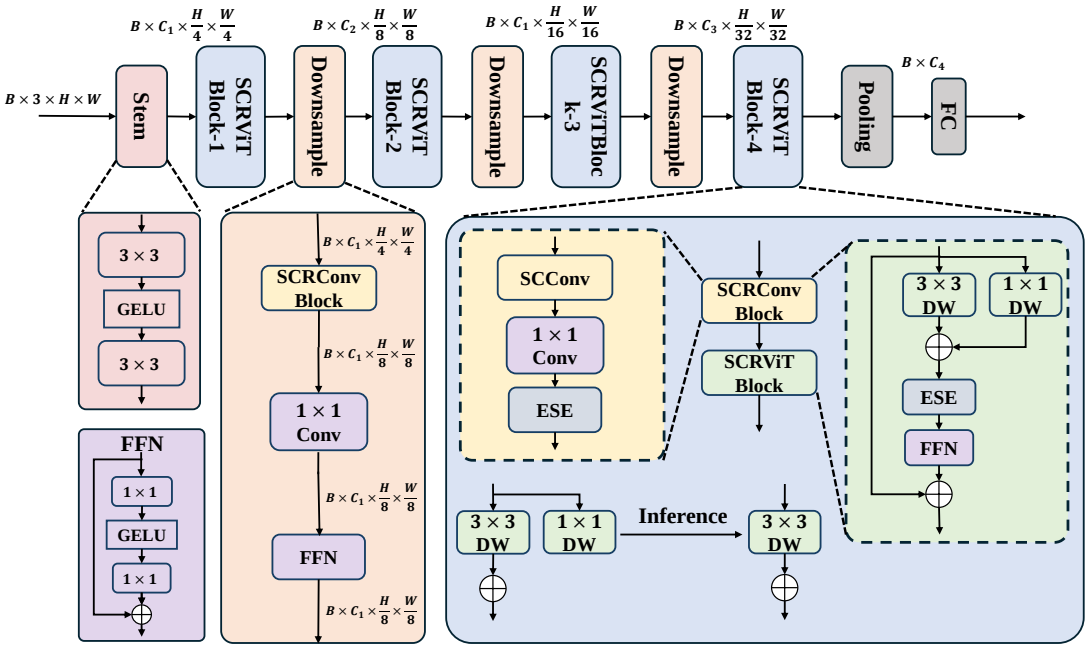

Figure 1: The overall architecture of SCRViT. The Stem block performs initial feature extraction through strided convolutions and FFN layers. SCRViT Block integrates SCConv for feature reconstruction and ESE for adaptive channel attention. The SCConv Block contains a 1Œ1 convolution followed by feature reconstruction units, while the inference path shows the model's simplified structure during deployment. Each component processes features with specific spatial dimensions (HŒW) and channel numbers (C_i), enabling progressive feature refinement across different scales.

This figure provides a comprehensive visualization of our proposed method, clearly illustrating:

1. The complete architectural pipeline from input to output

2. The internal structure of key components (Stem block, SCRViT Block, SCConv Block)

3. The feature processing flow at different scales

4. The model's deployment structure

We believe this figure effectively helps readers understand the complex concepts and relationships between different components of our method.

3. Reviewer #3

3.1. Reply 1

RC: ***The study uses a simulated UAV dataset for testing. While the authors attempted to simulate realistic imaging conditions by introducing distortions to ideal images, demonstrating the fidelity of these distortions to real-world EL image characteristics would strengthen the study.***

AR: Thank you for this constructive comment about our dataset simulation approach. We chose to create a simulated dataset because UAV-based EL imaging for PV inspection is still an emerging field without publicly available datasets for outdoor conditions. Real data collection faces significant practical challenges due to equipment costs, environmental variations, and the need for expert annotations. We acknowledge that our current simulation approach has limitations in fully capturing all real-world environmental effects and UAV flight conditions. These limitations point to important directions for future research, where we plan to collect more real-world UAV-based EL images and develop more comprehensive environmental simulation methods as the field continues to develop.

3.2. Reply 2

RC: ***The manuscript also mentions an Industrial Practice Dataset comprising images captured with high-precision CCD cameras. Including examples of these images would enhance clarity.***

AR: Thank you for this constructive suggestion. This Industrial Practice Dataset was collected by our research team in collaboration with PV manufacturers using high-precision CCD cameras. While we currently cannot share these images due to ongoing research and data processing requirements, we plan to make this dataset publicly available along with our source code once all necessary preparations and permissions are completed. This will help promote reproducible research and benefit the broader community working on PV defect detection.

3.3. Reply 3

RC: ***The final paragraph on page 11 repeatedly references Figure 9; this should be corrected to Figure 8.***

AR: Thank you for catching this inconsistency. We have corrected all references to Figure 8 in the text. The revised paragraph now consistently references the correct figure number:

> As shown in Figure 8(a)[R3], mono-crystalline samples demonstrate linear attention patterns with uniform distribution across busbar regions (Mean=0.078) in normal samples and focused attention on anomalous areas (Max=0.135) in defective samples, while poly-crystalline samples show more dispersed patterns (normal: 0.082, defective: 0.092) due to complex grain boundaries. The quantitative metrics in Figure 8[R3](b) illustrate that attention concentration decreases with defect occurrence (nor-

mal mono: 11.605, defect mono: 8.389, defect poly: 5.556), supported by key region ratios (defective: 0.161/0.144 vs. normal: 0.1411/0.1118 for mono/poly). Figure 8(c,d)[R3] reveals SCRViT's adaptive attention mechanism through spatial variance changes (mono: 0.008→0.003, poly: stable at 0.005) and comprehensive radar analysis.

3.4. Reply 4

RC: ***The manuscript should clarify the implementation details of the proposed network. Making the code, including the testing procedures, publicly available would significantly benefit the scientific community.***

AR: Thank you for this valuable suggestion. We agree that sharing implementation details and code is crucial for research reproducibility. We plan to release our complete source code, including model implementation, training and testing procedures on Github once our work is published. This will help promote transparent research and enable other researchers to build upon our work for further advancements in PV defect detection.

Lightweight Vision Architecture with Mutual Distillation for Robust Photovoltaic Defect Detection in Complex EnvironmentsLightweight Vision Architecture with Mutual Distillation for Robust Photovoltaic Defect Detection in Complex Environments

Haoran Zhang[a], Boao Gong[a], Bohan Ma[a], Zhiyong Tao[a,*], Shi Wang[a]

[a]*School of Electronics and Information Engineering,Liaoning University of Engineering and Technology, 188 Longwan South Street, Longwan Campus, Huludao, 125105, Liaoning, China*

Abstract

With the rapid growth of solar photovoltaic installations, defect detection in PV power stations has become crucial for ensuring operational safety and economic efficiency, as undetected defects can lead to significant performance degradation and potential hazards. Unmanned Aerial Vehicle (UAV)-based Electroluminescence (EL) imaging offers an efficient solution for large-scale inspection. However, the harsh environmental conditions and complex imaging scenarios pose significant challenges to detection models, while edge computing deployment demands strict resource constraints. This study introduces SCRViT, a lightweight deep learning model that substantially improves detection performance on low-quality EL images through a spatial-channel reconstruction mechanism and a peer network co-learning strategy. Experimental results demonstrate that the proposed method achieves 88.19% detection accuracy on simulated outdoor environment datasets, surpassing state-of-the-art approaches by 4.77% while reducing model parameters by 55.6%. Through multi-dimensional interpretability studies—including Shapley value feature attribution, GradCAM attention pattern analysis, and information-theoretic mechanism analysis—this research systematically elucidates the model's environmental adaptation mechanisms. This lightweight yet robust solution enables real-time defect detection on edge devices, improving inspection efficiency and reducing operational costs while providing reliable decision support for practical applications in complex outdoor environments.

Keywords: Photovoltaic defect detection, Lightweight deep learning, Mutual distillation, Edge computing, Electroluminescence imaging

1. Introduction

Over the past decade, global solar photovoltaic (PV) installed capacity has experienced remarkable growth, driven by both increasing demand for low-carbon energy and technological advancements [1]. However, PV power stations are predominantly constructed in remote areas with harsh environmental conditions, making them vulnerable to various degradation factors such as snow accumulation and dust deposition. Research shows that PV module failures make up over 70% of total system failures [2], causing both economic losses and potential safety risks. This challenge is particularly pronounced in utility-scale PV installations—individual power stations typically span tens to hundreds of hectares and comprise hundreds of thousands of modules, generating hundreds of gigabytes of inspection data. These characteristics pose significant challenges to fault detection systems in terms of real-time performance, accuracy, and scalability.

Among detection methods, current-voltage (IV) analysis struggles with minor defects [3], infrared (IR) imaging shows inconsistencies between hotspots and actual defects [4], while electroluminescence (EL) imaging, a non-destructive inspection technique that captures light emission from solar cells under forward bias conditions to reveal various types of defects including microcracks and inactive areas,[R2] offers superior reliability in capturing microscopic module features [5].Automated current injection through dedicated inverters has further enhanced EL detection efficiency. For efficient inspection of large-scale PV plants, intelligent monitoring systems integrating Unmanned Aerial Vehicles (UAVs) and IoT devices have demonstrated significant advantages [6]. These systems, combining UAV-mounted EL imaging equipment with edge computing for fault detection, effectively address the data latency issues inherent in traditional cloud-based solutions [7]. Figure 1 illustrates the proposed IoT-based architecture for EL image defect detection.

Internet of Things (IoT) technology enables efficient data collection and transmission through integrated sensors and communication systems [8], providing a foundation for large-scale fault detection. While cloud computing can process massive amounts of data, it faces bandwidth limitations and latency issues [9], leading to the adoption of Mobile Edge Computing (MEC) [10] solutions.

However, UAV-based EL detection in outdoor environments faces severe technical challenges that significantly impact image quality[11][R2]. Ambient light interference leads to substantial

*Corresponding author

Email addresses: 2206030423@stu.lntu.edu.cn (Haoran Zhang), 2206030403@stu.lntu.edu.cn (Boao Gong), 2306030115@stu.lntu.edu.cn (Bohan Ma), taozhiyong@lntu.edu.cn (Zhiyong Tao), wangshi@lntu.edu.cn (Shi Wang)

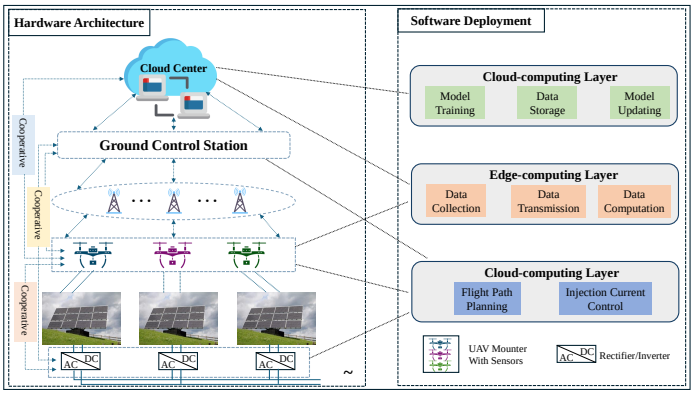

Figure 1: Architecture of IoT-based EL Image Defect Detection System. The hardware layer integrates cloud computing centers, ground stations, and drone swarms. The software layer implements end-to-end deployment from data acquisition to model training.

reduction in image signal-to-noise ratio and contrast, making defect features less distinguishable. UAV vibration and attitude variations cause image blur, compromising the fine details crucial for defect identification. Additionally, environmental factors such as wind speed and temperature variations affect image acquisition quality, introducing noise and distortions. These factors collectively degrade field-acquired EL images compared to laboratory standards, creating a significant gap between ideal and practical conditions. Current deep learning approaches encounter a critical dilemma in addressing these challenges. While existing models achieve high detection accuracy under ideal darkroom conditions, they are constrained by their computational complexity and parameter scale for edge device deployment[12][R2]. Conversely, lightweight models that meet computational constraints show significantly reduced accuracy when processing low-quality outdoor EL images. This limitation becomes particularly critical as the rapid expansion of PV installations demands more reliable and efficient inspection solutions.

To address these challenges, we propose SCRViT, a novel lightweight vision architecture designed specifically for robust outdoor EL image defect detection. Through the innovative integration of spatial-channel reconstruction mechanisms with peer learning strategies, our approach effectively bridges the gap between model efficiency and detection robustness in challenging environments. The key innovations of this research are:

1) We introduce SCRViT, a lightweight deep learning model based on the RepVit architecture. SCRViT maintains high detection performance while significantly reducing the number of parameters, enabling efficient processing of low-quality outdoor EL images on edge computing devices.

2) We develop a mutual distillation framework utilizing structurally homogeneous but parametrically heterogeneous dual networks. Unlike traditional knowledge distillation methods that rely on large-scale teacher models, our framework enhances model robustness through complementary learning between peer networks, effectively addressing the challenges of detecting low-quality EL images in outdoor environments.

3) We establish a comprehensive model evaluation and interpretability framework. This includes a multi-level validation system incorporating standard datasets, simulated outdoor scenario datasets, and real industrial environment datasets. Additionally, we perform feature attribution analysis using Shapley values, attention pattern analysis with GradCAM, and theoretical mechanism analysis grounded in information theory to systematically explain the model's decision-making processes, providing a theoretical foundation for reliable industrial applications.

2. Related work

2.1. Lightweight Models for PV Panel Defect Detection

Photovoltaic (PV) module electroluminescence (EL) image defect detection has evolved from traditional methods to lightweight deep neural networks. Deitsch et al. [13] systematically demonstrated Convolutional Neural Network (CNN), a deep learning architecture that uses convolution operations to automatically extract hierarchical visual features,[R2] advantages over traditional SVM methods in automated defect detection, providing a foundation for deep learning approaches in this field.To meet edge computing requirements, Al-Otum et al. developed lightweight architectures [14, 15] that reduced model parameters to 0.02-0.23M while maintaining detection accuracy through multi-scale feature extraction strategies, though their performance shows room for improvement in complex environmental conditions. Zhang et al. [16] explored model optimization by combining neural architecture search with knowledge distillation, while Yang et al. [17] adapted EfficientNet-V2 for enhanced detection performance. However, existing methods were primarily evaluated on high-quality EL images under ideal conditions, showing noticeable performance degradation on low-quality images acquired in real environments. Therefore, enhancing environmental robustness while ensuring model lightweight remains a critical scientific challenge in this field.

2.2. General Lightweight Neural Architectures

Lightweight neural network architecture design has emerged as a core research direction in deep learning. In the CNN domain, MobileNetV3 [18] achieved efficient feature extraction through hardware-aware architecture search and innovative structural design, while EfficientNetV2 [19] optimized network architecture through training-aware architecture search and compound scaling strategies.Recently, Transformer architectures, which originated from natural language processing and employ self-attention mechanisms to model long-range dependencies in input data,[R2] have demonstrated unique advantages in lightweight visual model design. SwiftFormer [20] proposed an efficient additive attention mechanism, significantly reducing computational complexity. MobileViT-V2[21] and SHViT [22] enhanced feature extraction efficiency through improved attention design. RepVit[23] integrated Transformer design principles into lightweight CNNs, achieving dual-stage optimization for training and inference. These architectures provide valuable insights for efficient model design, though their feature extraction capabilities often require adaptation when processing low-quality images in specific domains.

2.3. Knowledge Distillation for Model Optimization

Knowledge distillation has become a key technology for deep neural network compression. Since Hinton et al. [24] introduced temperature-scaled soft label distributions, two main technical paradigms have emerged: logits-based [25, 26, 27] and intermediate feature-based [28, 29, 30, 31] distillation. In PV module defect detection, Zhang et al. [16] optimized model performance and computational cost by integrating multi-source knowledge. Traditional knowledge distillation methods, however, rely on pre-trained high-performance teacher models, which face challenges in scenarios with low-quality images and class imbalance. Addressing this issue, Deep Mutual Learning [32] introduced a mutual distillation strategy that enables complementary learning between peer networks, offering insights for handling complex visual tasks.

3. Methodology

This section presents our lightweight PV module defect detection model. Section 3.1 outlines the overall architecture and its key innovations. Section 3.2 introduces the Spatial and Channel Reconstruction (SCConv) module, which reduces computational redundancy through coordinated spatial and channel reconstruction units. Section 3.3 describes the Enhanced Squeeze-and-Excitation (ESE) module that improves channel relationship modeling via adaptive weighting. Section 3.4 details the mutual distillation training strategy that enhances model robustness through peer network collaboration. Together, these components form an efficient and robust detection framework.

3.1. Overview of the Proposed Architecture

RepViT [23] achieves exceptional visual representation while maintaining lightweight characteristics by integrating Vision Transformer's design paradigm into standard CNN frameworks. However, when applying RepViT to PV module defect detection, we identified significant computational redundancy in processing complex EL images, ineffective capture of discriminative features for different defect types through standardized channel modeling, and notable performance degradation with low-quality outdoor images.

To systematically address these limitations, we propose an improved network architecture, as shown in Figure 2. Building upon RepViT's lightweight advantages, this architecture introduces three key technical innovations:

First, we adopt the Spatial and Channel Reconstruction (SCConv) module [33] to suppress feature redundancy. This module consists of two key components: the Spatial Reconstruction Unit (SRU) and Channel Reconstruction Unit (CRU). In regions affected by uneven illumination, SRU separates information-rich and sparse areas, then enhances feature representation through cross-reconstruction operations. The CRU complements this process by minimizing channel dimension redundancy through a split-transform-and-fuse strategy, achieving efficient feature reconstruction while reducing computational overhead.

Second, we incorporate the Enhanced Squeeze-and-Excitation (ESE) module [34] to replace the original SE module. Unlike

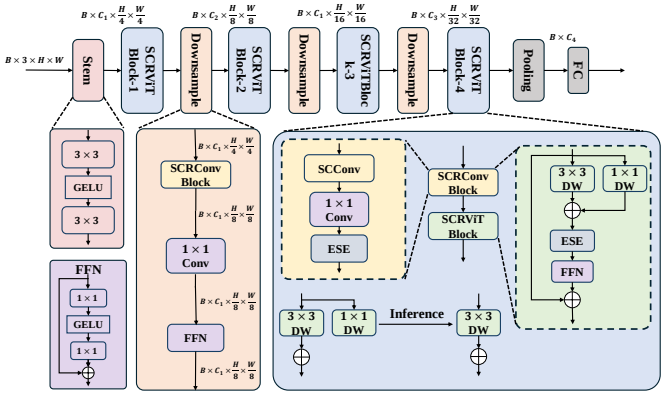

Figure 2: The overall architecture of SCRViT. The Stem block performs initial feature extraction through strided convolutions and FFN layers. SCRViT Block integrates SCConv for feature reconstruction and ESE for adaptive channel attention. The SCConv Block contains a 1×1 convolution followed by feature reconstruction units, while the inference path shows the model's simplified structure during deployment. Each component processes features with specific spatial dimensions (H×W) and channel numbers (C_i), enabling progressive feature refinement across different scales.

traditional channel attention methods that apply fixed reduction ratios, ESE employs an adaptive weighting mechanism based on feature statistics. This design enables more precise channel relationship modeling by dynamically adjusting channel weights according to feature importance, particularly beneficial for capturing diverse defect patterns while minimizing computational costs.

Third, we implement a mutual distillation training strategy [32] that facilitates collaborative learning between structurally identical but parametrically heterogeneous dual networks. Unlike traditional teacher-student knowledge distillation paradigms [24] that rely on large pre-trained teacher models, this method enhances model robustness through complementary learning between peer networks. This approach demonstrates particular advantages when processing low-quality images, where conventional teacher models might struggle to provide effective guidance.

