Attention-Enhanced Deep U-Net for Chip Defect Segmentation: Performance and Limitations

Yusen Fu *

Khoury College of Computer Science, Northeastern University, Massachusetts, United States

* Corresponding Author Email: fu.yus@northeastern.edu

Abstract. In semiconductor manufacturing, reliable quality control means finding defects in very large, high-resolution micrographs that only affect a few pixels. This study reevaluates U-Net for chip-defect segmentation and methodically investigates the efficacy of attention mechanisms in enhancing performance. A deep U-Net with a ResNet-34 encoder is trained on 525 annotated 1024 × 1024 images of metal interconnects and vias (80/20 split). The training process employs Augmentations for data augmentation and a hybrid loss (Focal + Generalized Dice) to mitigate extreme class imbalance. Three attention modules—Squeeze-and-Excitation (SE), Efficient Channel Attention (ECA), and CBAM—are integrated into the decoder and compared against a non-attention baseline. Performance is evaluated using Intersection-over-Union (IoU), Dice coefficient, background accuracy, and visual inspection of boundary continuity. All variants achieve strong results: the baseline reaches IoU 0.976, Dice 0.983, and background accuracy 0.995; CBAM slightly improves aggregate scores (IoU 0.978, Dice 0.985), while SE provides smoother edges. However, qualitative analysis indicates that ECA and CBAM often suppress low-contrast traces and cause breaks in thin, elongated interconnects; metric differences across models remain within ±0.003 IoU. These results indicate that for structured, low-texture chip imagery with constrained data, a meticulously calibrated U-Net effectively captures critical cues, while generic attention modules yield minimal or detrimental impacts online continuity. Future research should focus on multiscale fusion and boundary-aware objectives instead of additional attention mechanisms and investigate transfer learning and domain adaptation to enhance applicability.

Keywords: Deep U-Net, semiconductor defect segmentation, attention mechanism.

1. Introduction

Quality control in semiconductor manufacturing remains difficult, particularly when detecting tiny defects in very large, high-resolution imagery. Rare, few-pixel anomalies are easily missed, and manual review is resource-intensive and inconsistent. These pressures have accelerated the use of deep networks for semantic segmentation; U-Net is a natural choice because its encoder–decoder with skip connections preserves detail while aggregating context [1]. A U-Net variant has previously been constructed with a pretrained ResNet-34 encoder, yielding higher Intersection over Union (IoU) and Dice than common baselines on metal interconnect and via segmentation [1].

A limitation persists, however: vanilla U-Net lacks an explicit mechanism to emphasize informative features while suppressing spurious patterns. Channel attention addresses this limitation. According to Hu et al., Squeeze-and-Excitation Networks (SENet) explicitly models the interdependencies between channels to adaptively recalibrate channel-wise feature responses [2]. They also explain that the block acquires the ability to selectively emphasize informative features and suppress less useful ones by utilizing global information. Beyond channels, joint channel–spatial attention refines both what and where to focus. Woo et al. demonstrate that the Convolutional Block Attention Module (CBAM) sequentially infers a one-dimensional channel attention map and a 2D spatial attention map and applies them multiplicatively for adaptive refinement [3]. They empirically verify that utilizing both is more advantageous than relying solely on channel-wise attention.

For resource-constrained deployment, lightweight channel attention is attractive. Qilong et al. introduce Efficient Channel Attention Net (ECA-Net) [4], a method that efficiently captures cross-channel interaction and avoids dimensionality reduction. The method captures local cross-channel interaction by considering every channel and its k neighbors after global average pooling. They also

state that the ECA module is effective. For instance, the parameters and computations against the backbone of ResNet50 are 80 vs. 24.37M and 4.7e-4 GFLOPs vs. 3.86 GFLOPs, respectively. The performance boost in terms of Top-1 accuracy is over 2% [4].

