Defects Detection in Electroluminescence Images of Photovoltaic Modules Using Deep Semantic Segmentation

Kamalesh Barman
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India 400076
Email: 21d070034@iitb.ac.in

Amit Sethi
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India 400076
Email: asethi@iitb.ac.in

***Abstract*—This paper presents deep learning–based approaches for the detection and segmentation of cracks and other defects in electroluminescence (EL) images of photovoltaic (PV) modules. We study three semantic segmentation architectures—U-Net with ResNet-50 encoder, DeepLabV3+ with Xception backbone, and DeepLabV3+ with ResNet-101 backbone—to evaluate their capability in accurately identifying both structural features and defect regions. A class-indexed annotation scheme is employed to distinguish between defective and non-defective areas, followed by rigorous preprocessing and augmentation to enhance dataset diversity. On a benchmark dataset, the DeepLabV3+ model with a ResNet-101 encoder achieves the best results, attaining 92.22% overall accuracy, mean Intersection-over-Union (mIoU) of 0.4761, and mean recall of 0.774. Further testing on a smaller custom dataset annotated with the Computer Vision Annotation Tool (CVAT) yields 91.21% accuracy and 0.3943 mIoU, demonstrating strong generalization despite limited training data. These results highlight the potential of deep encoder–decoder architectures for automated PV module defect inspection.**

***Index Terms*—Electroluminescence imaging, photovoltaic modules, semantic segmentation, DeepLabV3+, ResNet-101, defect detection, cracks.**

I. Introduction

Electroluminescence (EL) imaging has become an indispensable diagnostic technique for assessing the structural integrity, material uniformity, and electrical performance of PV modules [2]. When a module is forward biased, it emits weak infrared radiation whose spatial distribution reflects the underlying electrical activity of individual cells. Capturing this emission allows visualization of hidden or microscopic defects that are otherwise invisible to the naked eye—including microcracks, inactive or poorly conducting regions, shunt pathways, gridline discontinuities, and ribbon breaks [1].

Traditionally, EL image interpretation has relied on manual inspection by domain experts, which is subjective, time-consuming, and difficult to scale. As solar manufacturing scales, there is a growing need for fast, objective, and automated EL-based inspection methods. Deep learning, and in particular convolutional neural networks (CNNs), offers a powerful framework to learn discriminative representations from raw EL data and to perform segmentation or classification with high accuracy and consistency.

In this work we design and evaluate deep learning models for semantic segmentation of EL images, focusing on accurate localization and classification of defects in polycrystalline silicon PV modules. We investigate encoder–decoder architectures using three variants: U-Net with ResNet-50 encoder, DeepLabV3+ with Xception backbone, and DeepLabV3+ with ResNet-101 backbone. We train and compare these models on a public benchmark dataset [1] as well as a smaller custom dataset annotated using CVAT.

Our contributions are:

- A semantic segmentation pipeline for EL images with seven classes covering both structural features and critical defect types.
- A thorough comparison of U-Net and DeepLabV3+ variants (Xception and ResNet-101) with class-weighted training to address imbalance.
- Experimental results showing that DeepLabV3+ with ResNet-101 outperforms a strong published DeepLabV3+ benchmark [1] in terms of mIoU and recall on key defects.
- Analysis of generalization to a smaller, custom CVAT-annotated dataset.

II. Related Work

Early EL-based defect detection methods relied on classical image processing, including edge detection, morphological operations, and texture analysis. While computationally efficient, these approaches struggle under low signal-to-noise ratios and varying illumination, and they often exhibit limited robustness to diverse defect morphologies.

Recent work has focused on deep learning. Chindarkkar *et al.* [2] used CNN-based approaches to detect cracks in EL images of fielded PV modules, demonstrating strong performance for binary “good/bad” classification. However, binary classification provides only a coarse indication of whether a defect exists, without specifying defect type or location.

Pratt *et al.* [1] introduced a benchmark dataset of 593 EL images annotated into 24 semantic classes covering multiple defect and feature types. They evaluated U-Net, PSPNet, and DeepLabV3+ architectures with various class-weighting schemes. Their best-performing configuration, DeepLabV3+ with custom class weights, achieved a median IoU of approximately 0.28 and crack recall of about 0.86 on key defects (crack, inactive, gridline), emphasizing the importance of median-based metrics under heavy class imbalance.