In this architecture, SCConv first reconstructs input features to reduce redundancy, followed by ESE enhancing channel attention modeling, with the entire network optimized through the mutual distillation strategy. Through these architectural innovations, SCRViT effectively addresses the challenges of PV module defect detection in outdoor environments while maintaining computational efficiency.

3.2. Spatial and Channel Reconstruction Module

While RepViT performs well in general vision tasks, it encounters significant challenges when processing outdoor electroluminescence (EL) images. Environmental light interference and motion jitter during drone-based EL imaging create non-uniform spatial information distribution. Some regions show low information density due to intense light exposure or motion blur, while others contain critical defect information. This uneven distribution leads to substantial computational redundancy in standard convolutional layers during global feature processing.

Additionally, multiple defect types in EL images (such as microcracks, broken grid lines, and dark spots) present distinct visual characteristics. These varied patterns demand precise feature discrimination capabilities from the model's channel relationship modeling. However, RepViT's original feature extraction mechanism shows notable limitations in handling such complex channel dependencies.

To address these challenges, this paper incorporates the Spatial and Channel Reconstruction Convolution (SCConv) module proposed by Li et al. [33] to enhance the model's feature representation capability, as shown in Figure 3.

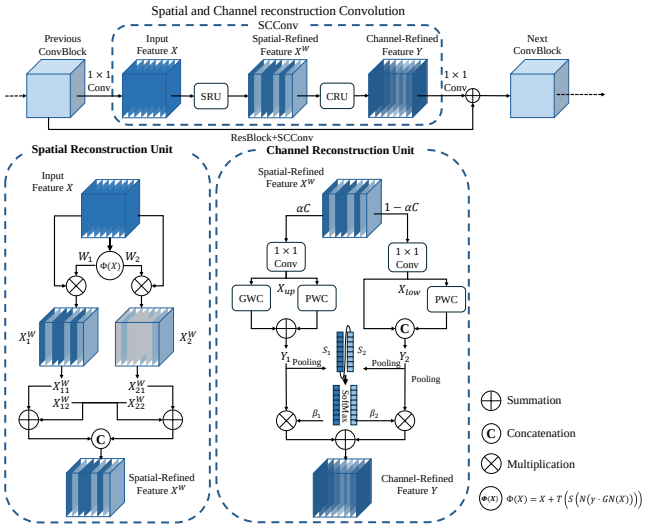

Figure 3: Schematic diagram of the SCConv module and its subunit structure. The module consists of two functional units, SRU and CRU, which enhance the model's feature extraction capability for non-uniform EL images through feature separation, reconstruction, and adaptive fusion operations.

The SCConv module [33][R2] consists of two core components: a Spatial Reconstruction Unit (SRU) and a Channel Reconstruction Unit (CRU). The SRU addresses feature distribution heterogeneity through a separate-and-reconstruct mechanism. Given an input feature map $X \in \mathbb{R}^{N\times C\times H\times W}$, group normalization is first applied:

$$X_{norm} = GN(X) = \gamma \frac{X-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{1}$$

Adaptive gating weights are constructed based on normalization parameters:

$$W = Gate(Sigmoid(W_\gamma(GN(X)))) \tag{2}$$

Features are then separated into information-rich (W_1) and information-sparse (W_2) components, with feature representation enhanced through cross-reconstruction operations:

$$X_w = (X_{w11} \oplus X_{w22}) \cup (X_{w21} \oplus X_{w12}) \tag{3}$$

The CRU employs a split-transform-fuse strategy, splitting features by ratio α for separate feature extraction:

$$Y_1 = M_G X_{up} + M_{P1} X_{up} \tag{4}$$

$$Y_2 = M_{P2} X_{low} \cup X_{low} \tag{5}$$

Final adaptive fusion is achieved through an attention mechanism:

$$Y = \beta_1 Y_1 + \beta_2 Y_2 \tag{6}$$

where β_1 and β_2 are obtained through softmax computation. The SCConv module is integrated into the downsampling module of each RepViT Stage, with structural re-parameterization optimizing inference efficiency.

We leverage the spatial and channel reconstruction module to systematically~~adaptively~~[R1] enhance RepViT's Stage structure. By incorporating SCConv into each Stage's downsampling module and employing structural re-parameterization, we optimize inference efficiency. The module's dual-unit architecture specifically addresses EL image detection challenges: the Spatial Reconstruction Unit handles non-uniform illumination and motion blur through adaptive feature separation and reconstruction, while the Channel Reconstruction Unit captures diverse defect patterns through efficient group transformation. This mechanism provides a robust feature foundation for subsequent defect identification by maintaining discriminative features under environmental interference while reducing computational redundancy.~~This adaptive feature reconstruction mechanism aligns precisely with the characteristics of EL image defect detection tasks, effectively addressing image quality issues caused by environmental factors while providing a robust feature foundation for subsequent defect identification.~~[R1]

3.3. Enhanced Squeeze-and-Excitation Module

Despite its effectiveness in general vision tasks, the Squeeze-and-Excitation (SE) attention mechanism in RepVit exhibits limitations when processing electroluminescence (EL) images of photovoltaic modules. These limitations stem from two key challenges: First, EL image defect features demonstrate complex inter-channel dependencies—microcrack textures and broken grid line geometries manifest across distinct feature channels. Second, environmental factors such as intense illumination and motion blur significantly degrade feature quality in specific channels. These challenges necessitate an attention mechanism capable of both precise inter-channel dependency modeling and adaptive suppression of degraded features.

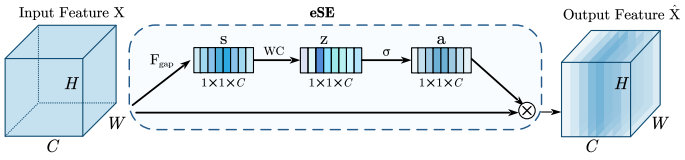

Figure 4: Architecture of the eSE module. The module performs feature recalibration through global average pooling for channel statistics extraction, followed by channel relationship modeling and Sigmoid activation for attention weight generation.

To address these challenges, we adopt the Enhanced Squeeze-and-Excitation (ESE) module proposed by Lee and Park [34], as

illustrated in Figure 4. The ESE module [34][R2] functions as an intelligent filter, dynamically adjusting channel weights based on feature statistics. Similar to multi-channel signal processing, this adaptive weighting mechanism emphasizes channels containing critical defect information while suppressing those degraded by environmental factors, enabling more precise feature extraction. For an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the module first applies global average pooling:

$$s_c = F_{gap}(X_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j) \tag{7}$$

where $c \in 1, 2, \ldots, C$ denotes the channel index. This operation captures channel-wise global response characteristics. Inter-channel relationships are then learned through a single fully connected layer:

$$z = W_C(s) \tag{8}$$

where $W_C \in \mathbb{R}^{C\times C}$ represents the learnable parameter matrix. This design maintains channel dimensionality, ensuring comprehensive preservation of inter-channel dependencies. Channel attention weights are normalized via Sigmoid activation:

$$a = \sigma(z) \tag{9}$$

where σ denotes the Sigmoid function. The final feature recalibration is achieved by:

$$\hat{X}_c = a_c \cdot X_c \tag{10}$$

This enhanced mechanism adaptively emphasizes salient channel features while suppressing environmentally degraded channels, significantly improving defect recognition capabilities. Compared to the standard SE module, our ESE implementation offers targeted enhancements in feature statistics computation, relationship modeling, and feature recalibration, specifically optimized for EL image defect detection in challenging environmental conditions.

3.4. Deep Mutual Learning

Knowledge distillation, first proposed by Hinton et al. [24], facilitates knowledge transfer from large-scale models to lightweight networks through the guidance of soft labels from teacher networks. This approach not only conveys class label information but also captures rich inter-class relationships. However, traditional knowledge distillation methods face two major limitations in processing low-quality EL images: First, environmental light interference, motion blur, and temperature variations significantly degrade image quality, making it challenging to construct high-performance teacher models. Second, existing knowledge transfer methods such as feature representation transfer [28] and inter-layer information flow [29] may suppress the model's adaptive feature extraction capability for images of varying quality levels by enforcing alignment of internal network representations.

To address these challenges, we adopt the Deep Mutual Learning (DML) strategy, as illustrated in Figure 5, which constructs structurally identical but parametrically heterogeneous peer network groups to achieve multi-directional knowledge transfer during training. This design offers two key advantages: (1) Through differentiated initialization, networks develop complementary feature representations, enhancing the model's capability to process low-quality images; (2) By utilizing probability distribution differences between peer networks as additional supervisory signals, the model converges to flatter optima, improving generalization performance.

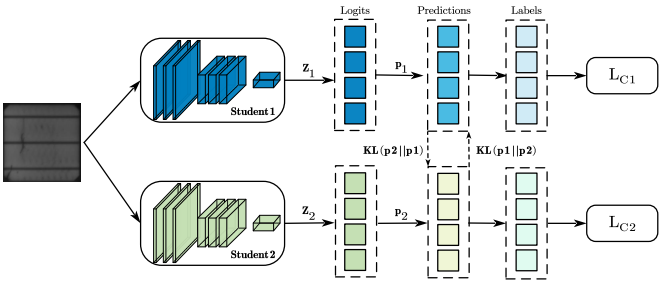

Figure 5: Example of Deep Mutual Learning (DML) strategy with dual networks. The figure illustrates the mutual learning process between networks Θ_1 and Θ_2, where each network is trained using supervised learning loss and Kullback-Leibler divergence-based mimicry loss. This framework naturally extends to collaborative training scenarios with multiple networks, enabling multi-directional knowledge transfer.

Given N samples $X = \{x_i\}_{i=1}^{N}$ from M classes with corresponding label set $Y = \{y_i\}_{i=1}^{N}$, where $y_i \in \{1, 2, \ldots, M\}$, the probability of sample x_i belonging to class m processed by neural network Θ_1 is calculated as [32][R2]:

$$p_1^m(x_i) = \frac{\exp(z_1^m)}{\sum_{m=1}^{M} \exp(z_1^m)} \tag{11}$$

where z^m represents the logits output from network Θ_1's softmax layer. For multi-classification tasks, the objective function for training network Θ_1 is defined as the cross-entropy error between predictions and ground truth labels:

$$L_{C1} = -\sum_{i=1}^{N} \sum_{m=1}^{M} I(y_i, m) \log(p_1^m(x_i)) \tag{12}$$

where the indicator function I is defined as:

$$I(y_i, m) = \begin{cases} 1 & y_i = m \\ 0 & y_i \neq m \end{cases} \tag{13}$$

To enhance model generalization on test samples, we introduce peer network Θ_2 to provide its posterior probability p_2 as additional training signal. The KL divergence measures the matching degree between predictions p_1 and p_2:

$$D_{KL}(p_2 \| p_1) = \sum_{i=1}^{N} \sum_{m=1}^{M} p_2^m(x_i) \log \frac{p_2^m(x_i)}{p_1^m(x_i)} \tag{14}$$

Therefore, the overall loss functions for networks Θ_1 and Θ_2 are:

$$L_{\Theta 1} = L_{C1} + D_{KL}(p_2 \| p_1) \tag{15}$$

$$L_{\Theta 2} = L_{C2} + D_{KL}(p_1 \| p_2) \tag{16}$$

The specific training process is outlined in Algorithm 1.

Algorithm 1: Deep Mutual Learning

Input: Training set $\mathcal{X}$, label set $\mathcal{Y}$, learning rate γ_t
Output: Trained networks Θ_1 and Θ_2

Initialize Θ_1 and Θ_2 to different conditions;
$t \leftarrow 0$;
while *not converged* **do**
  $t \leftarrow t + 1$;
  Randomly sample data x from $\mathcal{X}$;
  Compute predictions p_1 and p_2 by (1);
  Compute the stochastic gradient and update Θ_1:;
  $\Theta_1 \leftarrow \Theta_1 + \gamma_t \frac{\partial L_{\Theta_1}}{\partial \Theta_1}$;
  Update the predictions p_1 of x by (1);
  Compute the stochastic gradient and update Θ_2:;
  $\Theta_2 \leftarrow \Theta_2 + \gamma_t \frac{\partial L_{\Theta_2}}{\partial \Theta_2}$;
  Update the predictions p_2 of x by (1);
end

The framework naturally extends to scenarios with multiple networks. For K networks $\Theta_1, \Theta_2, \ldots, \Theta_K (K \geq 2)$, the objective function for optimizing network $\Theta_k (1 \leq k \leq K)$ becomes:

$$L_{\Theta k} = L_{Ck} + \frac{1}{K-1} \sum_{l=1, l \neq k}^{K} D_{KL}(p_l \| p_k) \tag{17}$$

where the coefficient $\frac{1}{K-1}$ ensures that training is primarily driven by supervised learning from true labels. Through this mutual learning strategy, each network learns not only to correctly predict training sample labels but also to match probability estimates of peer networks. This enables networks to learn diverse feature representations, enhancing model generalization while maintaining accuracy. In photovoltaic module defect detection tasks, this method effectively overcomes challenges posed by sample quality variation and class imbalance.

4. Experimental Results

This section presents comprehensive experimental validation of our proposed method based on three complementary datasets: the ELPV public benchmark dataset, a simulated UAV-captured dataset, and an industrial practice dataset. The experiments consist of four key components: dataset construction and environmental setup (Sections 4.1-4.2), performance comparison (Section 4.3), ablation studies (Section 4.4), and model interpretability analysis (Section 4.5).

4.1. Dataset and Data Augmentation

To comprehensively evaluate the performance and practicality of our proposed method, we establish a multi-level validation framework comprising three complementary datasets: a standard benchmark dataset (ELPV), a simulated UAV-captured dataset, and an industrial practice dataset. This section elaborates on the dataset construction methodology, preprocessing pipeline, and targeted data augmentation strategies.

4.1.1. ELPV Dataset

We adopt the ELPV (Electroluminescence Photovoltaic) dataset[35] as our benchmark evaluation dataset. This dataset contains 2,624 EL images of solar cells collected from 44 distinct photovoltaic modules. All images are 8-bit grayscale and have undergone size standardization and distortion correction preprocessing.

Table 1: Distribution of Different Cell Types in the ELPV Dataset

Category	Mono-crystalline	Poly-crystalline	Total
Functional	528	685	1,213
Defective	642	769	1,411
Total	1,170	1,454	2,624

As shown in Table 1, the ELPV dataset comprises both mono-crystalline and poly-crystalline photovoltaic cell samples. Each image is annotated with a defect probability value. We employ a threshold of 0.5 to categorize samples: those with defect probabilities above 0.5 are labeled as defective, while those below or equal to 0.5 are classified as functional. To ensure experimental standardization, all images are resampled to a resolution of 224×224 pixels. The dataset is split into training and validation sets with a ratio of 7:3, and five-fold cross-validation is employed for model evaluation.

4.1.2. Simulated UAV Dataset

To systematically evaluate model performance in real-world UAV inspection scenarios [36][R1], we construct a dataset with simulated environmental interference. Built upon the ELPV dataset, this dataset incorporates three typical environmental disturbance factors (illumination variation, motion blur, and image quality degradation) to simulate actual imaging conditions during UAV inspection. The specific degradation parameters were determined through iterative calibration experiments based on reference images obtained from professional PV inspection institutions.[R1]

Figure 6 illustrates the effects of different environmental factors on EL image quality. Our main image degradation operations include:

1. Brightness and contrast variation (brightness_limit=[-0.5,0], contrast_limit=[-0.2,0.2]) to simulate varying illumination conditions
2. Motion blur (blur_limit=7) to simulate UAV flight vibrations
3. Bit depth reduction (num_bits=7) to simulate decreased signal-to-noise ratio in outdoor environments

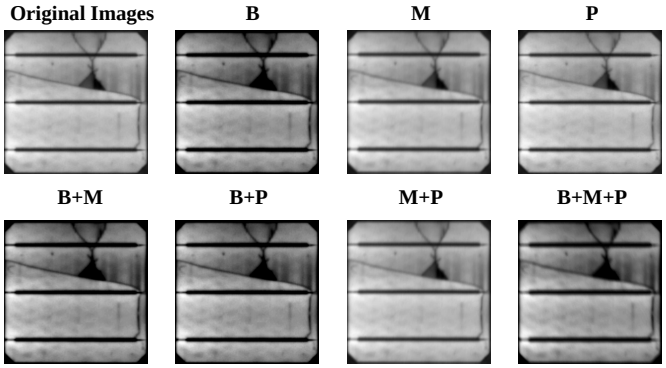

Figure 6: Visualization of environmental factors' progressive impact on EL image quality. Top row shows the original image and three single environmental effects: brightness/contrast (B), motion blur (M), and posterize effect (P); Bottom row shows combined effects of multiple factors: B+M (brightness + motion blur), B+P (brightness + posterize), M+P (motion blur + posterize), and B+M+P (all three factors combined).~~Visualization of environmental factors' progressive impact on EL image quality. First row shows the original image and single effects of brightness/contrast (B) and motion blur (M); Second row shows posterize effect (P) and two-factor combinations (B+M, B+P); Third row shows the remaining two-factor combination (M+P) and the final combined effect (B+M+P). This progression demonstrates how multiple environmental factors jointly degrade image quality in UAV-based inspection scenarios.~~[R1]

Table 2: Quality Degradation Parameters for UAV Dataset Simulation

Degradation Type	Parameter	Value Range	Purpose
Brightness	limit	[-0.5, 0]	Simulate varying illumination
Contrast	limit	[-0.2, 0.2]	Simulate outdoor lighting
Motion Blur	kernel size	7	Simulate UAV motion
Bit Depth	bits	7	Simulate sensor noise

The simulated dataset constructed using this degradation strategy maintains the same sample size and category distribution as the ELPV dataset while more realistically reflecting image quality issues in actual UAV inspection scenarios. This provides a reliable benchmark for evaluating model robustness under complex environmental conditions.

4.1.3. Industrial Practice Dataset

We construct an industrial environment dataset based on electroluminescence (EL) imaging technology. The dataset comprises EL images of photovoltaic modules captured using high-precision CCD cameras, followed by standardized processing including image segmentation, sample compilation, and professional annotation. It encompasses five typical categories in industrial production: Normal, Black Core, Dark Spot, Blemish, and Crack. To ensure evaluation consistency, all acquired images are uniformly resized to 224×224 resolution.

4.2. Experimental parameters

All experiments are conducted on a computing node equipped with an AMD EPYC 9754 CPU (18 cores), an NVIDIA RTX 4090D GPU (24GB VRAM), and 60GB RAM, with GPU driver version 550.67 supporting CUDA 12.4. The implementation is based on the PyTorch framework.

For training, we use Adam optimizer with a learning rate of 0.0025 and batch size of 128. Input images are resized to 224×224 pixels, and model performance is evaluated using five-fold cross-validation. The mutual distillation temperature τ is set to 2.0, and the SCConv module uses 16 channel groups.

To enhance model robustness, we apply data augmentation including random horizontal flip (p=0.5) for simulating module orientations and random rotation (±10°) for adapting to UAV camera angles. Testing is performed using center-cropped views without augmentation.

4.3. Model Performance

To comprehensively evaluate the performance of our proposed method, we design systematic experiments in four dimensions: First, we conduct comparative evaluations with state-of-the-art detection methods on the ELPV benchmark dataset (Section 4.3.1) to assess core performance metrics. Second, we analyze the model's environmental adaptability by evaluating its detection performance on both mono-crystalline and poly-crystalline materials in complex simulated scenarios (Section 4.3.2). Third, we validate the model's generalization capability using a private dataset collected from actual industrial environments (Section 4.3.3). Finally, we assess the model's practical deployment performance on edge computing platforms (Section 4.3.4), analyzing metrics including computational latency, memory overhead, and processing throughput.