Edge awareness is also crucial when defects coincide with sharp microstructures. BETAM, as proposed by Wang et al., is able to enhance edge features and pay more attention to them while capturing feature dependencies in the three dimensions of position, space, and channel [5]. These trade-offs are especially relevant for chip imagery: wafer layouts are highly regular with sharp edges, while defects are tiny and sparse against repetitive backgrounds. Channel recalibration can surface rare cues but may amplify periodic textures if mis-tuned; spatial attention can improve boundary localization with some latency; and lightweight designs such as ECA reduce overhead while preserving gains.

Guided by these considerations, the present study integrates SE, CBAM, and ECA into U-Net and compares them against a no-attention baseline. Beyond IoU and Dice, the evaluation considers boundary quality and qualitative error modes on chip-defect datasets to clarify when attention helps, which variant is most robust to noise and class imbalance, and what level of computational overhead is justified for high-precision inspection.

2. Method

2.1. Dataset Preparation

Beijing Jingxin Semiconductor Technology Co., Ltd. gave the dataset (Table 1) used in this study. This dataset has high-resolution microscopic pictures taken during the making of semiconductors. They show the structure of the metal interconnect layer and the through-hole layer of integrated circuits. These images capture microscopic details of the manufacturing process and can effectively identify process defects such as microcracks, wire deformation, and through-hole blockages. Because these defects only take up a few pixels in a large image, checking them manually is not very effective and can lead to errors. This makes this dataset particularly useful for testing automated segmentation methods in industrial settings.

During data preparation, each microscopic image was matched with binary masks that were manually marked to show two different types of structure: metal interconnect lines and via (hole) regions. To make it easier to train and test models again, all the images were cropped, normalized, and saved with the same size. To fix the class imbalance, the dataset was split into two parts: training and testing, with an 8:2 ratio. A foreground-aware splitting strategy was used to keep the number of defective and non-defective samples in both sets the same.

Table 1. Overview of the dataset

Property	Description
Source	Beijing Jingxin Semiconductor Technology Co., Ltd.
Image count	525 samples
Resolution	1024 × 1024 pixels
Channels	RGB (3)
Mask types	Binary segmentation of metal line and via regions
Foreground ratio	2–5% of total pixels (sparse targets)
Acquisition method	Optical or SEM-based microscopic imaging
Annotation	Manually labeled by trained engineers
Split ratio	80% training / 20% testing (foreground-aware)

Typical examples show metal interconnects as thin, grid-like conductive paths and vias as circular or ring-shaped features. Defective regions often appear as intensity variations or discontinuities within these structures, forming the segmentation targets analyzed in the subsequent experiments.

2.2. Data Preprocessing

All images were resized to 1024 × 1024 pixels and normalized using the following ImageNet-compatible statistics:

$$\text{mean} = (0.485, 0.456, 0.406), \text{std} = (0.229, 0.224, 0.225). \quad (1)$$

Data augmentation was implemented via the Albumentations library, including:

- Random 90° rotation (RandomRotate90 (p=0.3))
- Random horizontal flip (HorizontalFlip (p=0.5))
- Resizing and normalization (Resize + Normalize)
- Tensor conversion (ToTensorV2())

To address class imbalance, a foreground-aware splitting strategy was employed. Images containing visible metal or via regions were identified and split separately from background-only images, ensuring both training and validation sets contained similar proportions of defective samples.

2.3. Model Architecture

The proposed segmentation network is a modified Deep U-Net consisting of an encoder–decoder structure with skip connections for multi-scale feature fusion. The encoder is based on the ResNet34 architecture without allowing the network to learn directly from the chip imaging data. The decoder reconstructs spatial details through up sampling and convolutional refinement, optionally integrating attention modules to enhance feature weighting.

The encoder comprises five convolutional stages derived from ResNet34: encoder0: initial convolution layer for low-level edge and texture extraction; encoder1–4: four residual blocks capturing increasingly abstract structural and semantic features.