In contrast to the standard DeepLabV3+ backbone used in [1], we integrate a deeper ResNet-101 encoder [4] pretrained on ImageNet, which captures multi-scale context and fine defect boundaries more effectively. We show that this modification, coupled with carefully tuned class weights, yields substantial performance gains, reaching median IoU of 0.38 and recall of 0.94 for the same critical defects.

III. Problem Formulation

We consider the task of semantic segmentation of EL images of PV modules. Given an input image $\mathbf{I} \in R^{H\times W\times 1}$, the goal is to produce a pixel-wise label map $\mathbf{Y} \in \{0,\ldots,C-1\}^{H\times W}$, where C is the number of semantic classes.

The classes cover both non-defective features and defects, including:

- Background and padding regions,
- Ribbons and gridlines,
- Cracks,
- Inactive regions,
- Shunt defects,
- Ribbon defects and related artefacts.

The main challenges are:

- High intra-class variability and irregular geometries of defects (e.g., microcracks, shunts).
- Low contrast and overlapping appearance between defects and structural elements.
- Severe class imbalance, with some defect classes being rare.
- Limited dataset size for training high-capacity deep networks.

IV. Dataset and Preprocessing

A. Datasets

We use two datasets in this study.

1) Benchmark Dataset: The benchmark dataset is that of Pratt *et al.* [1], consisting of EL images of PV modules annotated into 24 feature and defect classes. We follow their evaluation protocol for fair comparison.

2) Custom CVAT-Annotated Dataset: The custom dataset was provided by Dr. Rajiv and consists of EL images of polycrystalline PV modules. A total of 124 individual cells from 27 modules were annotated using CVAT into 7 semantic classes: three non-defective (background, ribbons, gridlines) and four defective (crack, inactive, shunt error, ribbon defect). Each mask was converted into a grayscale, class-indexed format for pixel-wise supervision.

TABLE I
Summary of Available Datasets

Dataset	**#Cells**	**#Modules**	**Cell Type / Location**
Binary EL dataset (good/bad)	5428	139	Mono- and poly-crystalline; Drive / VIP Lab
CVAT-annotated subset	124	27	Mono- and poly-crystalline; Local / Drive / VIP Lab

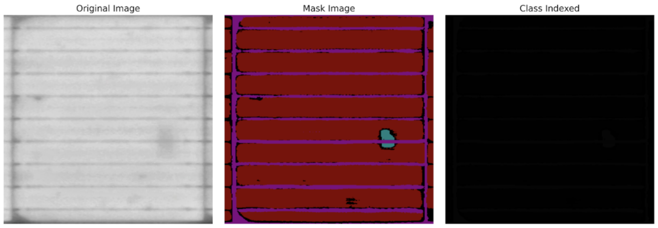

Fig. 1. Data preparation pipeline: original EL image, pixel-wise annotation, and class-indexed mask representation.

For broader context, Table I summarizes both the full binary classification dataset and the CVAT-annotated subset.

B. Data Augmentation

Given the limited size of the custom dataset, we employ an extensive augmentation pipeline to mitigate overfitting and improve robustness. Augmentations include:

- Horizontal and vertical flipping,
- Rotations by 90°, 180°, and 270°,
- Scaling and minor geometric transforms (when applicable).

This procedure increases the dataset size by approximately a factor of eight and introduces significant spatial and orientation diversity.

The augmented data are split into 80 % training, 10 % validation, and 10 % test sets, ensuring that test samples remain unseen during training and hyperparameter tuning.

C. Preprocessing

All images and masks are resized to 512 × 512 pixels, intensity-normalized, and converted to tensors. Class-indexed masks enable straightforward use of pixel-wise cross-entropy loss.

Figure 1 illustrates the preprocessing pipeline from raw EL image to class-indexed mask.

V. Network Architectures

A. Model 1: U-Net With ResNet-50 Encoder

We first implement a U-Net architecture with a ResNet-50 encoder [4] as a baseline. The encoder captures hierarchical features through successive convolutional and residual blocks. The decoder consists of transposed convolutions for upsampling, with skip connections bridging encoder and decoder layers to recover fine spatial details.