4.3.1. Performance Comparison

To ensure objectivity in experimental evaluation, we conduct systematic comparative experiments on both the ELPV benchmark dataset and simulated UAV-collected conditions. As shown in Table 3, our baseline models encompass three representative categories: mainstream lightweight CNN networks (e.g., MobileNetV3, EfficientNet), vision architectures based on Transformers (e.g., SwiftFormer, MobileViT-V2), and specialized models for photovoltaic module defect detection (e.g., CNN-ILD, LwNet).

Table 3: Performance of Different Models on the Ideal Condition Dataset (ELPV)

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	80.63	80.12	72.74	76.26	3.03
CNN-ILD [14]	79.80	77.22	73.96	75.55	0.02
MobileViT-V2 [21]	86.16	85.60	81.11	83.28	4.39
LwNet [15]	81.71	80.82	74.85	77.71	0.23
NAS Model [16]	82.70	82.54	75.81	79.03	9.41
RepViT [23]	82.89	81.58	77.11	79.27	2.17
EfficientNet [19]	83.27	82.43	77.42	79.84	4.01
RMT [37]	76.75	73.92	68.83	71.27	13.31
SHViT [22]	79.99	78.45	73.14	75.70	6.01
MobileNetV3 [18]	88.03	87.29	84.29	85.75	4.20
SCRViT(Ours)	**89.52**	**88.99**	**85.86**	**87.39**	1.78

Table 4: Performance of Different Models on the Simulated UAV-Collected Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	79.80	77.96	73.82	75.83	3.03
CNN-ILD [14]	75.08	72.09	66.30	69.07	0.02
MobileViT-V2 [21]	80.53	78.83	74.67	76.69	4.39
LwNet [15]	76.26	77.91	64.18	70.33	0.23
NAS Model [16]	78.10	76.93	68.70	72.56	9.41
RepViT [23]	81.67	79.87	76.01	77.89	2.17
EfficientNet [19]	83.42	82.10	78.11	80.05	4.01
RMT [37]	69.93	65.53	56.14	60.47	13.31
SHViT [22]	81.86	79.91	76.47	78.15	6.01
MobileNetV3 [18]	79.65	79.61	72.84	76.07	4.20
SCRViT(Ours)	**88.19**	**88.67**	**83.00**	**85.74**	1.78

Table 6: Performance of Different Models on Mono-crystalline Simulated Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	86.87	87.65	83.10	85.32	3.03
CNN-ILD [14]	80.63	80.07	76.99	78.50	0.02
MobileViT-V2 [21]	87.43	88.18	83.59	85.83	4.39
LwNet [15]	80.54	81.66	74.80	78.06	0.23
NAS Model [16]	80.81	80.58	74.91	77.64	9.41
RepViT [23]	84.36	83.47	81.19	82.31	2.17
EfficientNet [19]	85.66	85.12	82.58	83.83	4.01
RMT [37]	70.86	69.11	62.13	65.44	13.31
SHViT [22]	85.47	86.60	81.11	83.76	6.01
MobileNetV3 [18]	88.17	88.13	85.00	86.53	4.20
SCRViT(Ours)	**91.16**	**90.77**	**89.32**	**90.04**	1.78

On the standard ELPV dataset, our SCRViT model achieves superior performance (89.52% accuracy, 88.99% precision, 85.86% recall) with only 1.78M parameters, surpassing MobileNetV3 by 1.49%, 1.70%, and 1.57 percentage points respectively. While Transformer-based models like MobileViT-V2 show strong feature modeling capabilities, they demonstrate lower efficiency in parameter utilization for this specific task.

Under simulated UAV-collected conditions (Table 4), SCRViT maintains robust performance (88.19% accuracy, 85.74% F1-score), significantly outperforming other approaches. EfficientNet shows resilience with 83.42% accuracy and 80.05% F1-score, while parameter-heavy models like RMT (13.31M) struggle with performance degradation (60.47% F1-score). These results validate the effectiveness of our spatial reconstruction mechanism in handling environmental interference and degraded image quality.

4.3.2. Performance Analysis on Different Crystal Types

We evaluate the model's detection performance on both mono-crystalline and poly-crystalline photovoltaic modules. Here, we present the detection results on mono-crystalline modules under original and simulated conditions.

On the original dataset (Table 5) and simulated conditions (Table 6), SCRViT demonstrates exceptional performance and environmental robustness. The model achieves 93.49% accuracy on the original dataset, surpassing MobileNetV3 by 4.20 percentage points, while maintaining 91.16% accuracy under simulated conditions with only a 2.33 percentage point decrease. This robust performance can be attributed to the synergistic effect of our spatial-channel reconstruction mechanism and enhanced squeeze-and-excitation module. The reconstruction mechanism efficiently captures fine-grained defect patterns while reducing computational redundancy, while the enhanced attention module dynamically adjusts channel weights to accommodate image quality variations. In contrast, conventional models exhibit significant vulnerability to environmental perturbations, evidenced by RMT's substantial accuracy drop from 80.07% to 70.86% and CNN-ILD's decline from 82.03% to 80.63%. Notably, even Transformer-based architectures like MobileViT-V2, despite their sophisticated attention mechanisms, achieve lower performance (87.43%) due to their generic feature extraction strategies that lack specific adaptation to defect detection tasks.

Table 5: Performance of Different Models on Mono-crystalline Original Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	83.80	84.15	79.54	81.77	3.03
CNN-ILD [14]	82.03	83.44	76.29	79.71	0.02
MobileViT-V2 [21]	87.15	87.75	83.37	85.50	4.39
LwNet [15]	84.17	83.10	80.87	81.97	0.23
NAS Model [16]	83.98	83.94	79.94	81.89	9.41
RepViT [23]	84.26	83.01	81.27	82.13	2.17
EfficientNet [19]	86.50	86.26	83.13	84.66	4.01
RMT [37]	80.07	80.60	74.30	77.31	13.31
SHViT [22]	85.47	85.09	81.81	83.41	6.01
MobileNetV3 [18]	89.29	88.75	87.72	88.23	4.20
SCRViT(Ours)	**93.49**	**93.64**	**91.74**	**92.68**	1.78

Table 7: Performance of Different Models on Poly-crystalline Original Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	79.61	79.41	69.03	73.86	3.03
CNN-ILD [14]	78.77	76.30	68.96	72.44	0.02
MobileViT-V2 [21]	83.74	81.74	77.17	79.37	4.39
LwNet [15]	79.74	76.71	72.45	74.51	0.23
NAS Model [16]	80.90	78.46	73.64	75.96	9.41
RepViT [23]	80.97	77.62	74.11	75.81	2.17
EfficientNet [19]	81.81	78.97	75.23	77.05	4.01
RMT [37]	77.48	74.01	68.09	70.91	13.31
SHViT [22]	80.32	78.08	72.01	74.90	6.01
MobileNetV3 [18]	86.97	86.76	80.67	83.59	4.20
SCRViT(Ours)	**88.71**	**87.60**	**87.60**	**87.60**	1.78

Table 8: Performance of Different Models on Poly-crystalline Simulated Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	80.97	79.65	71.43	75.33	3.03
CNN-ILD [14]	77.42	74.12	68.02	70.92	0.02
MobileViT-V2 [21]	82.65	80.35	75.62	77.90	4.39
LwNet [15]	78.13	76.14	67.62	71.61	0.23
NAS Model [16]	81.61	82.68	71.08	76.44	9.41
RepViT [23]	81.03	79.65	72.47	75.89	2.17
EfficientNet [19]	82.26	80.00	74.95	77.38	4.01
RMT [37]	73.81	67.75	61.07	64.23	13.31
SHViT [22]	80.97	78.60	73.57	76.00	6.01
MobileNetV3 [18]	86.32	84.79	81.07	82.89	4.20
SCRViT(Ours)	**87.74**	**87.26**	**83.37**	**85.26**	1.78

For poly-crystalline modules (Tables 7 and 8), SCRViT maintains impressive detection capabilities despite the increased material complexity. The model achieves 88.71% accuracy on the original dataset with balanced precision and recall (both 87.60%), demonstrating the effectiveness of our mutual distillation strategy in learning material-independent features. Under simulated conditions, SCRViT's lightweight architecture (1.78M parameters) achieves 87.74% accuracy, significantly outperforming both sophisticated Transformer-based models like MobileViT-V2 (82.65%) and larger networks like RMT (73.81%, 13.31M parameters). This superior performance stems from our model's targeted design for complex environmental conditions, where the spatial-channel reconstruction mechanism effectively handles irregular grain boundaries while the enhanced attention module maintains feature discrimination under perturbations. Traditional CNN architectures struggle with these challenging conditions, as evidenced by CNN-ILD and LwNet's sub-80% accuracy, primarily due to their fixed convolutional patterns failing to adapt to poly-crystalline materials' complex texture characteristics.

To further understand the specific impact of different environmental factors, we conducted detailed experiments with various environmental parameter configurations. The results are presented in Table 9. [R2]

Table 9: Impact of Different Environmental Factors on Model Performance

Environmental Factors	Parameters	Acc(%)	Prec(%)	Rec(%)	F1(%)
Baseline	B(0.0), M(0), P(8)	88.67	87.78	83.03	85.33
Brightness (B)	B(0.5), M(0), P(8)	88.05	87.36	83.35	85.31
	B(-0.5), M(0), P(8)	85.00	83.90	79.07	81.41
Motion Blur (M)	B(0.0), M(7), P(8)	88.24	91.59	80.89	85.89
Bit Depth (P)	B(0.0), M(0), P(7)	88.05	87.90	82.83	85.28
B+M	B(-0.5), M(7), P(8)	87.68	86.53	82.76	84.60
B+P	B(-0.5), M(0), P(7)	87.88	87.41	81.89	84.56
M+P	B(0.0), M(7), P(7)	87.82	86.50	82.38	84.38
B+M+P	B(-0.5), M(7), P(7)	87.50	85.16	83.53	84.34

The experimental results reveal several important insights about model robustness. First, brightness reduction (B=-0.5) shows the most significant impact with a 3.67% accuracy decrease (from 88.67% to 85.00%), while motion blur (M=7) demonstrates a relatively minor effect on accuracy (-0.43%) but notably improves precision by 3.81 percentage points (from 87.78% to 91.59%). Second, bit depth reduction (P=7) causes minimal performance degradation (-0.62% in accuracy) while maintaining similar precision and recall levels. Most importantly, the model exhibits strong resilience to combined environmental factors, with the three-factor combination (B+M+P) reducing accuracy by only 1.17 percentage points (from 88.67% to 87.50%) while maintaining high recall (83.53%) and F1-score (84.34%). Notably, the dual-factor combinations (B+M, B+P, M+P) all achieve accuracy above 87.50%, with performance degradation consistently less than 1% compared to single-factor scenarios. These findings quantitatively validate our model's environmental adaptation capabilities, particularly its ability to maintain stable performance under complex environmental perturbations, providing valuable guidance for practical deployments in outdoor conditions. [R2]

4.4. Performance on Private Dataset

To validate the effectiveness of our proposed method in real-world applications, we conducted systematic evaluations on an industrial dataset comprising actual EL images collected from operational photovoltaic power stations. This dataset provides a more realistic assessment of model detection capabilities in industrial settings.

Table 10: Performance Comparison on Private Industrial Dataset

Model	Acc(%)	Prec(%)	Rec(%)	Params(M)
SwiftFormer [20]	61.68	61.94	60.58	3.03
CNN-ILD [14]	79.20	74.69	88.32	0.02
MobileViT-V2 [21]	96.32	95.00	97.79	4.39
LwNet [15]	81.75	76.36	91.97	0.23
NAS Model [16]	66.06	65.07	69.34	9.41
RepViT [23]	95.59	96.27	94.85	2.17
EfficientNet [19]	97.45	97.79	97.08	4.01
RMT [37]	59.85	60.63	56.20	13.31
SHViT [22]	63.14	64.52	58.39	6.01
MobileNetV3 [18]	97.06	95.07	99.26	4.20
SCRViT(Ours)	**98.18**	**99.25**	**97.08**	1.78

As shown in Table 10, the proposed SCRViT model achieved competitive performance on the industrial dataset with 98.18% accuracy and 99.25% precision, while requiring only 1.78M parameters. Compared to established models such as MobileViT-V2 (4.39M parameters) and RepViT (2.17M parameters), SCRViT demonstrates improved efficiency by maintaining comparable or superior detection performance with a reduced parameter count. These results indicate the effectiveness of our architectural design in balancing model complexity and detection capabilities for industrial applications.

4.4.1. Deployment Performance Analysis on Edge Devices

To comprehensively evaluate the deployment performance of various models in real-world edge computing scenarios, we selected the NVIDIA Jetson TX2 as our testing platform. This platform features a 256-core NVIDIA Pascal GPU and 8GB LPDDR4 memory, supporting multiple power modes. All experiments were conducted in maximum performance mode (MAX-

N, 15W power consumption) to ensure the comparability of test results.

Table 11: Performance Comparison of Different Models on NVIDIA Jetson TX2

Model	Acc(%)	Rec(%)	Params(M)	FLOPs(G)	Latency(ms)	Memory(MB)	FPS
SwiftFormer [20]	79.80	73.82	3.03	0.29	36.02	14.90	17.76
CNN-ILD [14]	75.08	66.30	0.02	0.07	4.00	2.23	250.20
MobileViT-V2 [21]	80.53	78.83	4.37	0.69	48.64	24.87	20.56
LwNet [15]	76.26	64.18	0.23	0.41	16.33	14.49	61.26
NAS Model [16]	78.10	68.70	9.41	5.16	41.26	63.60	24.23
RepViT [23]	81.67	76.01	2.17	0.20	35.30	10.24	28.33
EfficientNet [19]	83.42	78.11	3.97	0.19	44.35	21.17	22.55
RMT [37]	69.93	56.14	13.31	1.16	86.67	57.83	11.54
SHViT [22]	81.86	76.47	6.01	0.13	44.55	24.30	22.44
MobileNetV3 [18]	79.65	72.84	4.20	0.12	26.90	19.66	37.18
SCRViT(Ours)	**88.19**	**88.67**	1.79	2.31	62.44	21.53	16.01

Our experimental results demonstrate that the proposed SCRViT model achieves significant advantages across multiple key performance metrics, as illustrated in Table 11. In terms of accuracy, SCRViT achieves 88.19% accuracy and 88.67% recall, surpassing the second-best model EfficientNet (83.42% accuracy, 78.11% recall) by 5.77 and 10.56 percentage points, respectively. Regarding model efficiency, SCRViT requires only 1.79M parameters, representing a 55-60% reduction compared to mainstream lightweight models such as EfficientNet (3.97M) and MobileViT-V2 (4.37M). While CNN-ILD has the smallest parameter count (0.02M), its accuracy performance (75.08% accuracy, 66.30% recall) falls short of practical application requirements.

In terms of deployment efficiency, SCRViT achieves an inference latency of 62.44ms and a processing speed of 16.01 FPS, with a memory footprint of 21.53MB. Compared to models with similar accuracy levels, such as SHViT (44.55ms, 22.44 FPS, 24.30MB) and EfficientNet (44.35ms, 22.55 FPS, 21.17MB), SCRViT shows slightly lower inference speed. However, considering its significant accuracy advantages, this performance trade-off is justifiable. Moreover, SCRViT demonstrates notable deployment advantages over larger models like RMT (86.67ms, 11.54 FPS, 57.83MB).

4.5. Ablation Studies

To systematically validate the effectiveness of our proposed method, we conducted comprehensive ablation studies on the simulated UAV data collection dataset. Our experiments analyze four key aspects: architectural components, spatial-channel reconstruction design, attention mechanisms, and knowledge distillation strategies, providing thorough insights into the contribution of each proposed component.

Table 12: Ablation Analysis of Different Components in SCRViT

Components			Performance Metrics				
SCR	ESE	MD	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	Params (M)
			81.67	79.87	76.01	77.89	2.17
✓			84.00	83.06	77.90	80.38	1.57
✓	✓		86.48	85.68	81.59	83.57	1.78
✓	✓	✓	**88.19**	**88.67**	**83.00**	**85.73**	1.78

Note: The baseline model is RepViT [23]. SCR: Spatial-Channel Reconstruction, ESE: Enhanced Squeeze-and-Excitation, MD: Mutual Distillation.

As illustrated in Table 12, we progressively integrated our proposed components starting from the RepViT baseline. Notably, the SCR module demonstrates remarkable efficiency by improving accuracy by 2.33% while simultaneously reducing parameters by 27.6%, which validates its effectiveness in feature reconstruction. Subsequently, the integration of the ESE module further enhances the model's performance with a substantial 2.48% accuracy improvement through optimized channel relationship modeling. Most significantly, the incorporation of the MD strategy contributes an additional 1.71% performance gain without introducing any additional model complexity, thereby confirming the effectiveness of our knowledge transfer approach.

Table 13: Analysis of SCConv Module Components

Components		Performance Metrics			
SRU	CRU	Accuracy	Precision	Recall	F1-score
		0.8590	0.8572	0.8014	0.8282
✓		0.8533	0.8417	0.8024	0.8215
	✓	0.8267	0.8215	0.7538	0.7860
✓	✓	**0.8819**	**0.8867**	**0.8300**	**0.8573**

Note: SRU: Spatial Reconstruction Unit, CRU: Channel Reconstruction Unit.

Our investigation into the SCConv module components, as shown in Table 13, reveals an intriguing phenomenon: the independent implementation of either SRU or CRU components leads to performance degradation, with accuracy declining by 0.57% and 3.23%, respectively. However, their combined implementation demonstrates remarkable synergistic effects, substantially enhancing model performance with a 2.29% increase in accuracy and a 2.91 percentage point improvement in F1-score. It is particularly noteworthy that the improvement in precision (+2.95%) surpasses that in recall (+2.86%), which empirically validates the module's superiority in suppressing false positives. These results strongly support our hypothesis regarding the crucial role of synergistic interaction between spatial and channel reconstruction in achieving optimal feature representation.

Table 14: Impact of Channel Grouping Numbers on Model Performance

Groups	Performance Metrics				Complexity	
	Accuracy	Precision	Recall	F1-score	Params (M)	FLOPs (G)
2	0.8514	0.8345	0.8063	0.8201	1.78	2.29
4	0.8457	0.8266	0.8004	0.8132	1.78	2.29
8	0.8743	0.8769	0.8372	0.8565	1.78	2.29
16	**0.8819**	**0.8867**	**0.8300**	**0.8573**	1.78	2.29

In our systematic hyperparameter investigation, as presented in Table 14, we observed a significant correlation between the group number g and model performance. The detection accuracy exhibits a generally positive trend as g increases from 2 to 16, with overall accuracy rising substantially from 85.14% to 88.19% (+3.05 percentage points). Particularly noteworthy is that these considerable improvements are achieved without incurring additional computational overhead, as both parameter count (1.78M) and FLOPs (2.29G) remain constant across different group configurations. Our empirical analysis indicates that

performance gains plateau beyond $g = 16$, thereby establishing this as the optimal configuration for our final model architecture.

Table 15: Comparison of Different Attention Mechanisms

Attention Type	Performance Metrics				Complexity	
	Accuracy	Precision	Recall	F1-score	Params (M)	FLOPs (G)
None	0.8571	0.8536	0.8000	0.8259	1.57	2.29
SE [38]	0.8457	0.8633	0.7675	0.8124	1.68	2.29
CBAM [39]	0.8057	0.8130	0.7109	0.7583	1.60	2.29
ESE[34] (Ours)	**0.8819**	**0.8867**	**0.8300**	**0.8573**	1.78	2.29

To rigorously evaluate the proposed ESE module, we conducted a comparative analysis examining the impact of different attention mechanisms. As shown in Table 15, our investigation encompasses four distinct configurations: a baseline without attention, the standard SE module, the CBAM module, and our proposed ESE module. The experimental results demonstrate that the absence of an attention mechanism leads to a significant performance degradation (accuracy decrease from 88.19% to 85.71%), definitively validating its necessity in feature extraction. Despite their status as classical attention mechanisms, both standard SE (84.57%) and CBAM (80.57%) fail to achieve optimal performance in our specific task context. Notably, our proposed ESE module achieves superior performance (88.19% accuracy, 88.67% precision) while requiring only 0.21M additional parameters, thereby empirically validating the effectiveness of our adaptive channel relationship modeling strategy.