The decoder consists of four transposed convolutions up sampling layers (upconv4 → upconv1). Each level includes up sampling (ConvTranspose2d), skip connection, feature concatenation, and triple convolutional refinement (DoubleConv). The final output is bilinearly interpolated to match the input resolution and passed through a 1×1 convolution to produce three segmentation classes. A rough network architecture is shown in Fig. 1 below, which is from Ranneberger et al. [5].

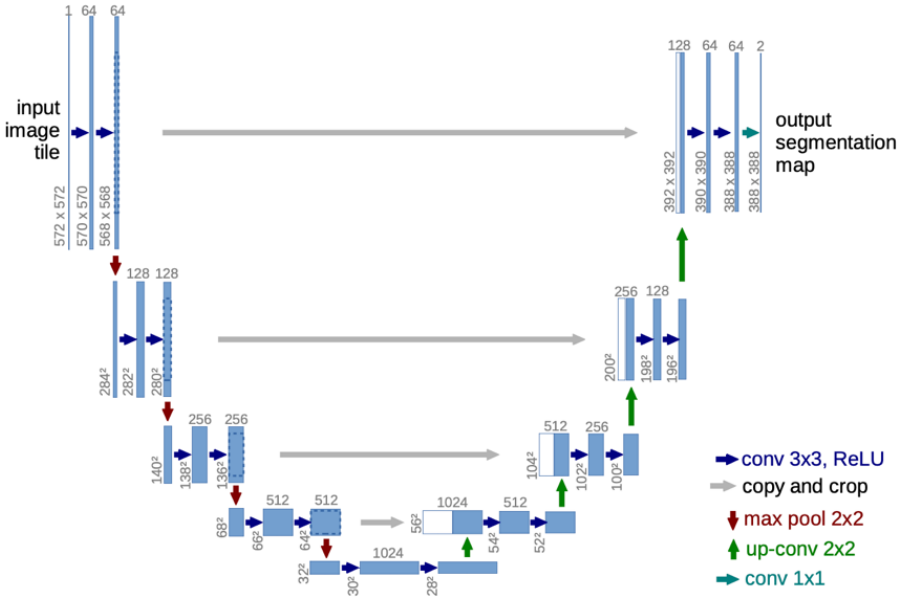

Figure 1. U-Net architecture (adapted from [1]).

Three attention modules were examined to assess their impact on feature representation: Squeeze-and-excitation, which performs channel-wise reweighting via global pooling; Efficient channel attention, a lightweight module that models local cross-channel interaction; and convolutional block attention module, which combines channel and spatial attention. In the proposed Deep U-Net, the selected attention module is inserted only in the decoder stage: after each up-sampling block's

convolutional refinement (DoubleConv) and before the resulting feature map is fed into the next up sampling layer. The module type is selected via the --attn argument (none, se, cbam, or eca).

2.4. Training Configuration

The training process was conducted on an Apple M2 platform using the Metal Performance Shader (MPS) backend for acceleration. Table 2 summarizes the main hyperparameter settings used in the experiments.

Table 2. Training configuration

Parameter	Description
Optimizer	Configuration
Initial learning rate	AdamW
Learning rate scheduler	3×10^{-4}
Batch size	ReduceLROnPlateau (factor = 0.5, patience = 3)
Number of epochs	4
Gradient clipping	25
Device	Max-norm = 0.1

To address class imbalance and enhance sensitivity to minority classes, a hybrid loss combining Focal Loss and Generalized Dice Loss was employed:

$$L = L_{focal} + L_{dice} \tag{2}$$

Focal Loss suppresses the contribution of easily classified pixels, emphasizing rare foreground classes. Generalized Dice Loss applies inverse-squared volume weighting to mitigate class imbalance across categories.

Model performance was assessed using the following metrics: Validation Loss, Background Accuracy (BA), Intersection-over-Union (IoU), Dice Coefficient (F1 Score). These metrics jointly evaluate pixel-level precision and region-level consistency of segmentation outputs.