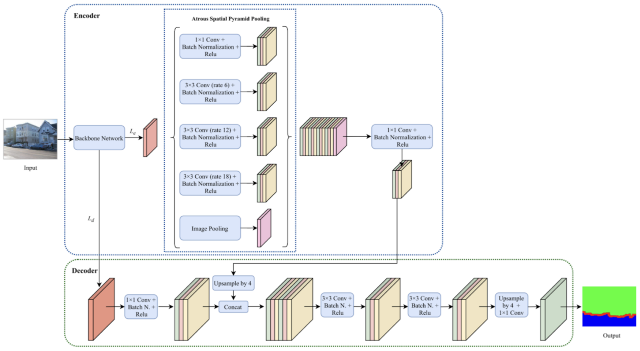
Encoder
Atrous Spatial Pyramid Pooling
1×1 Conv + Batch Normalization + Relu
3×3 Conv (rate 6) + Batch Normalization + Relu
3×3 Conv (rate 12) + Batch Normalization + Relu
3×3 Conv (rate 18) + Batch Normalization + Relu
Image Pooling
Input
Backbone Network
L_e
L_d
1×1 Conv + Batch Normalization + Relu
Decoder
1×1 Conv + Batch N. + Relu
Upsample by 4
Concat
3×3 Conv + Batch N. + Relu
3×3 Conv + Batch N. + Relu
Upsample by 4 + 1×1 Conv
Output

Fig. 2. Conceptual architecture of the DeepLabV3+ model with a ResNet-101 backbone. The encoder uses residual blocks and ASPP; the decoder refines the segmentation at higher resolution.

Each decoder block includes two convolution–batch normalization–ReLU layers. The model outputs class scores at the original resolution using bilinear upsampling.

B. Model 2: DeepLabV3+ With Xception Backbone

The second model is DeepLabV3+ [3] using an Xception backbone. The Xception network employs depthwise separable convolutions to decouple spatial and channel-wise filtering, providing a good trade-off between accuracy and efficiency. The encoder utilizes Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale context via parallel dilated convolutions.

The decoder refines coarse semantic features by merging them with low-level encoder features, yielding sharper segmentation boundaries. This configuration has been successfully used for road boundary estimation and related tasks [5].

C. Model 3: DeepLabV3+ With ResNet-101 Backbone

Our third and best-performing model uses DeepLabV3+ with a ResNet-101 backbone [4], as depicted conceptually in Fig. 2. The deep residual encoder captures both low-level edges and high-level semantics, while ASPP aggregates multi-scale context. The decoder progressively recovers spatial detail and produces refined segmentation masks.

All encoders are initialized with ImageNet-pretrained weights to accelerate convergence and improve generalization.

D. Loss Function and Optimization

To handle class imbalance, we employ class-weighted cross-entropy loss. For a single pixel with one-hot label vector $\mathbf{y}$ and predicted probability vector $\hat{\mathbf{y}}$, the loss is

$$\ell = -\sum_{c=1}^{C} w_c \, y_c \log \hat{y}_c, \tag{1}$$

where w_c is the class weight inversely proportional to the class frequency. The total loss is the average of ℓ over all pixels in the batch.

Class weights are derived from empirical frequencies and further tuned to emphasize defect classes while preserving structural feature importance. Table II summarizes the final weights used for the CVAT dataset.

TABLE II
EXAMPLE CLASS WEIGHTS FOR CVAT-ANNOTATED DATASET

Class	Type	Weight
Background	Feature	0.32
Ribbons	Feature	1.80
Gridline defect	Defect	1.20
Crack	Defect	3.40
Inactive area	Defect	5.10
Shunt error	Defect	6.00
Ribbon defect	Defect	3.50

We use the AdamW optimizer with learning rate 1×10^{-4} and weight decay for regularization. Mixed-precision training (AMP) is used when available to reduce memory usage and training time.

VI. Evaluation Metrics

We evaluate segmentation performance using:

- **Mean IoU (mIoU):** Average Intersection-over-Union across classes.
- **Median IoU:** Median of class-wise IoUs to reduce influence of outliers under imbalance.
- **Median recall:** Median of class-wise recalls, highlighting defect sensitivity.
- **Overall pixel accuracy:** Fraction of correctly labeled pixels over the full image.

These metrics follow the protocol of [1] for direct comparison.

VII. Experiments on Benchmark Dataset

A. U-Net (ResNet-50 Encoder)

The U-Net model with ResNet-50 encoder exhibits poor convergence on the benchmark dataset. Most feature and defect classes show near-zero IoU and recall, with the background class being the only exception (mean IoU ≈ 0.67). The overall accuracy is 67.18 %, and the mean mIoU is only 0.03, indicating that the model fails to learn discriminative masks for complex EL imagery.