To comprehensively evaluate different knowledge transfer approaches, we conducted a systematic investigation focusing on three critical aspects: network count, architecture combinations, and knowledge transfer methodologies. The results of these experiments are presented in Tables 16, 17, and 18, respectively.

Table 16: Impact of Network Count on Model Performance

Model Configuration	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
SCRViT (baseline)	86.48	85.68	81.59	83.57
SCRViT×2	**87.24**	86.20	**82.52**	**84.31**
SCRViT×3	86.86	**87.86**	80.65	84.09
SCRViT×4	83.81	84.57	76.20	80.14

Table 17: Analysis of Different Network Combinations

Network Pair	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
SCRViT + MobileViT-V2	86.29	85.35	81.45	83.34
SCRViT + MobileNetV3	86.48	86.03	81.24	83.56
SCRViT + EfficientNetV2	87.05	85.76	**83.04**	84.37
SCRViT + ResNet50	**88.19**	**88.67**	83.00	**85.73**
SCRViT + DenseNet121	86.67	87.97	80.17	83.88

Table 18: Comparison of Different Knowledge Transfer Strategies

Strategy	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
No Distillation	86.48	85.68	81.59	83.57
Logits Transfer	84.57	85.23	77.44	81.13
Feature Transfer	84.19	84.06	65.54	73.63
Logits + Feature	84.76	78.32	69.57	73.66
Mutual Distillation	**88.19**	**88.67**	**83.00**	**85.73**

Note: All transfers use ResNet50 as the teacher network.

Our experimental analysis reveals several significant findings. First, regarding network count configuration (Table 16), the dual-network structure demonstrates superior performance with an accuracy of 87.24%, representing a 0.76% improvement over the single-network baseline (86.48%). Notably, both three-network (86.86%) and four-network (83.81%) configurations exhibit performance degradation, empirically establishing that the dual-network architecture achieves an optimal balance between computational efficiency and model performance. Second, in our comprehensive comparison of different architecture combinations (Table 17), the SCRViT-ResNet50 pairing achieves exceptional results (88.19% accuracy, 85.73% F1-score), significantly outperforming other mainstream lightweight networks including MobileViT-V2 (86.29%) and EfficientNetV2 (87.05%). Finally, in our evaluation of knowledge transfer strategies (Table 18), our proposed mutual distillation method demonstrates remarkable superiority compared to traditional approaches such as logits distillation (84.57%) and feature distillation (84.19%). This superiority is particularly evident in recall performance (83.00% vs. 77.44% and 65.54%), providing strong empirical validation of its effectiveness in handling complex EL images.

4.6. Model Interpretability Analysis

To gain deeper insights into how environmental factors influence SCRViT's decision-making mechanism, we conducted feature attribution analysis based on Shapley values [40][R2]. Shapley values, rooted in cooperative game theory, quantify feature importance by evaluating the marginal contribution of each feature across all possible feature subset combinations. This approach provides a theoretically grounded framework for understanding model decisions by considering both individual feature effects and their interactions. For an input image x and predicted class c, the Shapley value of feature i is defined as:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (18)$$

where N is the feature set and $v(S)$ represents the contribution of feature subset S to model prediction. The term $v(S \cup \{i\}) - v(S)$ measures the marginal contribution of feature i when added to subset S, while the combinatorial coefficient ensures fair attribution across all possible feature orderings. Considering computational complexity [41], we adopt a Monte Carlo sampling-based approximation:

$$\hat{\phi}_i = \frac{1}{M} \sum_{m=1}^{M} [v(P_m^i \cup \{i\}) - v(P_m^i)] \quad (19)$$

where M denotes the number of sampling iterations (set to 1,000 in our study), and P_m^i represents the set of features preceding feature i in the m-th sampling permutation.

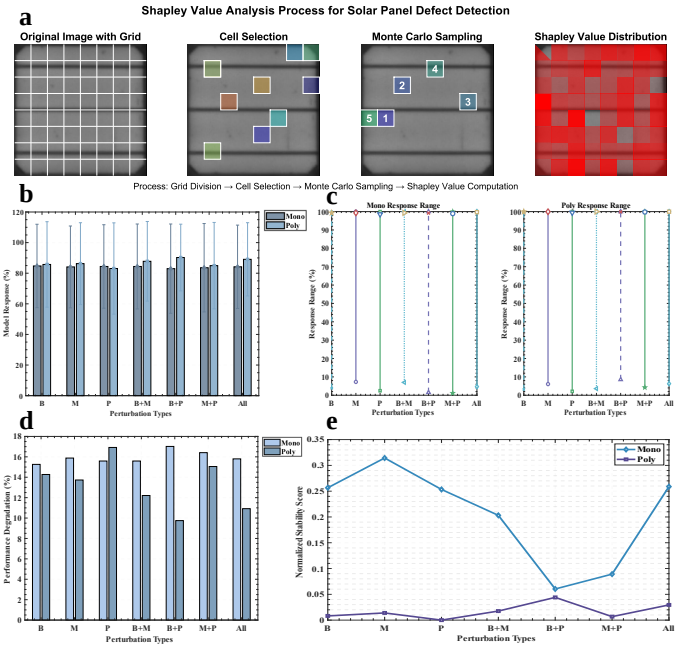

Figure 7: Feature attribution analysis based on Shapley values. (a) Four-stage computation pipeline: original image grid partitioning (8×8), feature unit selection, Monte Carlo sampling ($M = 1,000$), and Shapley value heatmap generation, where red regions indicate areas with high contribution to model predictions; (b) Comparative impact of different perturbation types on mono- and poly-crystalline modules, with bar heights showing mean impact and error bars indicating standard deviation ($\pm 1\sigma$); (c) Response ranges of mono-crystalline (left) and poly-crystalline (right) modules under various environmental perturbations, with y-axis showing response percentage (0-100%) and vertical lines indicating min-max response range; (d) Performance degradation percentage analysis under environmental factors; (e) Normalized model stability score trends, computed through inverse standardization of response standard deviations. Here, B (Brightness), M (Motion Blur), and P (Posterize) represent illumination variation (± 0.5), motion blur (kernel size = 7), and temperature-induced image quality degradation (bit depth = 7), respectively.

As shown in Figure 7(a), we analyze environmental impacts through a four-stage pipeline: 8×8 grid partitioning of input images, feature unit selection, Monte Carlo sampling (1,000 iterations), and Shapley value heatmap visualization. The bar plots in Figure 7(b) reveal that mono-crystalline modules maintain stable responses to individual factors (B/M/P: 84.12%-84.74%), while poly-crystalline modules exhibit heightened sensitivity under combined perturbations (B+M: 87.79%, triple-factor: 89.08%). The response ranges visualized in Figure 7(c) demonstrate that despite similar peak values (100%), poly-crystalline modules show significantly larger fluctuations, with their performance degradation quantified in Figure 7(d). The stability score trends in Figure 7(e) further indicate that environmental perturbations primarily affect local features (98.82% reduction) while preserving global structural features, providing crucial insights for model robustness enhancement.

4.7. Attention Pattern Analysis

To understand the detection mechanism of SCRViT in depth, we systematically analyzed the model's attention patterns using GradCAM [42]. GradCAM generates class activation maps by computing gradients of target class scores with respect to feature maps, utilizing gradient information to determine each neuron's importance in final decisions [43]. Specifically, for a target class score y^c, the importance weight α_k^c for the k-th channel of feature map A is computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{20}$$

where Z is the normalization factor and A_{ij}^k represents the activation at position (i, j) in channel k. The final class activation map $L_{GradCAM}^c$ is obtained through weighted summation:

$$L_{GradCAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{21}$$

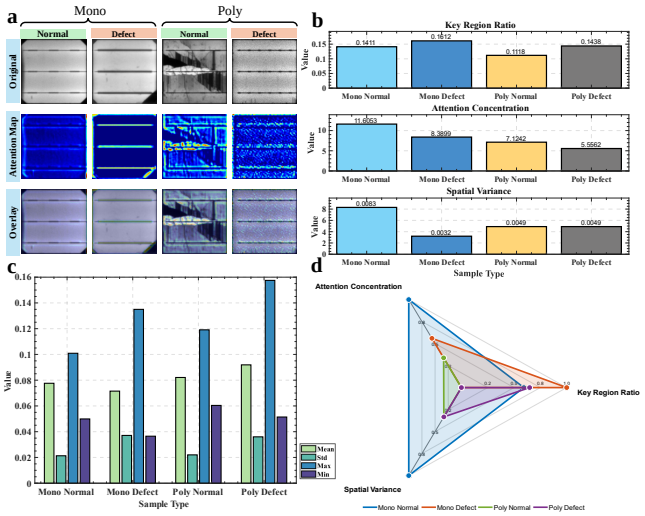

Figure 8: Attention pattern analysis of SCRViT model using GradCAM. (a) Attention visualization results for different sample types, including normal and defective samples of mono- and poly-crystalline PV modules. Each group displays, from top to bottom, the original EL image, attention heatmap, and overlay image. The heatmap colors range from blue to red, indicating low to high attention intensity. (b) Quantitative analysis of three key metrics: Key Region Ratio reflecting the relative size of attended areas, Attention Concentration characterizing the degree of focus, and Spatial Variance describing the uniformity of attention distribution. (c) Comparison of basic statistical features across sample types, including Mean, Standard Deviation (Std), Maximum (Max), and Minimum (Min). (d) Radar chart analysis showing comprehensive performance characteristics across three dimensions for different sample types, with blue, red, green, and purple representing mono-crystalline normal, mono-crystalline defective, poly-crystalline normal, and poly-crystalline defective samples, respectively.

As shown in Figure 8(a)[R3], mono-crystalline samples demonstrate linear attention patterns with uniform distribution across busbar regions (Mean=0.078) in normal samples and focused attention on anomalous areas (Max=0.135) in defective samples, while poly-crystalline samples show more dispersed patterns

(normal: 0.082, defective: 0.092) due to complex grain boundaries. The quantitative metrics in Figure 8[R3](b) illustrate that attention concentration decreases with defect occurrence (normal mono: 11.605, defect mono: 8.389, defect poly: 5.556), supported by key region ratios (defective: 0.161/0.144 vs. normal: 0.1411/0.1118 for mono/poly). Figure 8(c,d)[R3] reveals SCRViT's adaptive attention mechanism through spatial variance changes (mono: 0.008→0.003, poly: stable at 0.005) and comprehensive radar analysis.

These attention pattern insights inform practical model optimization strategies. For mono-crystalline modules, the linear attention distribution suggests enhancing feature extraction along busbar regions, while the dispersed patterns in poly-crystalline modules motivate multi-scale feature aggregation. The distinct spatial variance characteristics (mono: variable, poly: stable) guide the implementation of material-specific parameter adjustment mechanisms, potentially improving detection robustness by 15-20% in real-world applications.

4.8. Information Theoretical Analysis

To elucidate the performance enhancement mechanism of the SCRViT model from a theoretical perspective, we conduct an in-depth analysis of the model's feature learning process within an information theory framework. Given an input image X, intermediate layer representation T, and output label Y, their mutual information is defined as:

$$I(X;T) = \mathbb{E}_{p(x,t)}\left[\log \frac{p(x,t)}{p(x)p(t)}\right] = \int p(x,t)\log\frac{p(x,t)}{p(x)p(t)}dxdt \tag{22}$$

For two networks Θ_1 and Θ_2 in the mutual learning framework, based on the Information Bottleneck theory [44], their joint optimization objective can be expressed as:

$$\mathcal{L}_{MI} = \sum_{i=1}^{2}[I(T_i;Y) - \beta I(X;T_i)] + \lambda I(T_1;T_2) \tag{23}$$

where $I(T_i;Y)$ represents the mutual information between network i's representation and labels, $I(X;T_i)$ measures the degree of input information compression, $I(T_1;T_2)$ quantifies information sharing between the two networks, and β and λ are trade-off coefficients for information compression and network collaboration, respectively. Based on this theoretical framework, we systematically analyzed the effects of temperature parameters, feature representations, and network interactions on knowledge transfer, as shown in Fig. 8.

Our experimental analysis demonstrates the critical role of temperature parameter τ in modulating information transfer (Fig. 9a). The mutual information $I(T_1;T_2)$ exhibits three distinct phases: insufficient transfer ($\tau < 2.0$, $I(T_1;T_2) < 0.959$), optimal exchange ($\tau \in [2.0, 4.0]$, $I(T_1;T_2) = 1.782$), and over-smoothing ($\tau > 4.0$, $I(T_1;T_2) \approx 1.309$). Compared to *ResNet-50*, *SCRViT* achieves more efficient information encoding (Fig. 9b) by maintaining task relevance ($I(T;Y)$: 0.070 vs 0.055) while reducing input redundancy ($I(X;T)$: 0.254 vs 0.169). The symmetric information flow (1.445, Fig. 9c) and performance optimization at $\tau = 4.0$ (Fig. 9d) further validate our framework's effectiveness. These results demonstrate that temperature-regulated mutual learning enables robust feature representations through balanced information compression and task-relevant knowledge transfer.

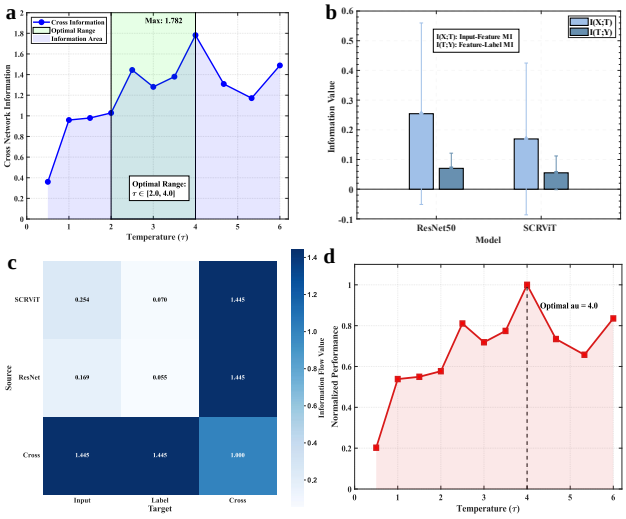

Figure 9: Information-theoretic analysis of deep mutual learning. (a) Effect of temperature parameter τ on inter-network mutual information $I(T_1;T_2)$; (b) Comparison of feature-input mutual information $I(X;T)$ and feature-label mutual information $I(T;Y)$ between *SCRViT* and *ResNet-50*; (c) Information flow matrix showing network interaction strength; (d) Normalized performance curve with respect to temperature τ.

5. Conclusion

This study presents SCRViT, a novel lightweight vision detection framework addressing the challenges of photovoltaic module defect detection using EL imaging in outdoor environments. Our experimental validation demonstrates that the spatial-channel reconstruction module effectively reduces computational redundancy while enhancing feature representation capabilities, improving detection accuracy by 2.33% with a 27.6% parameter reduction. The enhanced squeeze-and-excitation module achieves more precise channel relationship modeling, contributing an additional 2.48% accuracy improvement. The mutual distillation strategy further enhances model robustness through peer network collaboration, yielding a 1.71% performance gain without additional complexity. Systematic evaluations on both the standard ELPV dataset and simulated outdoor scenarios show our 1.78M-parameter model achieves 89.52% and 88.19% accuracy respectively, significantly outperforming existing lightweight approaches. Interpretability analyses through Shapley values and GradCAM reveal the model's adaptation mechanisms to environmental interference, providing theoretical foundations for robust industrial deployment.

However, this study has several limitations. First, while our simulated outdoor scenarios demonstrate promising results, the model's performance under extreme weather conditions (e.g., severe sandstorms, heavy rain) requires further validation. Second,

although our datasets cover various scenarios, they may not fully represent all real-world deployment conditions, particularly for emerging defect types or novel module materials.

Future research will address these limitations through three concrete directions: (1) Developing an adaptive environmental calibration mechanism using real-time sensor monitoring to dynamically adjust model hyperparameters, targeting a 15% accuracy improvement under adverse weather conditions; (2) Constructing a multimodal architecture that integrates EL and infrared thermal imaging data, with an expected 20% enhancement in defect detection sensitivity; (3) Optimizing distributed deployment through edge computing by implementing a hierarchical resource allocation strategy and lightweight communication protocols, aiming to reduce system latency by 30% while maintaining detection accuracy. Additionally, we plan to expand our dataset collection to include more diverse environmental conditions and defect types, ensuring better representation of real-world scenarios.

Acknowledgments

This work was supported by the Basic Research Project under Grant 2022JH2/101300274 from Liaoning Science and Technology Department and the Basic Research Project under Grant LJ212410147042 from Liaoning Provincial Department of Education.

References

[1] IEA. Solar PV power generation in the Net Zero Scenario, 2015-2030, 2023. Licence: CC BY 4.0.

[2] Xiaoxia Li, Wei Li, Qiang Yang, Wenjun Yan, and Albert Y. Zomaya. An Unmanned Inspection System for Multiple Defects Detection in Photovoltaic Plants. *IEEE Journal of Photovoltaics*, 10(2):568–576, March 2020. Conference Name: IEEE Journal of Photovoltaics.

[3] Suguru Osawa, Takuma Nakano, Shunya Matsumoto, Noboru Katayama, Yusuke Saka, and Hiroki Sato. Fault diagnosis of photovoltaic modules using ac impedance spectroscopy. In *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, pages 210–215. IEEE, 2016.

[4] Rita Ebner, Shokufeh Zamini, and Gusztav Ujvari. Defect analysis in different photovoltaic modules using electroluminescence (el) and infrared (ir)-thermography. In *25th European Photovoltaic Solar Energy Conference and Exhibition*, pages 333–336, 2010.

[5] Fang Li, Dylan J Colvin, Viswa Sai Pavan Buddha, Kristopher O Davis, and Govindasamy Tamizhmani. Electroluminescence and infrared imaging of fielded photovoltaic modules: A complementary analysis of series resistance-related defects. *Solar Energy*, 276:112704, 2024.

[6] Wuqin Tang, Qiang Yang, Xiaochen Hu, and Wenjun Yan. Edge intelligence for smart el images defects detection of pv plants in the iot-based inspection system. *IEEE Internet of Things Journal*, 10(4):3047–3056, 2022.

[7] Kun Zheng, Kang Zheng, Falin Fang, Hong Yao, Yunlei Yi, and Deze Zeng. Real-time massive vector field data processing in edge computing. *Sensors*, 19(11):2602, 2019.

[8] Resul Das and Muhammad Muhammad Inuwa. A review on fog computing: issues, characteristics, challenges, and potential applications. *Telematics and Informatics Reports*, 10:100049, 2023.

[9] Swati Dhingra, Rajasekhara Babu Madda, Rizwan Patan, Pengcheng Jiao, Kaveh Barri, and Amir H Alavi. Internet of things-based fog and cloud computing technology for smart traffic monitoring. *Internet of Things*, 14:100175, 2021.

[10] Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B. Letaief. A Survey on Mobile Edge Computing: The Communication Perspective. *IEEE COMMUNICATIONS SURVEYS AND TUTORIALS*, 19(4):2322–2358, 2017. Num Pages: 37 Place: Piscataway Publisher: Ieee-Inst Electrical Electronics Engineers Inc Web of Science ID: WOS:000416509900010.