3. Experiments and Discussion

3.1. Quantitative Results

All experiments were performed under identical data splits and hyperparameter configurations to ensure fair comparison. Table 3 presents the validation performance of different attention mechanisms.

Table 3. Quantitative comparison of model variants

Model	Validation Loss	IoU	Dice	Background Acc
Baseline (No Attention)	1.00	0.976	0.983	0.995
SE [6, 7]	1.01	0.977	0.984	0.994
ECA [8, 9]	1.02	0.975	0.982	0.997
CBAM [10]	1.01	0.978	0.985	0.996

All models achieved strong segmentation performance, with the baseline Deep U-Net already demonstrating excellent accuracy. The inclusion of attention mechanisms (SE, ECA, CBAM) produced only marginal improvements, indicating that the baseline model was sufficient for this dataset.

3.2. Qualitative Results

The following Fig. 2 presents the visual comparison of segmentation predictions from different model variants. The baseline model (without attention) produced the clearest and most continuous segmentation, particularly for the long, thin metal interconnect lines. The SE-enhanced model slightly smoothed the mask edges but occasionally blurred the boundaries. In contrast, both CBAM and ECA

modules significantly weakened the continuity of line structures—several connections were broken or nearly invisible in the output masks.

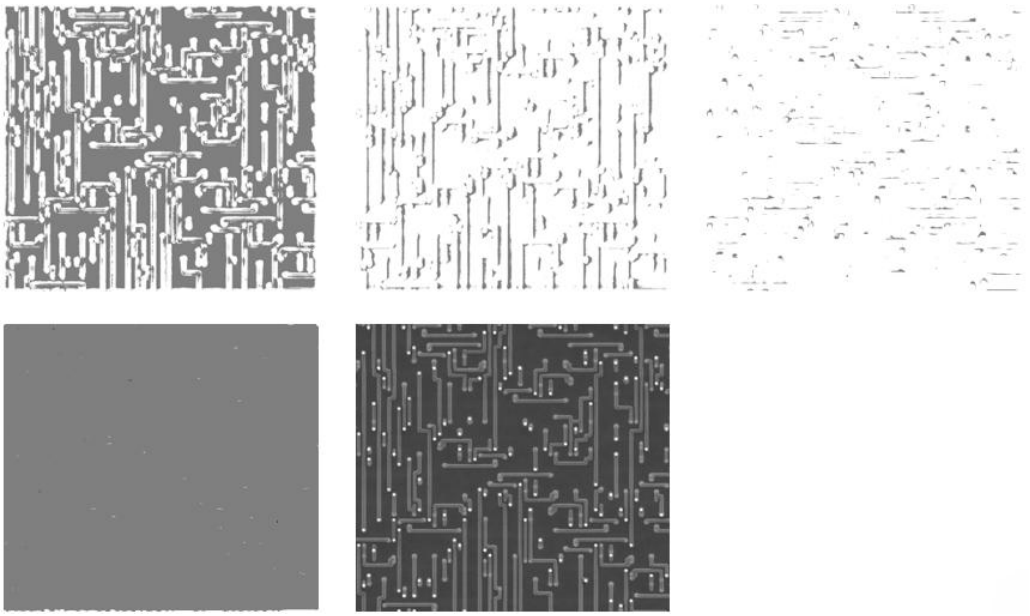

Figure 2. Qualitative comparison of segmentation results using different attention mechanisms. From left to right: (a) None (baseline Deep U-Net without attention), (b) SE (Squeeze-and-Excitation module), (c) CBAM (Convolutional Block Attention Module), (d) ECA (Efficient Channel Attention), and (e) Original image (Picture credit: Original).

These results indicate that while attention mechanisms often help highlight salient regions in complex natural images, they may instead suppress small-scale or low-intensity features in datasets dominated by highly regular, low-texture patterns such as chip microscopy imagery.

3.3. Discussion

The unexpected degradation of segmentation quality after introducing CBAM and ECA attention modules suggests that attention mechanisms are not universally beneficial.