B. DeepLabV3+ (Xception Backbone)

Replacing the encoder with Xception and adding ASPP substantially improves performance. The model achieves 88.34 % overall accuracy and mean mIoU of 0.36. Feature classes such as padding, background, and special dogbone regions exhibit strong segmentation quality (median IoU up to $\approx$ 0.99). Among defects, corrosion ribbon, edge darkening, and inactive regions are segmented reasonably well (median IoUs in the range 0.56–0.73). Thin cracks and splice defects remain challenging.

C. DeepLabV3+ (ResNet-101 Backbone)

The DeepLabV3+ model with ResNet-101 backbone delivers the highest segmentation performance. It attains 92.22 % overall accuracy, mean mIoU of 0.476, and mean recall of 0.774. Feature classes such as padding, background, border,

TABLE III
COMPARISON ON BENCHMARK EL DATASET

Model	Accuracy	Median mIoU	Median Reca
U-Net (ResNet-50)	67.18 %	0.03	0.01
DeepLabV3+ (Xception)	88.34 %	0.32	0.69
DeepLabV3+ (ResNet-101)	92.22 %	0.38	0.94
DeepLabV3+ (Benchmark [1])	–	0.28	0.86

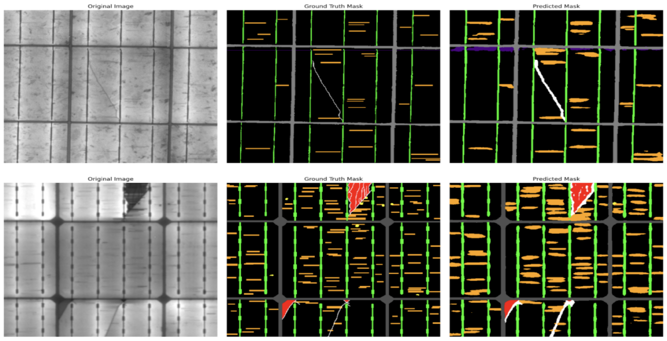

Fig. 3. Representative qualitative results on the benchmark dataset using DeepLabV3+ (ResNet-101). Columns show original EL image, ground-truth mask, and predicted mask.

and dogbone are segmented almost perfectly (median IoUs near 0.8–1.0). Defect classes including inactive regions, rings, brightening, and corrosion ribbon also exhibit strong localization.

For the three critical defects emphasized in [1]—crack, inactive, and gridline—we obtain median IoUs of 0.228, 0.601, and 0.300, respectively, yielding an average median IoU of 0.376 and average median recall of 0.94. Both values surpass the published benchmark.

D. Comparative Summary

Table III summarizes the performance of all three models on the benchmark dataset and compares with the DeepLabV3+ configuration in [1].

Qualitative examples of successful defect localization using the ResNet-101 model are shown in Fig. 3.

VIII. Experiments on CVAT-Annotated Dataset

A. Overall Performance

We further evaluate the DeepLabV3+ (ResNet-101) model on the custom CVAT-annotated dataset of 124 cells. After augmentation and 80/10/10 splitting, the model achieves:

- 91.21 % overall pixel accuracy,
- Mean IoU of 0.3943,
- Mean recall of 0.5554.

These values are slightly lower than those observed on the benchmark dataset, reflecting the smaller size and reduced diversity of the custom data.

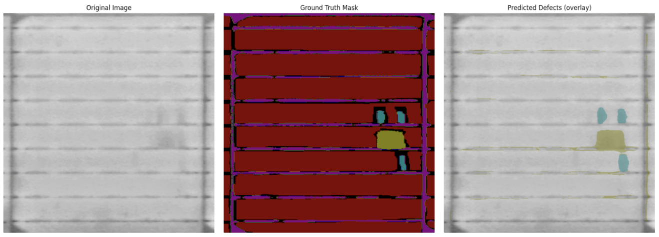

Fig. 4. Qualitative results on the CVAT-annotated dataset. Top: original image, ground truth, and predicted mask. Middle: per-class softmax confidence maps. Bottom: feature activation maps from encoder layers.

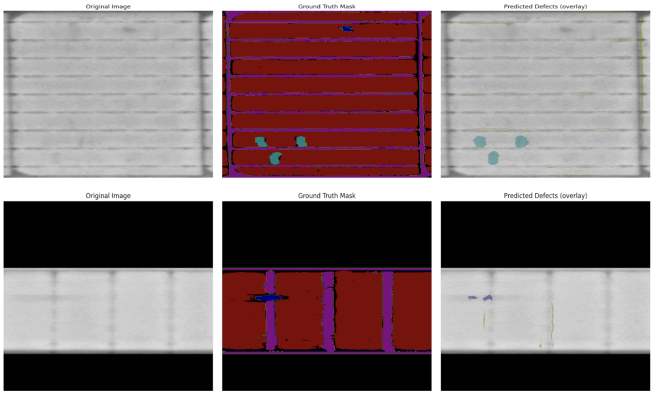

Fig. 5. Example failure cases on the CVAT dataset where subtle defects and low-contrast inactive regions are partially or fully missed.