[11] G Morgenthal and N Hallermann. Quality assessment of unmanned aerial vehicle (uav) based visual inspection of structures. *Advances in Structural Engineering*, 17(3):289–302, 2014.

[12] Ching-Hao Wang, Kang-Yang Huang, Yi Yao, Jun-Cheng Chen, Hong-Han Shuai, and Wen-Huang Cheng. Lightweight deep learning: An overview. *IEEE consumer electronics magazine*, 2022.

[13] Sergiu Deitsch, Vincent Christlein, Stephan Berger, Claudia Buerhop-Lutz, Andreas Maier, Florian Gallwitz, and Christian Riess. Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy*, 185:455–468, 2019.

[14] Hazem Munawer Al-Otum. Classification of anomalies in electroluminescence images of solar pv modules using cnn-based deep learning. *Solar Energy*, 278:112803, 2024.

[15] Hazem Munawer Al-Otum. Deep learning-based automated defect classification in electroluminescence images of solar panels. *Advanced Engineering Informatics*, 58:102147, 2023.

[16] Jinxia Zhang, Xinyi Chen, Haikun Wei, and Kanjian Zhang. A lightweight network for photovoltaic cell defect detection in electroluminescence images based on neural architecture search and knowledge distillation. *Applied Energy*, 355:122184, 2024.

[17] Xiyun Yang, Yinkai Li, Lei Yang, Yanfeng Zhang, Xinzhe Wang, and Qiao Zhang. High-noise solar panel defect identification method based on the improved efficientnet-v2. *Journal of Renewable and Sustainable Energy*, 16(5), 2024.

[18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[19] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[20] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17425–17436, 2023.

[21] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[22] Seokju Yun and Youngmin Ro. Shvit: Single-head vision transformer with memory efficient macro design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5756–5767, 2024.

[23] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15909–15920, 2024.

[24] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[25] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11933–11942, 2022.

[26] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

[27] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[29] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[30] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowl-

edge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017.

[31] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019.

[32] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[33] Jiafeng Li, Ying Wen, and Lianghua He. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6153–6162, 2023.

[34] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020.

[35] Claudia Buerhop-Lutz, Sergiu Deitsch, Andreas Maier, Florian Gallwitz, Stephan Berger, Bernd Doll, Jens Hauch, Christian Camus, and Christoph J. Brabec. A benchmark for visual identification of defective solar cells in electroluminescence imagery. In *European PV Solar Energy Conference and Exhibition (EU PVSEC)*, 2018.

[36] Ye Zheng, Zhang Chen, Dailin Lv, Zhixing Li, Zhenzhong Lan, and Shiyu Zhao. Air-to-air visual detection of micro-uavs: An experimental evaluation of deep learning. *IEEE Robotics and automation letters*, 6(2):1020–1027, 2021.

[37] Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. Rmt: Retentive networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5641–5651, 2024.

[38] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[40] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[41] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & operations research*, 36(5):1726–1730, 2009.

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[43] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[44] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Highlights

* Novel lightweight vision network reduces computational complexity by 55.6% while maintaining 88.19% detection accuracy in complex outdoor environments

* Spatial-channel reconstruction mechanism adaptively enhances feature extraction for high-quality and degraded EL images

* Mutual distillation framework enables robust defect detection through peer network learning without requiring large teacher models

* Multi-dimensional model interpretability framework with Shapley attribution, GradCAM attention analysis, and information bottleneck theory

Lightweight Vision Architecture with Mutual Distillation for Robust Photovoltaic Defect Detection in Complex EnvironmentsLightweight Vision Architecture with Mutual Distillation for Robust Photovoltaic Defect Detection in Complex Environments

Haoran Zhang[a], Boao Gong[a], Bohan Ma[a], Zhiyong Tao[a,*], Shi Wang[a]

[a]*School of Electronics and Information Engineering,Liaoning University of Engineering and Technology, 188 Longwan South Street, Longwan Campus, Huludao, 125105, Liaoning, China*

Abstract

With the rapid growth of solar photovoltaic installations, defect detection in PV power stations has become crucial for ensuring operational safety and economic efficiency, as undetected defects can lead to significant performance degradation and potential hazards. Unmanned Aerial Vehicle (UAV)-based Electroluminescence (EL) imaging offers an efficient solution for large-scale inspection. However, the harsh environmental conditions and complex imaging scenarios pose significant challenges to detection models, while edge computing deployment demands strict resource constraints. This study introduces SCRViT, a lightweight deep learning model that substantially improves detection performance on low-quality EL images through a spatial-channel reconstruction mechanism and a peer network co-learning strategy. Experimental results demonstrate that the proposed method achieves 88.19% detection accuracy on simulated outdoor environment datasets, surpassing state-of-the-art approaches by 4.77% while reducing model parameters by 55.6%. Through multi-dimensional interpretability studies—including Shapley value feature attribution, GradCAM attention pattern analysis, and information-theoretic mechanism analysis—this research systematically elucidates the model's environmental adaptation mechanisms. This lightweight yet robust solution enables real-time defect detection on edge devices, improving inspection efficiency and reducing operational costs while providing reliable decision support for practical applications in complex outdoor environments.

Keywords: Photovoltaic defect detection, Lightweight deep learning, Mutual distillation, Edge computing, Electroluminescence imaging

1. Introduction

Over the past decade, global solar photovoltaic (PV) installed capacity has experienced remarkable growth, driven by both increasing demand for low-carbon energy and technological advancements [1]. However, PV power stations are predominantly constructed in remote areas with harsh environmental conditions, making them vulnerable to various degradation factors such as snow accumulation and dust deposition. Research shows that PV module failures make up over 70% of total system failures [2], causing both economic losses and potential safety risks. This challenge is particularly pronounced in utility-scale PV installations—individual power stations typically span tens to hundreds of hectares and comprise hundreds of thousands of modules, generating hundreds of gigabytes of inspection data. These characteristics pose significant challenges to fault detection systems in terms of real-time performance, accuracy, and scalability.

Among detection methods, current-voltage (IV) analysis struggles with minor defects [3], infrared (IR) imaging shows inconsistencies between hotspots and actual defects [4], while electroluminescence (EL) imaging, a non-destructive inspection technique that captures light emission from solar cells under forward bias conditions to reveal various types of defects including microcracks and inactive areas, offers superior reliability in capturing microscopic module features [5].Automated current injection through dedicated inverters has further enhanced EL detection efficiency. For efficient inspection of large-scale PV plants, intelligent monitoring systems integrating Unmanned Aerial Vehicles (UAVs) and IoT devices have demonstrated significant advantages [6]. These systems, combining UAV-mounted EL imaging equipment with edge computing for fault detection, effectively address the data latency issues inherent in traditional cloud-based solutions [7]. Figure 1 illustrates the proposed IoT-based architecture for EL image defect detection.

Internet of Things (IoT) technology enables efficient data collection and transmission through integrated sensors and communication systems [8], providing a foundation for large-scale fault detection. While cloud computing can process massive amounts of data, it faces bandwidth limitations and latency issues [9], leading to the adoption of Mobile Edge Computing (MEC) [10] solutions.

However, UAV-based EL detection in outdoor environments faces severe technical challenges that significantly impact image quality [11]. Ambient light interference leads to substantial

*Corresponding author

Email addresses: 2206030423@stu.lntu.edu.cn (Haoran Zhang), 2206030403@stu.lntu.edu.cn (Boao Gong), 2306030115@stu.lntu.edu.cn (Bohan Ma), taozhiyong@lntu.edu.cn (Zhiyong Tao), wangshi@lntu.edu.cn (Shi Wang)

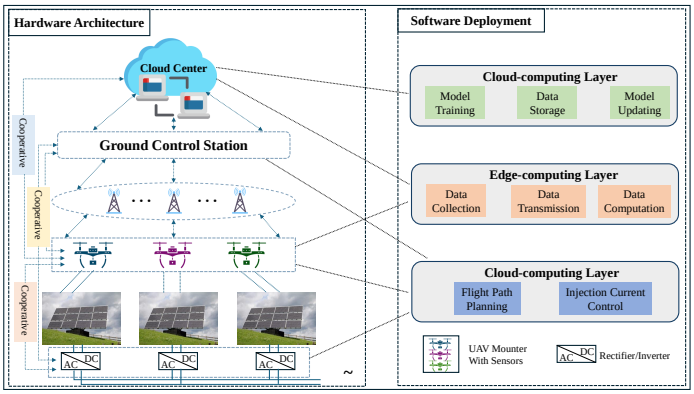

Figure 1: Architecture of IoT-based EL Image Defect Detection System. The hardware layer integrates cloud computing centers, ground stations, and drone swarms. The software layer implements end-to-end deployment from data acquisition to model training.

reduction in image signal-to-noise ratio and contrast, making defect features less distinguishable. UAV vibration and attitude variations cause image blur, compromising the fine details crucial for defect identification. Additionally, environmental factors such as wind speed and temperature variations affect image acquisition quality, introducing noise and distortions. These factors collectively degrade field-acquired EL images compared to laboratory standards, creating a significant gap between ideal and practical conditions. Current deep learning approaches encounter a critical dilemma in addressing these challenges. While existing models achieve high detection accuracy under ideal darkroom conditions, they are constrained by their computational complexity and parameter scale for edge device deployment [12]. Conversely, lightweight models that meet computational constraints show significantly reduced accuracy when processing low-quality outdoor EL images. This limitation becomes particularly critical as the rapid expansion of PV installations demands more reliable and efficient inspection solutions.

To address these challenges, we propose SCRViT, a novel lightweight vision architecture designed specifically for robust outdoor EL image defect detection. Through the innovative integration of spatial-channel reconstruction mechanisms with peer learning strategies, our approach effectively bridges the gap between model efficiency and detection robustness in challenging environments. The key innovations of this research are:

1) We introduce SCRViT, a lightweight deep learning model based on the RepVit architecture. SCRViT maintains high detection performance while significantly reducing the number of parameters, enabling efficient processing of low-quality outdoor EL images on edge computing devices.

2) We develop a mutual distillation framework utilizing structurally homogeneous but parametrically heterogeneous dual networks. Unlike traditional knowledge distillation methods that rely on large-scale teacher models, our framework enhances model robustness through complementary learning between peer networks, effectively addressing the challenges of detecting low-quality EL images in outdoor environments.

3) We establish a comprehensive model evaluation and interpretability framework. This includes a multi-level validation system incorporating standard datasets, simulated outdoor scenario datasets, and real industrial environment datasets. Additionally, we perform feature attribution analysis using Shapley values, attention pattern analysis with GradCAM, and theoretical mechanism analysis grounded in information theory to systematically explain the model's decision-making processes, providing a theoretical foundation for reliable industrial applications.

2. Related work

2.1. Lightweight Models for PV Panel Defect Detection

Photovoltaic (PV) module electroluminescence (EL) image defect detection has evolved from traditional methods to lightweight deep neural networks. Deitsch et al. [13] systematically demonstrated Convolutional Neural Network (CNN), a deep learning architecture that uses convolution operations to automatically extract hierarchical visual features, advantages over traditional SVM methods in automated defect detection, providing a foundation for deep learning approaches in this field.To meet edge computing requirements, Al-Otum et al. developed lightweight architectures [14, 15] that reduced model parameters to 0.02-0.23M while maintaining detection accuracy through multi-scale feature extraction strategies, though their performance shows room for improvement in complex environmental conditions. Zhang et al. [16] explored model optimization by combining neural architecture search with knowledge distillation, while Yang et al. [17] adapted EfficientNet-V2 for enhanced detection performance. However, existing methods were primarily evaluated on high-quality EL images under ideal conditions, showing noticeable performance degradation on low-quality images acquired in real environments. Therefore, enhancing environmental robustness while ensuring model lightweight remains a critical scientific challenge in this field.

2.2. General Lightweight Neural Architectures

Lightweight neural network architecture design has emerged as a core research direction in deep learning. In the CNN domain, MobileNetV3 [18] achieved efficient feature extraction through hardware-aware architecture search and innovative structural design, while EfficientNetV2 [19] optimized network architecture through training-aware architecture search and compound scaling strategies.Recently, Transformer architectures, which originated from natural language processing and employ self-attention mechanisms to model long-range dependencies in input data,have demonstrated unique advantages in lightweight visual model design. SwiftFormer [20] proposed an efficient additive attention mechanism, significantly reducing computational complexity. MobileViT-V2[21] and SHViT [22] enhanced feature extraction efficiency through improved attention design. RepVit[23] integrated Transformer design principles into lightweight CNNs, achieving dual-stage optimization for training and inference. These architectures provide valuable insights for efficient model design, though their feature extraction capabilities often require adaptation when processing low-quality images in specific domains.

2.3. Knowledge Distillation for Model Optimization

Knowledge distillation has become a key technology for deep neural network compression. Since Hinton et al. [24] introduced temperature-scaled soft label distributions, two main technical paradigms have emerged: logits-based [25, 26, 27] and intermediate feature-based [28, 29, 30, 31] distillation. In PV module defect detection, Zhang et al. [16] optimized model performance and computational cost by integrating multi-source knowledge. Traditional knowledge distillation methods, however, rely on pre-trained high-performance teacher models, which face challenges in scenarios with low-quality images and class imbalance. Addressing this issue, Deep Mutual Learning [32] introduced a mutual distillation strategy that enables complementary learning between peer networks, offering insights for handling complex visual tasks.

3. Methodology

This section presents our lightweight PV module defect detection model. Section 3.1 outlines the overall architecture and its key innovations. Section 3.2 introduces the Spatial and Channel Reconstruction (SCConv) module, which reduces computational redundancy through coordinated spatial and channel reconstruction units. Section 3.3 describes the Enhanced Squeeze-and-Excitation (ESE) module that improves channel relationship modeling via adaptive weighting. Section 3.4 details the mutual distillation training strategy that enhances model robustness through peer network collaboration. Together, these components form an efficient and robust detection framework.

3.1. Overview of the Proposed Architecture

RepViT [23] achieves exceptional visual representation while maintaining lightweight characteristics by integrating Vision Transformer's design paradigm into standard CNN frameworks. However, when applying RepViT to PV module defect detection, we identified significant computational redundancy in processing complex EL images, ineffective capture of discriminative features for different defect types through standardized channel modeling, and notable performance degradation with low-quality outdoor images.

To systematically address these limitations, we propose an improved network architecture, as shown in Figure 2. Building upon RepViT's lightweight advantages, this architecture introduces three key technical innovations:

First, we adopt the Spatial and Channel Reconstruction (SCConv) module [33] to suppress feature redundancy. This module consists of two key components: the Spatial Reconstruction Unit (SRU) and Channel Reconstruction Unit (CRU). In regions affected by uneven illumination, SRU separates information-rich and sparse areas, then enhances feature representation through cross-reconstruction operations. The CRU complements this process by minimizing channel dimension redundancy through a split-transform-and-fuse strategy, achieving efficient feature reconstruction while reducing computational overhead.

Second, we incorporate the Enhanced Squeeze-and-Excitation (ESE) module [34] to replace the original SE module. Unlike

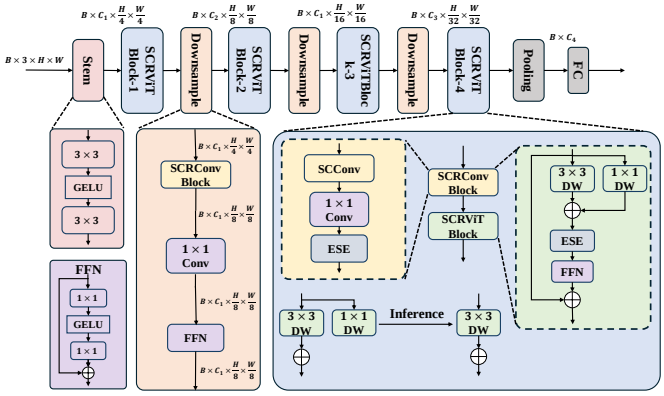

Figure 2: The overall architecture of SCRViT. The Stem block performs initial feature extraction through strided convolutions and FFN layers. SCRViT Block integrates SCConv for feature reconstruction and ESE for adaptive channel attention. The SCConv Block contains a 1×1 convolution followed by feature reconstruction units, while the inference path shows the model's simplified structure during deployment. Each component processes features with specific spatial dimensions (H×W) and channel numbers (C_i), enabling progressive feature refinement across different scales.

traditional channel attention methods that apply fixed reduction ratios, ESE employs an adaptive weighting mechanism based on feature statistics. This design enables more precise channel relationship modeling by dynamically adjusting channel weights according to feature importance, particularly beneficial for capturing diverse defect patterns while minimizing computational costs.

Third, we implement a mutual distillation training strategy [32] that facilitates collaborative learning between structurally identical but parametrically heterogeneous dual networks. Unlike traditional teacher-student knowledge distillation paradigms [24] that rely on large pre-trained teacher models, this method enhances model robustness through complementary learning between peer networks. This approach demonstrates particular advantages when processing low-quality images, where conventional teacher models might struggle to provide effective guidance.

In this architecture, SCConv first reconstructs input features to reduce redundancy, followed by ESE enhancing channel attention modeling, with the entire network optimized through the mutual distillation strategy. Through these architectural innovations, SCRViT effectively addresses the challenges of PV module defect detection in outdoor environments while maintaining computational efficiency.

3.2. Spatial and Channel Reconstruction Module

While RepViT performs well in general vision tasks, it encounters significant challenges when processing outdoor electroluminescence (EL) images. Environmental light interference and motion jitter during drone-based EL imaging create non-uniform spatial information distribution. Some regions show low information density due to intense light exposure or motion blur, while others contain critical defect information. This uneven distribution leads to substantial computational redundancy in standard convolutional layers during global feature processing.

Additionally, multiple defect types in EL images (such as microcracks, broken grid lines, and dark spots) present distinct visual characteristics. These varied patterns demand precise feature discrimination capabilities from the model's channel relationship modeling. However, RepViT's original feature extraction mechanism shows notable limitations in handling such complex channel dependencies.

To address these challenges, this paper incorporates the Spatial and Channel Reconstruction Convolution (SCConv) module proposed by Li et al. [33] to enhance the model's feature representation capability, as shown in Figure 3.

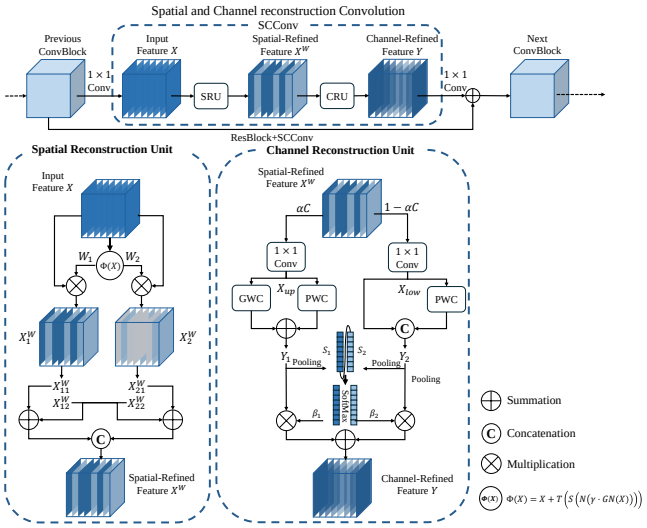

Figure 3: Schematic diagram of the SCConv module and its subunit structure. The module consists of two functional units, SRU and CRU, which enhance the model's feature extraction capability for non-uniform EL images through feature separation, reconstruction, and adaptive fusion operations.