Three factors likely contribute to this phenomenon: 1) Over-suppression of weak features. Chip microscopy images exhibit strong regularity and low texture variation. The model already learns stable global representations through convolutional and residual operations, so additional attention recalibration provides little new information and may distort existing feature maps. 2) Limited feature diversity. Chip microscopy images exhibit strong regularity and low texture variation. 3) Small dataset scale. About 500 samples make up the dataset, which is not enough to reliably train the additional parameters that attention modules introduce. Consequently, during training, attention weights fluctuate, resulting in uneven feature emphasis and unstable activation maps.

Conversely, the baseline Deep U-Net and SE variant achieved stable and interpretable results. The SE block focuses solely on channel-wise weighting through global pooling, which can subtly enhance feature discrimination without disturbing spatial coherence—explaining its slightly improved smoothness compared to the baseline.

Overall, these findings reveal that simpler models may generalize better when the dataset is small, structured, and lacks visual complexity. For such tasks, the key to effective segmentation lies not in introducing complex attention operations but in maintaining stable feature propagation across scales.

3.4. Limitations and Future Work

Despite achieving excellent segmentation accuracy, this study still has two primary limitations. The first limitation is limited dataset diversity: all samples were obtained from a single company and imaging configuration, potentially limiting their applicability to other chip manufacturing contexts.

The second limitation is single-domain evaluation: multi-layer or heterogeneous material structures, which arise in sophisticated semiconductor processes, have not yet been tested using the model.

Future work will focus on two main directions. The first is utilizing multi-scale feature fusion to increase the model's resilience to varying defect sizes. The second is investigating transfer learning techniques and domain adaptation to broaden the model's applicability across different imaging domains and manufacturing settings.

4. Conclusion

This study tackled the difficulty of isolating minute defects in high-resolution chip images, a significant impediment in semiconductor quality assurance. A Deep U-Net with a ResNet-34 encoder and three attention modules (SE, ECA, CBAM) was evaluated through ablation experiments. All models achieved high accuracy (up to IoU 0.978 and Dice 0.985), yet attention did not consistently outperform the baseline and occasionally disrupted the continuity of thin interconnects. For highly regular, low-texture data with limited samples, simpler architectures seem to be more dependable. The present study is constrained by reliance on single-source data and a narrow domain scope. Future improvements will focus on making datasets more diverse, adding multiscale and boundary-aware objectives, and using transfer learning and domain adaptation to make generalization better.

References

[1] Ranneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc Int Conf Med Image Comput Assist Interv (MICCAI). 2015; 9351: 234 – 41.

[2] Woo S, Park J, Lee J, Kweon IS. CBAM: Convolutional block attention module. In: Proc Eur Conf Comput Vis (ECCV). 2018; 3 – 19.

[3] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR). 2020; 11534 – 42.

[4] Wang G, Chen J, Mo L, Wu P, Yi X. Border-enhanced triple attention mechanism for high-resolution remote sensing images and application to land cover classification. Remote Sens. 2024; 16 (15): 2814.

[5] Ranneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc Int Conf Med Image Comput Comput Assist Interv (MICCAI). 2015; 9351: 234 – 41.

[6] Deng J, Ma Y, Li DA, Zhao J, Liu Y, Zhang H. Classification of breast density categories based on SE-attention neural networks. Comput Methods Programs Biomed. 2020; 193: 105489.

[7] Zhigang L, Baoshan S, Kaiyu B. Optimization of YOLOv7 based on PConv, SE attention and wise-IoU. Int J Comput Intell Appl. 2024; 23 (1): 2350033.

[8] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2020; 11534 – 42.

[9] Ni H, Shi Z, Karungaru S, Lv S, Li X, Wang X, Zhang J. Classification of typical pests and diseases of rice based on the ECA attention mechanism. Agriculture. 2023; 13 (5): 1066.

[10] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc Eur Conf Comput Vis (ECCV). 2018; 3 – 19.