B. Per-Class Behaviour

Feature classes such as background and ribbons achieve high mIoU (>0.75), while defect classes—especially shunt error and crack—exhibit lower IoUs due to scarcity and annotation difficulty. Despite this, crack recall remains relatively high (approximately 0.74), indicating that the network often detects crack regions but struggles to localize their exact boundaries.

Figure 4 illustrates qualitative predictions alongside ground-truth masks, softmax confidence maps, and intermediate feature activations.

Observed failure modes (Fig. 5) include missed subtle defects and inaccuracies in low-contrast inactive regions, primarily due to limited training diversity.

IX. Discussion

The experimental results indicate that more expressive encoder backbones lead to measurable gains in EL segmentation performance. The ResNet-101 backbone significantly improves both feature and defect segmentation relative to U-Net and DeepLabV3+ with Xception.

Comparing our ResNet-101 model with the DeepLabV3+ benchmark configuration of [1], we observe an absolute improvement of about 0.10 in median IoU on critical defects and a sizable increase in crack recall (0.94 vs. 0.86). We attribute this to:

- Richer multi-scale representation from the deeper residual encoder.
- Carefully tuned class weights emphasizing rare defect classes.
- Consistent use of median statistics to mitigate class imbalance.

On the smaller CVAT dataset, performance degrades modestly but remains strong. This suggests that high-capacity segmentation networks pretrained on large datasets can generalize well to limited EL data when combined with augmentation and balanced loss functions.

Remaining challenges include:

- Robust segmentation of thin cracks and subtle shunt defects.
- Handling label noise and limited annotation precision.
- Coping with variable imaging conditions across modules and setups.

X. Conclusion

We have presented a comprehensive study of deep learning–based semantic segmentation for crack and defect detection in EL images of PV modules. Among three evaluated architectures—U-Net (ResNet-50), DeepLabV3+ (Xception), and DeepLabV3+ (ResNet-101)—the DeepLabV3+ model with ResNet-101 backbone achieved the best performance, surpassing a strong published benchmark on a standard EL dataset.

On the benchmark dataset, the model attained 92.22 % accuracy, 0.4761 mean mIoU, and 0.774 mean recall; on a smaller custom CVAT-annotated dataset, it maintained 91.21 % accuracy and 0.3943 mean mIoU. These results demonstrate that deep encoder–decoder architectures can effectively learn both structural and defect features from EL imagery and form the basis for scalable, automated PV quality inspection systems.

Future work includes exploring transformer-based segmentation architectures (e.g., SegFormer, Swin-UNet), synthetic defect generation to enrich rare classes, and domain adaptation across different EL acquisition setups.

Acknowledgment

The author thanks Prof. Amit Sethi for his guidance and support throughout the entire project. The author also acknowledges the use of OpenAI ChatGPT [7] for minor language refinement and clarity.

References

[1] L. Pratt, J. Mattheus, and R. Klein, “A benchmark dataset for defect detection and classification in electroluminescence images of PV modules using semantic segmentation,” *Systems and Soft Computing*, vol. 5, 2023.

[2] A. Chindarkkar *et al.*, “Deep learning based detection of cracks in electroluminescence images of fielded PV modules,” in *Proc. 47th IEEE Photovoltaic Specialists Conf. (PVSC)*, 2020.

[3] L.-C. Chen *et al.*, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. European Conf. on Computer Vision (ECCV)*, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] S. Das, A. A. Fime, N. Siddique, and M. M. A. Hashem, “Estimation of road boundary for intelligent vehicles based on DeepLabV3+ architecture,” *IEEE Access*, vol. 9, pp. 121231–121242, 2021.

[6] E. Manos, C. Witharana, M. R. Udawalpola, A. Hasan, and A. K. Liljedahl, “Convolutional neural networks for automated built infrastructure detection in the Arctic using sub-meter spatial resolution satellite imagery,” *Remote Sensing*, vol. 14, no. 11, p. 2719, 2022.

[7] OpenAI, “ChatGPT [Large language model],” 2023. Used for rephrasing sentences and improving clarity. Accessed Oct. 23, 2025.