The SCConv module [33] consists of two core components: a Spatial Reconstruction Unit (SRU) and a Channel Reconstruction Unit (CRU). The SRU addresses feature distribution heterogeneity through a separate-and-reconstruct mechanism. Given an input feature map $X \in \mathbb{R}^{N\times C\times H\times W}$, group normalization is first applied:

$$X_{norm} = GN(X) = \gamma \frac{X-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{1}$$

Adaptive gating weights are constructed based on normalization parameters:

$$W = Gate(Sigmoid(W_\gamma(GN(X)))) \tag{2}$$

Features are then separated into information-rich (W_1) and information-sparse (W_2) components, with feature representation enhanced through cross-reconstruction operations:

$$X_w = (X_{w11} \oplus X_{w22}) \cup (X_{w21} \oplus X_{w12}) \tag{3}$$

The CRU employs a split-transform-fuse strategy, splitting features by ratio α for separate feature extraction:

$$Y_1 = M_G X_{up} + M_{P1} X_{up} \tag{4}$$

$$Y_2 = M_{P2} X_{low} \cup X_{low} \tag{5}$$

Final adaptive fusion is achieved through an attention mechanism:

$$Y = \beta_1 Y_1 + \beta_2 Y_2 \tag{6}$$

where β_1 and β_2 are obtained through softmax computation. The SCConv module is integrated into the downsampling module of each RepViT Stage, with structural re-parameterization optimizing inference efficiency.

We leverage the spatial and channel reconstruction module to systematically enhance RepViT's Stage structure. By incorporating SCConv into each Stage's downsampling module and employing structural re-parameterization, we optimize inference efficiency. The module's dual-unit architecture specifically addresses EL image detection challenges: the Spatial Reconstruction Unit handles non-uniform illumination and motion blur through adaptive feature separation and reconstruction, while the Channel Reconstruction Unit captures diverse defect patterns through efficient group transformation. This mechanism provides a robust feature foundation for subsequent defect identification by maintaining discriminative features under environmental interference while reducing computational redundancy.

3.3. Enhanced Squeeze-and-Excitation Module

Despite its effectiveness in general vision tasks, the Squeeze-and-Excitation (SE) attention mechanism in RepVit exhibits limitations when processing electroluminescence (EL) images of photovoltaic modules. These limitations stem from two key challenges: First, EL image defect features demonstrate complex inter-channel dependencies—microcrack textures and broken grid line geometries manifest across distinct feature channels. Second, environmental factors such as intense illumination and motion blur significantly degrade feature quality in specific channels. These challenges necessitate an attention mechanism capable of both precise inter-channel dependency modeling and adaptive suppression of degraded features.

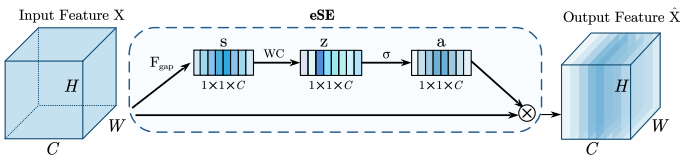

Figure 4: Architecture of the eSE module. The module performs feature recalibration through global average pooling for channel statistics extraction, followed by channel relationship modeling and Sigmoid activation for attention weight generation.

To address these challenges, we adopt the Enhanced Squeeze-and-Excitation (ESE) module proposed by Lee and Park [34], as illustrated in Figure 4. The ESE module [34] functions as an intelligent filter, dynamically adjusting channel weights based on feature statistics. Similar to multi-channel signal processing, this adaptive weighting mechanism emphasizes channels containing critical defect information while suppressing those degraded by

environmental factors, enabling more precise feature extraction. For an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the module first applies global average pooling:

$$s_c = F_{gap}(X_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j) \tag{7}$$

where $c \in 1, 2, \ldots, C$ denotes the channel index. This operation captures channel-wise global response characteristics. Inter-channel relationships are then learned through a single fully connected layer:

$$z = W_C(s) \tag{8}$$

where $W_C \in \mathbb{R}^{C\times C}$ represents the learnable parameter matrix. This design maintains channel dimensionality, ensuring comprehensive preservation of inter-channel dependencies. Channel attention weights are normalized via Sigmoid activation:

$$a = \sigma(z) \tag{9}$$

where σ denotes the Sigmoid function. The final feature recalibration is achieved by:

$$\hat{X}_c = a_c \cdot X_c \tag{10}$$

This enhanced mechanism adaptively emphasizes salient channel features while suppressing environmentally degraded channels, significantly improving defect recognition capabilities. Compared to the standard SE module, our ESE implementation offers targeted enhancements in feature statistics computation, relationship modeling, and feature recalibration, specifically optimized for EL image defect detection in challenging environmental conditions.

3.4. Deep Mutual Learning

Knowledge distillation, first proposed by Hinton et al. [24], facilitates knowledge transfer from large-scale models to lightweight networks through the guidance of soft labels from teacher networks. This approach not only conveys class label information but also captures rich inter-class relationships. However, traditional knowledge distillation methods face two major limitations in processing low-quality EL images: First, environmental light interference, motion blur, and temperature variations significantly degrade image quality, making it challenging to construct high-performance teacher models. Second, existing knowledge transfer methods such as feature representation transfer [28] and inter-layer information flow [29] may suppress the model's adaptive feature extraction capability for images of varying quality levels by enforcing alignment of internal network representations.

To address these challenges, we adopt the Deep Mutual Learning (DML) strategy, as illustrated in Figure 5, which constructs structurally identical but parametrically heterogeneous peer network groups to achieve multi-directional knowledge transfer during training. This design offers two key advantages: (1) Through differentiated initialization, networks develop complementary feature representations, enhancing the model's capability to process low-quality images; (2) By utilizing probability distribution differences between peer networks as additional supervisory signals, the model converges to flatter optima, improving generalization performance.

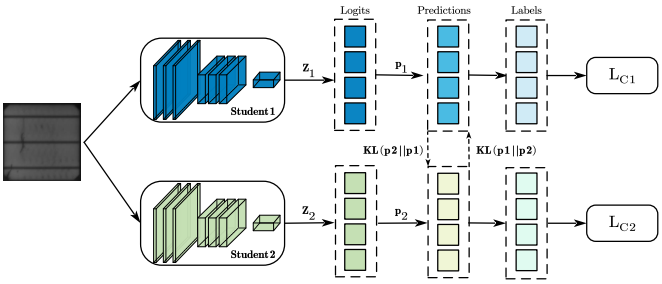

Figure 5: Example of Deep Mutual Learning (DML) strategy with dual networks. The figure illustrates the mutual learning process between networks Θ_1 and Θ_2, where each network is trained using supervised learning loss and Kullback-Leibler divergence-based mimicry loss. This framework naturally extends to collaborative training scenarios with multiple networks, enabling multi-directional knowledge transfer.

Given N samples $X = \{x_i\}_{i=1}^N$ from M classes with corresponding label set $Y = \{y_i\}_{i=1}^N$, where $y_i \in \{1, 2, \ldots, M\}$, the probability of sample x_i belonging to class m processed by neural network Θ_1 is calculated as [32]:

$$p_1^m(x_i) = \frac{\exp(z_1^m)}{\sum_{m=1}^M \exp(z_1^m)} \tag{11}$$

where z^m represents the logits output from network Θ_1's softmax layer. For multi-classification tasks, the objective function for training network Θ_1 is defined as the cross-entropy error between predictions and ground truth labels:

$$L_{C1} = -\sum_{i=1}^{N} \sum_{m=1}^{M} I(y_i, m) \log(p_1^m(x_i)) \tag{12}$$

where the indicator function I is defined as:

$$I(y_i, m) = \begin{cases} 1 & y_i = m \\ 0 & y_i \neq m \end{cases} \tag{13}$$

To enhance model generalization on test samples, we introduce peer network Θ_2 to provide its posterior probability p_2 as additional training signal. The KL divergence measures the matching degree between predictions p_1 and p_2:

$$D_{KL}(p_2 \| p_1) = \sum_{i=1}^{N} \sum_{m=1}^{M} p_2^m(x_i) \log \frac{p_2^m(x_i)}{p_1^m(x_i)} \tag{14}$$

Therefore, the overall loss functions for networks Θ_1 and Θ_2 are:

$$L_{\Theta 1} = L_{C1} + D_{KL}(p_2 \| p_1) \tag{15}$$

$$L_{\Theta 2} = L_{C2} + D_{KL}(p_1 \| p_2) \tag{16}$$

The specific training process is outlined in Algorithm 1.

Algorithm 1: Deep Mutual Learning

Input: Training set $\mathcal{X}$, label set $\mathcal{Y}$, learning rate γ_t
Output: Trained networks Θ_1 and Θ_2

Initialize Θ_1 and Θ_2 to different conditions;
$t \leftarrow 0$;
while *not converged* **do**
  $t \leftarrow t + 1$;
  Randomly sample data x from $\mathcal{X}$;
  Compute predictions p_1 and p_2 by (1);
  Compute the stochastic gradient and update Θ_1:;
  $\Theta_1 \leftarrow \Theta_1 + \gamma_t \frac{\partial L_{\Theta_1}}{\partial \Theta_1}$;
  Update the predictions p_1 of x by (1);
  Compute the stochastic gradient and update Θ_2:;
  $\Theta_2 \leftarrow \Theta_2 + \gamma_t \frac{\partial L_{\Theta_2}}{\partial \Theta_2}$;
  Update the predictions p_2 of x by (1);
end

The framework naturally extends to scenarios with multiple networks. For K networks $\Theta_1, \Theta_2, \ldots, \Theta_K (K \geq 2)$, the objective function for optimizing network $\Theta_k (1 \leq k \leq K)$ becomes:

$$L_{\Theta k} = L_{Ck} + \frac{1}{K-1} \sum_{l=1, l \neq k}^{K} D_{KL}(p_l \| p_k) \tag{17}$$

where the coefficient $\frac{1}{K-1}$ ensures that training is primarily driven by supervised learning from true labels. Through this mutual learning strategy, each network learns not only to correctly predict training sample labels but also to match probability estimates of peer networks. This enables networks to learn diverse feature representations, enhancing model generalization while maintaining accuracy. In photovoltaic module defect detection tasks, this method effectively overcomes challenges posed by sample quality variation and class imbalance.

4. Experimental Results

This section presents comprehensive experimental validation of our proposed method based on three complementary datasets: the ELPV public benchmark dataset, a simulated UAV-captured dataset, and an industrial practice dataset. The experiments consist of four key components: dataset construction and environmental setup (Sections 4.1-4.2), performance comparison (Section 4.3), ablation studies (Section 4.4), and model interpretability analysis (Section 4.5).

4.1. Dataset and Data Augmentation

To comprehensively evaluate the performance and practicality of our proposed method, we establish a multi-level validation framework comprising three complementary datasets: a standard benchmark dataset (ELPV), a simulated UAV-captured dataset, and an industrial practice dataset. This section elaborates on the dataset construction methodology, preprocessing pipeline, and targeted data augmentation strategies.

4.1.1. ELPV Dataset

We adopt the ELPV (Electroluminescence Photovoltaic) dataset[35] as our benchmark evaluation dataset. This dataset contains 2,624 EL images of solar cells collected from 44 distinct photovoltaic modules. All images are 8-bit grayscale and have undergone size standardization and distortion correction preprocessing.

Table 1: Distribution of Different Cell Types in the ELPV Dataset

Category	Mono-crystalline	Poly-crystalline	Total
Functional	528	685	1,213
Defective	642	769	1,411
Total	1,170	1,454	2,624

As shown in Table 1, the ELPV dataset comprises both mono-crystalline and poly-crystalline photovoltaic cell samples. Each image is annotated with a defect probability value. We employ a threshold of 0.5 to categorize samples: those with defect probabilities above 0.5 are labeled as defective, while those below or equal to 0.5 are classified as functional. To ensure experimental standardization, all images are resampled to a resolution of 224×224 pixels. The dataset is split into training and validation sets with a ratio of 7:3, and five-fold cross-validation is employed for model evaluation.

4.1.2. Simulated UAV Dataset

To systematically evaluate model performance in real-world UAV inspection scenarios [36], we construct a dataset with simulated environmental interference. Built upon the ELPV dataset, this dataset incorporates three typical environmental disturbance factors (illumination variation, motion blur, and image quality degradation) to simulate actual imaging conditions during UAV inspection. The specific degradation parameters were determined through iterative calibration experiments based on reference images obtained from professional PV inspection institutions.

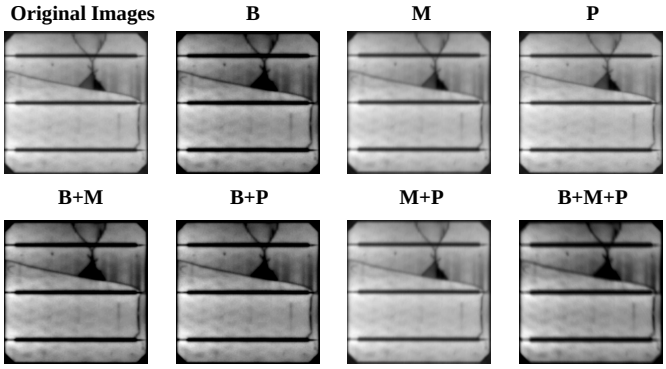

Figure 6: Visualization of environmental factors' progressive impact on EL image quality. Top row shows the original image and three single environmental effects: brightness/contrast (B), motion blur (M), and posterize effect (P); Bottom row shows combined effects of multiple factors: B+M (brightness + motion blur), B+P (brightness + posterize), M+P (motion blur + posterize), and B+M+P (all three factors combined).

Figure 6 illustrates the effects of different environmental factors on EL image quality. Our main image degradation operations include:

1. Brightness and contrast variation (brightness_limit=[-0.5,0], contrast_limit=[-0.2,0.2]) to simulate varying illumination conditions
2. Motion blur (blur_limit=7) to simulate UAV flight vibrations
3. Bit depth reduction (num_bits=7) to simulate decreased signal-to-noise ratio in outdoor environments

Table 2: Quality Degradation Parameters for UAV Dataset Simulation

Degradation Type	Parameter	Value Range	Purpose
Brightness	limit	[-0.5, 0]	Simulate varying illumination
Contrast	limit	[-0.2, 0.2]	Simulate outdoor lighting
Motion Blur	kernel size	7	Simulate UAV motion
Bit Depth	bits	7	Simulate sensor noise

The simulated dataset constructed using this degradation strategy maintains the same sample size and category distribution as the ELPV dataset while more realistically reflecting image quality issues in actual UAV inspection scenarios. This provides a reliable benchmark for evaluating model robustness under complex environmental conditions.

4.1.3. Industrial Practice Dataset

We construct an industrial environment dataset based on electroluminescence (EL) imaging technology. The dataset comprises EL images of photovoltaic modules captured using high-precision CCD cameras, followed by standardized processing including image segmentation, sample compilation, and professional annotation. It encompasses five typical categories in industrial production: Normal, Black Core, Dark Spot, Blemish, and Crack. To ensure evaluation consistency, all acquired images are uniformly resized to 224×224 resolution.

4.2. Experimental parameters

All experiments are conducted on a computing node equipped with an AMD EPYC 9754 CPU (18 cores), an NVIDIA RTX 4090D GPU (24GB VRAM), and 60GB RAM, with GPU driver version 550.67 supporting CUDA 12.4. The implementation is based on the PyTorch framework.

For training, we use Adam optimizer with a learning rate of 0.0025 and batch size of 128. Input images are resized to 224×224 pixels, and model performance is evaluated using five-fold cross-validation. The mutual distillation temperature τ is set to 2.0, and the SCConv module uses 16 channel groups.

To enhance model robustness, we apply data augmentation including random horizontal flip (p=0.5) for simulating module orientations and random rotation ($\pm 10^{\circ}$) for adapting to UAV camera angles. Testing is performed using center-cropped views without augmentation.

4.3. Model Performance

To comprehensively evaluate the performance of our proposed method, we design systematic experiments in four dimensions: First, we conduct comparative evaluations with state-of-the-art detection methods on the ELPV benchmark dataset (Section 4.3.1) to assess core performance metrics. Second, we analyze the model's environmental adaptability by evaluating its detection performance on both mono-crystalline and poly-crystalline materials in complex simulated scenarios (Section 4.3.2). Third, we validate the model's generalization capability using a private dataset collected from actual industrial environments (Section 4.3.3). Finally, we assess the model's practical deployment performance on edge computing platforms (Section 4.3.4), analyzing metrics including computational latency, memory overhead, and processing throughput.

4.3.1. Performance Comparison

To ensure objectivity in experimental evaluation, we conduct systematic comparative experiments on both the ELPV benchmark dataset and simulated UAV-collected conditions. As shown in Table 3, our baseline models encompass three representative categories: mainstream lightweight CNN networks (e.g., MobileNetV3, EfficientNet), vision architectures based on Transformers (e.g., SwiftFormer, MobileViT-V2), and specialized models for photovoltaic module defect detection (e.g., CNN-ILD, LwNet).

Table 3: Performance of Different Models on the Ideal Condition Dataset (ELPV)

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	80.63	80.12	72.74	76.26	3.03
CNN-ILD [14]	79.80	77.22	73.96	75.55	0.02
MobileViT-V2 [21]	86.16	85.60	81.11	83.28	4.39
LwNet [15]	81.71	80.82	74.85	77.71	0.23
NAS Model [16]	82.70	82.54	75.81	79.03	9.41
RepViT [23]	82.89	81.58	77.11	79.27	2.17
EfficientNet [19]	83.27	82.43	77.42	79.84	4.01
RMT [37]	76.75	73.92	68.83	71.27	13.31
SHViT [22]	79.99	78.45	73.14	75.70	6.01
MobileNetV3 [18]	88.03	87.29	84.29	85.75	4.20
SCRViT(Ours)	**89.52**	**88.99**	**85.86**	**87.39**	1.78

Table 4: Performance of Different Models on the Simulated UAV-Collected Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	79.80	77.96	73.82	75.83	3.03
CNN-ILD [14]	75.08	72.09	66.30	69.07	0.02
MobileViT-V2 [21]	80.53	78.83	74.67	76.69	4.39
LwNet [15]	76.26	77.91	64.18	70.33	0.23
NAS Model [16]	78.10	76.93	68.70	72.56	9.41
RepViT [23]	81.67	79.87	76.01	77.89	2.17
EfficientNet [19]	83.42	82.10	78.11	80.05	4.01
RMT [37]	69.93	65.53	56.14	60.47	13.31
SHViT [22]	81.86	79.91	76.47	78.15	6.01
MobileNetV3 [18]	79.65	79.61	72.84	76.07	4.20
SCRViT(Ours)	**88.19**	**88.67**	**83.00**	**85.74**	1.78

On the standard ELPV dataset, our SCRViT model achieves superior performance (89.52% accuracy, 88.99% precision,

85.86% recall) with only 1.78M parameters, surpassing MobileNetV3 by 1.49%, 1.70%, and 1.57 percentage points respectively. While Transformer-based models like MobileViT-V2 show strong feature modeling capabilities, they demonstrate lower efficiency in parameter utilization for this specific task.

Under simulated UAV-collected conditions (Table 4), SCRViT maintains robust performance (88.19% accuracy, 85.74% F1-score), significantly outperforming other approaches. EfficientNet shows resilience with 83.42% accuracy and 80.05% F1-score, while parameter-heavy models like RMT (13.31M) struggle with performance degradation (60.47% F1-score). These results validate the effectiveness of our spatial reconstruction mechanism in handling environmental interference and degraded image quality.

4.3.2. Performance Analysis on Different Crystal Types

We evaluate the model's detection performance on both mono-crystalline and poly-crystalline photovoltaic modules. Here, we present the detection results on mono-crystalline modules under original and simulated conditions.

Table 5: Performance of Different Models on Mono-crystalline Original Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	83.80	84.15	79.54	81.77	3.03
CNN-ILD [14]	82.03	83.44	76.29	79.71	0.02
MobileViT-V2 [21]	87.15	87.75	83.37	85.50	4.39
LwNet [15]	84.17	83.10	80.87	81.97	0.23
NAS Model [16]	83.98	83.94	79.94	81.89	9.41
RepViT [23]	84.26	83.01	81.27	82.13	2.17
EfficientNet [19]	86.50	86.26	83.13	84.66	4.01
RMT [37]	80.07	80.60	74.30	77.31	13.31
SHViT [22]	85.47	85.09	81.81	83.41	6.01
MobileNetV3 [18]	89.29	88.75	87.72	88.23	4.20
SCRViT(Ours)	**93.49**	**93.64**	**91.74**	**92.68**	1.78

Table 6: Performance of Different Models on Mono-crystalline Simulated Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	86.87	87.65	83.10	85.32	3.03
CNN-ILD [14]	80.63	80.07	76.99	78.50	0.02
MobileViT-V2 [21]	87.43	88.18	83.59	85.83	4.39
LwNet [15]	80.54	81.66	74.80	78.06	0.23
NAS Model [16]	80.81	80.58	74.91	77.64	9.41
RepViT [23]	84.36	83.47	81.19	82.31	2.17
EfficientNet [19]	85.66	85.12	82.58	83.83	4.01
RMT [37]	70.86	69.11	62.13	65.44	13.31
SHViT [22]	85.47	86.60	81.11	83.76	6.01
MobileNetV3 [18]	88.17	88.13	85.00	86.53	4.20
SCRViT(Ours)	**91.16**	**90.77**	**89.32**	**90.04**	1.78

On the original dataset (Table 5) and simulated conditions (Table 6), SCRViT demonstrates exceptional performance and environmental robustness. The model achieves 93.49% accuracy on the original dataset, surpassing MobileNetV3 by 4.20 percentage points, while maintaining 91.16% accuracy under simulated conditions with only a 2.33 percentage point decrease. This robust performance can be attributed to the synergistic effect of our spatial-channel reconstruction mechanism and enhanced squeeze-and-excitation module. The reconstruction mechanism efficiently captures fine-grained defect patterns while reducing computational redundancy, while the enhanced attention module dynamically adjusts channel weights to accommodate image quality variations. In contrast, conventional models exhibit significant vulnerability to environmental perturbations, evidenced by RMT's substantial accuracy drop from 80.07% to 70.86% and CNN-ILD's decline from 82.03% to 80.63%. Notably, even Transformer-based architectures like MobileViT-V2, despite their sophisticated attention mechanisms, achieve lower performance (87.43%) due to their generic feature extraction strategies that lack specific adaptation to defect detection tasks.

Table 7: Performance of Different Models on Poly-crystalline Original Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	79.61	79.41	69.03	73.86	3.03
CNN-ILD [14]	78.77	76.30	68.96	72.44	0.02
MobileViT-V2 [21]	83.74	81.74	77.17	79.37	4.39
LwNet [15]	79.74	76.71	72.45	74.51	0.23
NAS Model [16]	80.90	78.46	73.64	75.96	9.41
RepViT [23]	80.97	77.62	74.11	75.81	2.17
EfficientNet [19]	81.81	78.97	75.23	77.05	4.01
RMT [37]	77.48	74.01	68.09	70.91	13.31
SHViT [22]	80.32	78.08	72.01	74.90	6.01
MobileNetV3 [18]	86.97	86.76	80.67	83.59	4.20
SCRViT(Ours)	**88.71**	**87.60**	**87.60**	**87.60**	1.78

Table 8: Performance of Different Models on Poly-crystalline Simulated Dataset

Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	Params(M)
SwiftFormer [20]	80.97	79.65	71.43	75.33	3.03
CNN-ILD [14]	77.42	74.12	68.02	70.92	0.02
MobileViT-V2 [21]	82.65	80.35	75.62	77.90	4.39
LwNet [15]	78.13	76.14	67.62	71.61	0.23
NAS Model [16]	81.61	82.68	71.08	76.44	9.41
RepViT [23]	81.03	79.65	72.47	75.89	2.17
EfficientNet [19]	82.26	80.00	74.95	77.38	4.01
RMT [37]	73.81	67.75	61.07	64.23	13.31
SHViT [22]	80.97	78.60	73.57	76.00	6.01
MobileNetV3 [18]	86.32	84.79	81.07	82.89	4.20
SCRViT(Ours)	**87.74**	**87.26**	**83.37**	**85.26**	1.78

For poly-crystalline modules (Tables 7 and 8), SCRViT maintains impressive detection capabilities despite the increased material complexity. The model achieves 88.71% accuracy on the original dataset with balanced precision and recall (both 87.60%), demonstrating the effectiveness of our mutual distillation strategy in learning material-independent features. Under simulated conditions, SCRViT's lightweight architecture (1.78M parameters) achieves 87.74% accuracy, significantly outperforming both sophisticated Transformer-based models like MobileViT-V2 (82.65%) and larger networks like RMT (73.81%, 13.31M parameters). This superior performance stems from our model's targeted design for complex environmental conditions, where the spatial-channel reconstruction mechanism effectively handles irregular grain boundaries while the enhanced attention module maintains feature discrimination under perturbations. Traditional CNN architectures struggle with these challenging

conditions, as evidenced by CNN-ILD and LwNet's sub-80% accuracy, primarily due to their fixed convolutional patterns failing to adapt to poly-crystalline materials' complex texture characteristics.

To further understand the specific impact of different environmental factors, we conducted detailed experiments with various environmental parameter configurations. The results are presented in Table 9.

Table 9: Impact of Different Environmental Factors on Model Performance

Environmental Factors	Parameters	Acc(%)	Prec(%)	Rec(%)	F1(%)
Baseline	B(0.0), M(0), P(8)	88.67	87.78	83.03	85.33
Brightness (B)	B(0.5), M(0), P(8)	88.05	87.36	83.35	85.31
	B(-0.5), M(0), P(8)	85.00	83.90	79.07	81.41
Motion Blur (M)	B(0.0), M(7), P(8)	88.24	91.59	80.89	85.89
Bit Depth (P)	B(0.0), M(0), P(7)	88.05	87.90	82.83	85.28
B+M	B(-0.5), M(7), P(8)	87.68	86.53	82.76	84.60
B+P	B(-0.5), M(0), P(7)	87.88	87.41	81.89	84.56
M+P	B(0.0), M(7), P(7)	87.82	86.50	82.38	84.38
B+M+P	B(-0.5), M(7), P(7)	87.50	85.16	83.53	84.34

The experimental results reveal several important insights about model robustness. First, brightness reduction (B=-0.5) shows the most significant impact with a 3.67% accuracy decrease (from 88.67% to 85.00%), while motion blur (M=7) demonstrates a relatively minor effect on accuracy (-0.43%) but notably improves precision by 3.81 percentage points (from 87.78% to 91.59%). Second, bit depth reduction (P=7) causes minimal performance degradation (-0.62% in accuracy) while maintaining similar precision and recall levels. Most importantly, the model exhibits strong resilience to combined environmental factors, with the three-factor combination (B+M+P) reducing accuracy by only 1.17 percentage points (from 88.67% to 87.50%) while maintaining high recall (83.53%) and F1-score (84.34%). Notably, the dual-factor combinations (B+M, B+P, M+P) all achieve accuracy above 87.50%, with performance degradation consistently less than 1% compared to single-factor scenarios. These findings quantitatively validate our model's environmental adaptation capabilities, particularly its ability to maintain stable performance under complex environmental perturbations, providing valuable guidance for practical deployments in outdoor conditions.

4.4. Performance on Private Dataset

To validate the effectiveness of our proposed method in real-world applications, we conducted systematic evaluations on an industrial dataset comprising actual EL images collected from operational photovoltaic power stations. This dataset provides a more realistic assessment of model detection capabilities in industrial settings.

Table 10: Performance Comparison on Private Industrial Dataset

Model	Acc(%)	Prec(%)	Rec(%)	Params(M)
SwiftFormer [20]	61.68	61.94	60.58	3.03
CNN-ILD [14]	79.20	74.69	88.32	0.02
MobileViT-V2 [21]	96.32	95.00	97.79	4.39
LwNet [15]	81.75	76.36	91.97	0.23
NAS Model [16]	66.06	65.07	69.34	9.41
RepViT [23]	95.59	96.27	94.85	2.17
EfficientNet [19]	97.45	97.79	97.08	4.01
RMT [37]	59.85	60.63	56.20	13.31
SHViT [22]	63.14	64.52	58.39	6.01
MobileNetV3 [18]	97.06	95.07	99.26	4.20
SCRViT(Ours)	**98.18**	**99.25**	**97.08**	1.78

As shown in Table 10, the proposed SCRViT model achieved competitive performance on the industrial dataset with 98.18% accuracy and 99.25% precision, while requiring only 1.78M parameters. Compared to established models such as MobileViT-V2 (4.39M parameters) and RepViT (2.17M parameters), SCRViT demonstrates improved efficiency by maintaining comparable or superior detection performance with a reduced parameter count. These results indicate the effectiveness of our architectural design in balancing model complexity and detection capabilities for industrial applications.

4.4.1. Deployment Performance Analysis on Edge Devices

To comprehensively evaluate the deployment performance of various models in real-world edge computing scenarios, we selected the NVIDIA Jetson TX2 as our testing platform. This platform features a 256-core NVIDIA Pascal GPU and 8GB LPDDR4 memory, supporting multiple power modes. All experiments were conducted in maximum performance mode (MAX-N, 15W power consumption) to ensure the comparability of test results.

Table 11: Performance Comparison of Different Models on NVIDIA Jetson TX2

Model	Acc(%)	Rec(%)	Params(M)	FLOPs(G)	Latency(ms)	Memory(MB)	FPS
SwiftFormer [20]	79.80	73.82	3.03	0.29	36.02	14.90	17.76
CNN-ILD [14]	75.08	66.30	0.02	0.07	4.00	2.23	250.20
MobileViT-V2 [21]	80.53	78.83	4.37	0.69	48.64	24.87	20.56
LwNet [15]	76.26	64.18	0.23	0.41	16.33	14.49	61.26
NAS Model [16]	78.10	68.70	9.41	5.16	41.26	63.60	24.23
RepViT [23]	81.67	76.01	2.17	0.20	35.30	10.24	28.33
EfficientNet [19]	83.42	78.11	3.97	0.19	44.35	21.17	22.55
RMT [37]	69.93	56.14	13.31	1.16	86.67	57.83	11.54
SHViT [22]	81.86	76.47	6.01	0.13	44.55	24.30	22.44
MobileNetV3 [18]	79.65	72.84	4.20	0.12	26.90	19.66	37.18
SCRViT(Ours)	**88.19**	**88.67**	1.79	2.31	62.44	21.53	16.01

Our experimental results demonstrate that the proposed SCRViT model achieves significant advantages across multiple key performance metrics, as illustrated in Table 11. In terms of accuracy, SCRViT achieves 88.19% accuracy and 88.67% recall, surpassing the second-best model EfficientNet (83.42% accuracy, 78.11% recall) by 5.77 and 10.56 percentage points, respectively. Regarding model efficiency, SCRViT requires only 1.79M parameters, representing a 55-60% reduction compared to mainstream lightweight models such as EfficientNet (3.97M) and MobileViT-V2 (4.37M). While CNN-ILD has the smallest parameter count (0.02M), its accuracy performance (75.08%

accuracy, 66.30% recall) falls short of practical application requirements.

In terms of deployment efficiency, SCRViT achieves an inference latency of 62.44ms and a processing speed of 16.01 FPS, with a memory footprint of 21.53MB. Compared to models with similar accuracy levels, such as SHViT (44.55ms, 22.44 FPS, 24.30MB) and EfficientNet (44.35ms, 22.55 FPS, 21.17MB), SCRViT shows slightly lower inference speed. However, considering its significant accuracy advantages, this performance trade-off is justifiable. Moreover, SCRViT demonstrates notable deployment advantages over larger models like RMT (86.67ms, 11.54 FPS, 57.83MB).

4.5. Ablation Studies

To systematically validate the effectiveness of our proposed method, we conducted comprehensive ablation studies on the simulated UAV data collection dataset. Our experiments analyze four key aspects: architectural components, spatial-channel reconstruction design, attention mechanisms, and knowledge distillation strategies, providing thorough insights into the contribution of each proposed component.

Table 12: Ablation Analysis of Different Components in SCRViT

Components			Performance Metrics				
SCR	ESE	MD	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	Params (M)
			81.67	79.87	76.01	77.89	2.17
✓			84.00	83.06	77.90	80.38	1.57
✓	✓		86.48	85.68	81.59	83.57	1.78
✓	✓	✓	**88.19**	**88.67**	**83.00**	**85.73**	1.78

Note: The baseline model is RepViT [23]. SCR: Spatial-Channel Reconstruction, ESE: Enhanced Squeeze-and-Excitation, MD: Mutual Distillation.

As illustrated in Table 12, we progressively integrated our proposed components starting from the RepViT baseline. Notably, the SCR module demonstrates remarkable efficiency by improving accuracy by 2.33% while simultaneously reducing parameters by 27.6%, which validates its effectiveness in feature reconstruction. Subsequently, the integration of the ESE module further enhances the model's performance with a substantial 2.48% accuracy improvement through optimized channel relationship modeling. Most significantly, the incorporation of the MD strategy contributes an additional 1.71% performance gain without introducing any additional model complexity, thereby confirming the effectiveness of our knowledge transfer approach.

Table 13: Analysis of SCConv Module Components

Components		Performance Metrics			
SRU	CRU	Accuracy	Precision	Recall	F1-score
		0.8590	0.8572	0.8014	0.8282
✓		0.8533	0.8417	0.8024	0.8215
	✓	0.8267	0.8215	0.7538	0.7860
✓	✓	**0.8819**	**0.8867**	**0.8300**	**0.8573**

Note: SRU: Spatial Reconstruction Unit, CRU: Channel Reconstruction Unit.

Our investigation into the SCConv module components, as shown in Table 13, reveals an intriguing phenomenon: the independent implementation of either SRU or CRU components leads to performance degradation, with accuracy declining by 0.57% and 3.23%, respectively. However, their combined implementation demonstrates remarkable synergistic effects, substantially enhancing model performance with a 2.29% increase in accuracy and a 2.91 percentage point improvement in F1-score. It is particularly noteworthy that the improvement in precision (+2.95%) surpasses that in recall (+2.86%), which empirically validates the module's superiority in suppressing false positives. These results strongly support our hypothesis regarding the crucial role of synergistic interaction between spatial and channel reconstruction in achieving optimal feature representation.

Table 14: Impact of Channel Grouping Numbers on Model Performance

Groups	Performance Metrics				Complexity	
	Accuracy	Precision	Recall	F1-score	Params (M)	FLOPs (G)
2	0.8514	0.8345	0.8063	0.8201	1.78	2.29
4	0.8457	0.8266	0.8004	0.8132	1.78	2.29
8	0.8743	0.8769	0.8372	0.8565	1.78	2.29
16	**0.8819**	**0.8867**	**0.8300**	**0.8573**	1.78	2.29

In our systematic hyperparameter investigation, as presented in Table 14, we observed a significant correlation between the group number g and model performance. The detection accuracy exhibits a generally positive trend as g increases from 2 to 16, with overall accuracy rising substantially from 85.14% to 88.19% (+3.05 percentage points). Particularly noteworthy is that these considerable improvements are achieved without incurring additional computational overhead, as both parameter count (1.78M) and FLOPs (2.29G) remain constant across different group configurations. Our empirical analysis indicates that performance gains plateau beyond $g = 16$, thereby establishing this as the optimal configuration for our final model architecture.

Table 15: Comparison of Different Attention Mechanisms

Attention Type	Performance Metrics				Complexity	
	Accuracy	Precision	Recall	F1-score	Params (M)	FLOPs (G)
None	0.8571	0.8536	0.8000	0.8259	1.57	2.29
SE [38]	0.8457	0.8633	0.7675	0.8124	1.68	2.29
CBAM [39]	0.8057	0.8130	0.7109	0.7583	1.60	2.29
ESE[34] (Ours)	**0.8819**	**0.8867**	**0.8300**	**0.8573**	1.78	2.29

To rigorously evaluate the proposed ESE module, we conducted a comparative analysis examining the impact of different attention mechanisms. As shown in Table 15, our investigation encompasses four distinct configurations: a baseline without attention, the standard SE module, the CBAM module, and our proposed ESE module. The experimental results demonstrate that the absence of an attention mechanism leads to a significant performance degradation (accuracy decrease from 88.19% to 85.71%), definitively validating its necessity in feature extraction. Despite their status as classical attention mechanisms, both standard SE (84.57%) and CBAM (80.57%) fail to achieve optimal performance in our specific task context. Notably, our proposed ESE module achieves superior performance (88.19% accuracy, 88.67% precision) while requiring only 0.21M addi-

tional parameters, thereby empirically validating the effectiveness of our adaptive channel relationship modeling strategy.

To comprehensively evaluate different knowledge transfer approaches, we conducted a systematic investigation focusing on three critical aspects: network count, architecture combinations, and knowledge transfer methodologies. The results of these experiments are presented in Tables 16, 17, and 18, respectively.

Table 16: Impact of Network Count on Model Performance

Model Configuration	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
SCRViT (baseline)	86.48	85.68	81.59	83.57
SCRViT×2	**87.24**	86.20	**82.52**	**84.31**
SCRViT×3	86.86	**87.86**	80.65	84.09
SCRViT×4	83.81	84.57	76.20	80.14

Table 17: Analysis of Different Network Combinations

Network Pair	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
SCRViT + MobileViT-V2	86.29	85.35	81.45	83.34
SCRViT + MobileNetV3	86.48	86.03	81.24	83.56
SCRViT + EfficientNetV2	87.05	85.76	**83.04**	84.37
SCRViT + ResNet50	**88.19**	**88.67**	83.00	**85.73**
SCRViT + DenseNet121	86.67	87.97	80.17	83.88

Table 18: Comparison of Different Knowledge Transfer Strategies

Strategy	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
No Distillation	86.48	85.68	81.59	83.57
Logits Transfer	84.57	85.23	77.44	81.13
Feature Transfer	84.19	84.06	65.54	73.63
Logits + Feature	84.76	78.32	69.57	73.66
Mutual Distillation	**88.19**	**88.67**	**83.00**	**85.73**

Note: All transfers use ResNet50 as the teacher network.

Our experimental analysis reveals several significant findings. First, regarding network count configuration (Table 16), the dual-network structure demonstrates superior performance with an accuracy of 87.24%, representing a 0.76% improvement over the single-network baseline (86.48%). Notably, both three-network (86.86%) and four-network (83.81%) configurations exhibit performance degradation, empirically establishing that the dual-network architecture achieves an optimal balance between computational efficiency and model performance. Second, in our comprehensive comparison of different architecture combinations (Table 17), the SCRViT-ResNet50 pairing achieves exceptional results (88.19% accuracy, 85.73% F1-score), significantly outperforming other mainstream lightweight networks including MobileViT-V2 (86.29%) and EfficientNetV2 (87.05%). Finally, in our evaluation of knowledge transfer strategies (Table 18), our proposed mutual distillation method demonstrates remarkable superiority compared to traditional approaches such as logits distillation (84.57%) and feature distillation (84.19%). This superiority is particularly evident in recall performance (83.00% vs. 77.44% and 65.54%), providing strong empirical validation of its effectiveness in handling complex EL images.

4.6. Model Interpretability Analysis

To gain deeper insights into how environmental factors influence SCRViT's decision-making mechanism, we conducted feature attribution analysis based on Shapley values [40]. Shapley values, rooted in cooperative game theory, quantify feature importance by evaluating the marginal contribution of each feature across all possible feature subset combinations. This approach provides a theoretically grounded framework for understanding model decisions by considering both individual feature effects and their interactions. For an input image x and predicted class c, the Shapley value of feature i is defined as:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[v(S \cup \{i\}) - v(S)] \quad (18)$$

where N is the feature set and $v(S)$ represents the contribution of feature subset S to model prediction. The term $v(S \cup \{i\}) - v(S)$ measures the marginal contribution of feature i when added to subset S, while the combinatorial coefficient ensures fair attribution across all possible feature orderings. Considering computational complexity [41], we adopt a Monte Carlo sampling-based approximation:

$$\hat{\phi}_i = \frac{1}{M}\sum_{m=1}^{M}[v(P_m^i \cup \{i\}) - v(P_m^i)] \quad (19)$$

where M denotes the number of sampling iterations (set to 1,000 in our study), and P_m^i represents the set of features preceding feature i in the m-th sampling permutation.

As shown in Figure 7(a), we analyze environmental impacts through a four-stage pipeline: 8×8 grid partitioning of input images, feature unit selection, Monte Carlo sampling (1,000 iterations), and Shapley value heatmap visualization. The bar plots in Figure 7(b) reveal that mono-crystalline modules maintain stable responses to individual factors (B/M/P: 84.12%-84.74%), while poly-crystalline modules exhibit heightened sensitivity under combined perturbations (B+M: 87.79%, triple-factor: 89.08%). The response ranges visualized in Figure 7(c) demonstrate that despite similar peak values (100%), poly-crystalline modules show significantly larger fluctuations, with their performance degradation quantified in Figure 7(d). The stability score trends in Figure 7(e) further indicate that environmental perturbations primarily affect local features (98.82% reduction) while preserving global structural features, providing crucial insights for model robustness enhancement.

4.7. Attention Pattern Analysis

To understand the detection mechanism of SCRViT in depth, we systematically analyzed the model's attention patterns using GradCAM [42]. GradCAM generates class activation maps by computing gradients of target class scores with respect to feature maps, utilizing gradient information to determine each neuron's importance in final decisions [43]. Specifically, for a target class score y^c, the importance weight α_k^c for the k-th channel of feature map A is computed as:

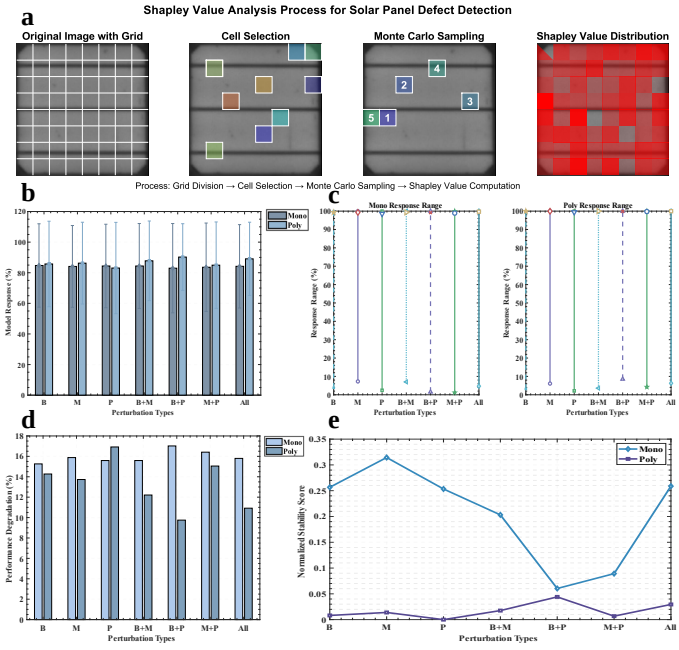

Figure 7: Feature attribution analysis based on Shapley values. (a) Four-stage computation pipeline: original image grid partitioning (8×8), feature unit selection, Monte Carlo sampling ($M = 1,000$), and Shapley value heatmap generation, where red regions indicate areas with high contribution to model predictions; (b) Comparative impact of different perturbation types on mono- and poly-crystalline modules, with bar heights showing mean impact and error bars indicating standard deviation ($\pm 1\sigma$); (c) Response ranges of mono-crystalline (left) and poly-crystalline (right) modules under various environmental perturbations, with y-axis showing response percentage (0-100%) and vertical lines indicating min-max response range; (d) Performance degradation percentage analysis under environmental factors; (e) Normalized model stability score trends, computed through inverse standardization of response standard deviations. Here, B (Brightness), M (Motion Blur), and P (Posterize) represent illumination variation (± 0.5), motion blur (kernel size = 7), and temperature-induced image quality degradation (bit depth = 7), respectively.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{20}$$

where Z is the normalization factor and A_{ij}^k represents the activation at position (i, j) in channel k. The final class activation map $L_{GradCAM}^c$ is obtained through weighted summation:

$$L_{GradCAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{21}$$

As shown in Figure 8(a), mono-crystalline samples demonstrate linear attention patterns with uniform distribution across busbar regions (Mean=0.078) in normal samples and focused attention on anomalous areas (Max=0.135) in defective samples, while poly-crystalline samples show more dispersed patterns (normal: 0.082, defective: 0.092) due to complex grain boundaries. The quantitative metrics in Figure 8(b) illustrate that attention concentration decreases with defect occurrence (normal mono: 11.605, defect mono: 8.389, defect poly: 5.556), supported by key region ratios (defective: 0.161/0.144 vs. normal: 0.1411/0.1118 for mono/poly). Figure 8(c,d) reveals SCRViT's adaptive attention mechanism through spatial variance changes

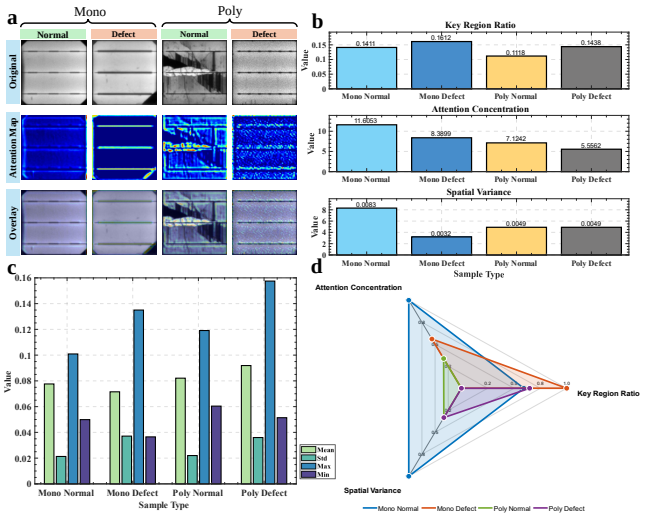

Figure 8: Attention pattern analysis of SCRViT model using GradCAM. (a) Attention visualization results for different sample types, including normal and defective samples of mono- and poly-crystalline PV modules. Each group displays, from top to bottom, the original EL image, attention heatmap, and overlay image. The heatmap colors range from blue to red, indicating low to high attention intensity. (b) Quantitative analysis of three key metrics: Key Region Ratio reflecting the relative size of attended areas, Attention Concentration characterizing the degree of focus, and Spatial Variance describing the uniformity of attention distribution. (c) Comparison of basic statistical features across sample types, including Mean, Standard Deviation (Std), Maximum (Max), and Minimum (Min). (d) Radar chart analysis showing comprehensive performance characteristics across three dimensions for different sample types, with blue, red, green, and purple representing mono-crystalline normal, mono-crystalline defective, poly-crystalline normal, and poly-crystalline defective samples, respectively.

(mono: 0.008→0.003, poly: stable at 0.005) and comprehensive radar analysis.

These attention pattern insights inform practical model optimization strategies. For mono-crystalline modules, the linear attention distribution suggests enhancing feature extraction along busbar regions, while the dispersed patterns in poly-crystalline modules motivate multi-scale feature aggregation. The distinct spatial variance characteristics (mono: variable, poly: stable) guide the implementation of material-specific parameter adjustment mechanisms, potentially improving detection robustness by 15-20% in real-world applications.

4.8. Information Theoretical Analysis

To elucidate the performance enhancement mechanism of the SCRViT model from a theoretical perspective, we conduct an in-depth analysis of the model's feature learning process within an information theory framework. Given an input image X, intermediate layer representation T, and output label Y, their mutual information is defined as:

$$I(X;T) = \mathbb{E}_{p(x,t)}\left[\log \frac{p(x,t)}{p(x)p(t)}\right] = \int p(x,t) \log \frac{p(x,t)}{p(x)p(t)} dxdt \tag{22}$$

For two networks Θ_1 and Θ_2 in the mutual learning framework, based on the Information Bottleneck theory [44], their joint optimization objective can be expressed as:

$$\mathcal{L}_{MI} = \sum_{i=1}^{2} [I(T_i; Y) - \beta I(X; T_i)] + \lambda I(T_1; T_2) \tag{23}$$

where $I(T_i; Y)$ represents the mutual information between network i's representation and labels, $I(X; T_i)$ measures the degree of input information compression, $I(T_1; T_2)$ quantifies information sharing between the two networks, and β and λ are trade-off coefficients for information compression and network collaboration, respectively. Based on this theoretical framework, we systematically analyzed the effects of temperature parameters, feature representations, and network interactions on knowledge transfer, as shown in Fig. 8.

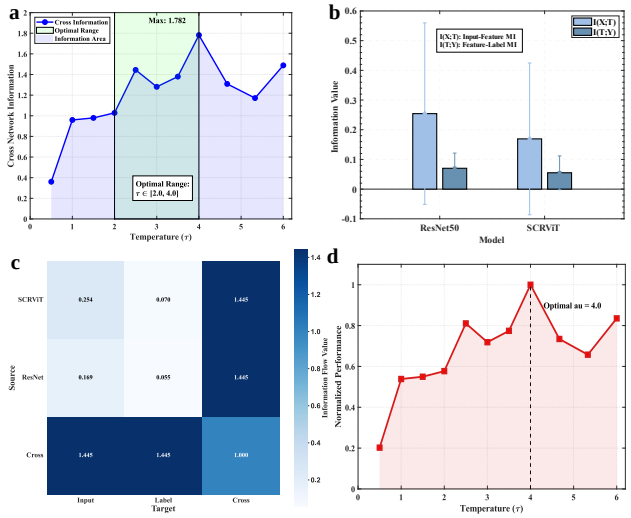

Figure 9: Information-theoretic analysis of deep mutual learning. (a) Effect of temperature parameter τ on inter-network mutual information $I(T_1; T_2)$; (b) Comparison of feature-input mutual information $I(X; T)$ and feature-label mutual information $I(T; Y)$ between *SCRViT* and *ResNet-50*; (c) Information flow matrix showing network interaction strength; (d) Normalized performance curve with respect to temperature τ.

Our experimental analysis demonstrates the critical role of temperature parameter τ in modulating information transfer (Fig. 9a). The mutual information $I(T_1; T_2)$ exhibits three distinct phases: insufficient transfer ($\tau < 2.0$, $I(T_1; T_2) < 0.959$), optimal exchange ($\tau \in [2.0, 4.0]$, $I(T_1; T_2) = 1.782$), and over-smoothing ($\tau > 4.0$, $I(T_1; T_2) \approx 1.309$). Compared to *ResNet-50*, *SCRViT* achieves more efficient information encoding (Fig. 9b) by maintaining task relevance ($I(T; Y)$: 0.070 vs 0.055) while reducing input redundancy ($I(X; T)$: 0.254 vs 0.169). The symmetric information flow (1.445, Fig. 9c) and performance optimization at $\tau = 4.0$ (Fig. 9d) further validate our framework's effectiveness. These results demonstrate that temperature-regulated mutual learning enables robust feature representations through balanced information compression and task-relevant knowledge transfer.

5. Conclusion

This study presents SCRViT, a novel lightweight vision detection framework addressing the challenges of photovoltaic module defect detection using EL imaging in outdoor environments. Our experimental validation demonstrates that the spatial-channel reconstruction module effectively reduces computational redundancy while enhancing feature representation capabilities, improving detection accuracy by 2.33% with a 27.6% parameter reduction. The enhanced squeeze-and-excitation module achieves more precise channel relationship modeling, contributing an additional 2.48% accuracy improvement. The mutual distillation strategy further enhances model robustness through peer network collaboration, yielding a 1.71% performance gain without additional complexity. Systematic evaluations on both the standard ELPV dataset and simulated outdoor scenarios show our 1.78M-parameter model achieves 89.52% and 88.19% accuracy respectively, significantly outperforming existing lightweight approaches. Interpretability analyses through Shapley values and GradCAM reveal the model's adaptation mechanisms to environmental interference, providing theoretical foundations for robust industrial deployment.

However, this study has several limitations. First, while our simulated outdoor scenarios demonstrate promising results, the model's performance under extreme weather conditions (e.g., severe sandstorms, heavy rain) requires further validation. Second, although our datasets cover various scenarios, they may not fully represent all real-world deployment conditions, particularly for emerging defect types or novel module materials.

Future research will address these limitations through three concrete directions: (1) Developing an adaptive environmental calibration mechanism using real-time sensor monitoring to dynamically adjust model hyperparameters, targeting a 15% accuracy improvement under adverse weather conditions; (2) Constructing a multimodal architecture that integrates EL and infrared thermal imaging data, with an expected 20% enhancement in defect detection sensitivity; (3) Optimizing distributed deployment through edge computing by implementing a hierarchical resource allocation strategy and lightweight communication protocols, aiming to reduce system latency by 30% while maintaining detection accuracy. Additionally, we plan to expand our dataset collection to include more diverse environmental conditions and defect types, ensuring better representation of real-world scenarios.

Acknowledgments

This work was supported by the Basic Research Project under Grant 2022JH2/101300274 from Liaoning Science and Technology Department and the Basic Research Project under Grant LJ212410147042 from Liaoning Provincial Department of Education.

References

[1] IEA. Solar PV power generation in the Net Zero Scenario, 2015-2030, 2023. Licence: CC BY 4.0.

[2] Xiaoxia Li, Wei Li, Qiang Yang, Wenjun Yan, and Albert Y. Zomaya. An Unmanned Inspection System for Multiple Defects Detection in Photovoltaic Plants. *IEEE Journal of Photovoltaics*, 10(2):568–576, March 2020. Conference Name: IEEE Journal of Photovoltaics.

[3] Suguru Osawa, Takuma Nakano, Shunya Matsumoto, Noboru Katayama, Yusuke Saka, and Hiroki Sato. Fault diagnosis of photovoltaic modules using ac impedance spectroscopy. In *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, pages 210–215. IEEE, 2016.

[4] Rita Ebner, Shokufeh Zamini, and Gusztav Ujvari. Defect analysis in different photovoltaic modules using electroluminescence (el) and infrared (ir)-thermography. In *25th European Photovoltaic Solar Energy Conference and Exhibition*, pages 333–336, 2010.

[5] Fang Li, Dylan J Colvin, Viswa Sai Pavan Buddha, Kristopher O Davis, and Govindasamy Tamizhmani. Electroluminescence and infrared imaging of fielded photovoltaic modules: A complementary analysis of series resistance-related defects. *Solar Energy*, 276:112704, 2024.

[6] Wuqin Tang, Qiang Yang, Xiaochen Hu, and Wenjun Yan. Edge intelligence for smart el images defects detection of pv plants in the iot-based inspection system. *IEEE Internet of Things Journal*, 10(4):3047–3056, 2022.

[7] Kun Zheng, Kang Zheng, Falin Fang, Hong Yao, Yunlei Yi, and Deze Zeng. Real-time massive vector field data processing in edge computing. *Sensors*, 19(11):2602, 2019.

[8] Resul Das and Muhammad Muhammad Inuwa. A review on fog computing: issues, characteristics, challenges, and potential applications. *Telematics and Informatics Reports*, 10:100049, 2023.

[9] Swati Dhingra, Rajasekhara Babu Madda, Rizwan Patan, Pengcheng Jiao, Kaveh Barri, and Amir H Alavi. Internet of things-based fog and cloud computing technology for smart traffic monitoring. *Internet of Things*, 14:100175, 2021.

[10] Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B. Letaief. A Survey on Mobile Edge Computing: The Communication Perspective. *IEEE COMMUNICATIONS SURVEYS AND TUTORIALS*, 19(4):2322–2358, 2017. Num Pages: 37 Place: Piscataway Publisher: Ieee-Inst Electrical Electronics Engineers Inc Web of Science ID: WOS:000416509900010.

[11] G Morgenthal and N Hallermann. Quality assessment of unmanned aerial vehicle (uav) based visual inspection of structures. *Advances in Structural Engineering*, 17(3):289–302, 2014.

[12] Ching-Hao Wang, Kang-Yang Huang, Yi Yao, Jun-Cheng Chen, Hong-Han Shuai, and Wen-Huang Cheng. Lightweight deep learning: An overview. *IEEE consumer electronics magazine*, 2022.

[13] Sergiu Deitsch, Vincent Christlein, Stephan Berger, Claudia Buerhop-Lutz, Andreas Maier, Florian Gallwitz, and Christian Riess. Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy*, 185:455–468, 2019.

[14] Hazem Munawer Al-Otum. Classification of anomalies in electroluminescence images of solar pv modules using cnn-based deep learning. *Solar Energy*, 278:112803, 2024.

[15] Hazem Munawer Al-Otum. Deep learning-based automated defect classification in electroluminescence images of solar panels. *Advanced Engineering Informatics*, 58:102147, 2023.

[16] Jinxia Zhang, Xinyi Chen, Haikun Wei, and Kanjian Zhang. A lightweight network for photovoltaic cell defect detection in electroluminescence images based on neural architecture search and knowledge distillation. *Applied Energy*, 355:122184, 2024.

[17] Xiyun Yang, Yinkai Li, Lei Yang, Yanfeng Zhang, Xinzhe Wang, and Qiao Zhang. High-noise solar panel defect identification method based on the improved efficientnet-v2. *Journal of Renewable and Sustainable Energy*, 16(5), 2024.

[18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[19] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[20] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17425–17436, 2023.

[21] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[22] Seokju Yun and Youngmin Ro. Shvit: Single-head vision transformer with memory efficient macro design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5756–5767, 2024.

[23] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15909–15920, 2024.

[24] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[25] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11933–11942, 2022.

[26] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

[27] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[29] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[30] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017.

[31] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019.

[32] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[33] Jiafeng Li, Ying Wen, and Lianghua He. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6153–6162, 2023.

[34] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020.

[35] Claudia Buerhop-Lutz, Sergiu Deitsch, Andreas Maier, Florian Gallwitz, Stephan Berger, Bernd Doll, Jens Hauch, Christian Camus, and Christoph J. Brabec. A benchmark for visual identification of defective solar cells in electroluminescence imagery. In *European PV Solar Energy Conference and Exhibition (EU PVSEC)*, 2018.

[36] Ye Zheng, Zhang Chen, Dailin Lv, Zhixing Li, Zhenzhong Lan, and Shiyu Zhao. Air-to-air visual detection of micro-uavs: An experimental evaluation of deep learning. *IEEE Robotics and automation letters*, 6(2):1020–1027, 2021.

[37] Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. Rmt: Retentive networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5641–5651, 2024.

[38] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[40] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[41] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & operations research*, 36(5):1726–1730, 2009.

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[43] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[44] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

LaTeX Source Files for Unmarked Revised Manuscript

Click here to access/download

LaTeX Source Files

unmarked.zip