Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 22, No. 2, 2026

Validation-uided Lightweight Attention for Small Chip Surface Defect Detection

Yaoyin Chen*

Southwest Minzu University, Chengdu 610000, China

Abstract: Small chip-surface defects are difficult to detect because they occupy few pixels and vary across acquisition domains. This study develops a compact YOLOv8n detector through controlled, validation-only ablations while preserving a held-out test set for later model development. The dataset comprised 3,511 real chip images from two source domains and three defect classes. Increasing the input resolution from 416 to 640 pixels improved validation mAP50 from 0.755 to 0.778, with the largest class-level gain for the smallest DIE_INK defects. Adding a stride-4 P2 head increased computation from 8.1 to 12.2 GFLOPs but reduced mAP50 to 0.729. In contrast, placing a convolutional block attention module on the P3 feature map increased mAP50 to 0.797 with a negligible parameter increase, although mAP50-95 decreased slightly from 0.468 to 0.463. Focal loss and a deterministic Dataset-B-style augmentation both degraded the relevant validation metrics. At a validation-selected confidence threshold of 0.15, the detector-supported sorting rule achieved 95.44% three-way decision accuracy and a 5.88% defect false-accept rate. These results show that targeted attention can improve moderate-overlap detection without the cost of an additional high-resolution head, but they also expose unresolved localization, source-domain, and false-accept limitations. The reported values are development-stage validation evidence and are not estimates of final held-out generalization.

Keywords: Chip surface defect detection; Small-object detection; YOLOv8n; CBAM; Validation ablation; Automated sorting.

1. Introduction

Visual inspection is central to quality control in semiconductor packaging, where ink contamination, package breakage, and cracks can affect downstream handling and product reliability. Automated inspection is attractive because manual screening is difficult to sustain at production speed and because different defect types lead to different dispositions, including acceptance, rework, and rejection. The imaging problem is nevertheless demanding: relevant defects can occupy only a small fraction of a chip image, while illumination, background structure, package edges, and acquisition conditions vary between production sources [1-3].

One-stage detectors offer a practical speed-accuracy compromise for industrial inspection, but small targets are vulnerable to information loss as spatial resolution decreases through a convolutional backbone. Previous chip-defect studies have therefore investigated shallow-feature fusion, attention, feature pyramids, cascaded screening, data generation, and detector ensembling [1-3,6]. These strategies are not uniformly beneficial. A higher-resolution prediction path can increase computational cost and false positives, while attention and hard-example reweighting remain sensitive to placement and training conditions [4,5]. Consequently, module selection should be supported by controlled ablation rather than by architectural accumulation.

A second challenge is evaluation discipline during ongoing model development. Repeatedly consulting a held-out test set converts it into an implicit validation set and weakens the credibility of the final performance estimate. For a short development study, a defensible alternative is to freeze the test set, report only validation evidence, retain negative ablations, and limit the claims accordingly. Such reporting cannot replace a final independent test, but it can establish which design changes merit that later evaluation.

Here, we use a fixed training-validation split to examine five compact detector configurations for three chip-surface defect classes. We test input resolution, a stride-4 P2 head, P3-level convolutional block attention, and focal loss, and we retain a negative appearance-transfer experiment prompted by a source-domain diagnostic. We further translate detector outputs into PASS, REWORK, and REJECT decisions and quantify the resulting accuracy-false-accept trade-off. The study asks a bounded engineering question: which low-cost modification provides the strongest validation evidence without opening the sealed test set?

2. Related Work

2.1. Small chip-surface defect detection

Chip defects differ from generic objects because defect regions are often tiny, weakly contrasted, and embedded in structured package surfaces. Wang et al. [1] combined spatial attention and feature fusion with YOLOv4 and introduced the real and synthetic chip-surface dataset underlying the present source domains. Huang et al. [3] expanded shallow-feature fusion and pruned prediction branches for small chip defects, reporting that shallow spatial information is important when defect regions occupy few pixels. These studies motivate preserving fine-scale information, but they do not imply that every additional shallow prediction head will improve a smaller detector.

2.2. Attention, loss reweighting, and controlled optimization

CBAM sequentially applies channel and spatial attention to refine intermediate feature maps and was designed as a lightweight, general module [4]. Its low overhead makes it suitable for a compact detector, but the appropriate feature level remains task dependent. Focal loss was introduced to reduce the influence of well-classified examples and focus dense-detector training on hard examples [5]. Because its benefit depends on imbalance and hyperparameters, it is treated here as a candidate rather than an assumed

improvement. In semiconductor inspection, Dehaerne et al. [6] similarly showed through controlled YOLOv7 experiments that only some hyperparameter changes improved mean AP, supporting an evidence-led ablation strategy.

2.3. Dataset expansion and production-oriented decisions

Zhu et al. [2] expanded the chip-surface dataset to 2,270 real non-defective, 1,241 real defective, and 7,250 synthetic defective samples, and combined a classifier with an object detector to improve cascade efficiency. Their work also emphasized false acceptance and false rejection as production-relevant outcomes. The present study uses only the 3,511 real images and evaluates a different lightweight path: a single YOLOv8n detector followed by a transparent three-way decision rule. Cross-paper accuracy values are not directly compared because the splits, detector versions, metrics, and deployment conditions differ.

3. Materials and Methods

3.1. Dataset and development protocol

The dataset contained 3,511 real chip images from Dataset A and Dataset B of the publicly accessible Chip-surface-defect-dataset [1,2]. The three object classes were DIE_INK, DIE_BROKEN, and DIE_CRACK. Synthetic images distributed with the source dataset were excluded. A deterministic stratified split with seed 20260806 assigned 2,458 images to training, 351 to validation, and 702 to a held-out test set. The test set remained sealed because model development was ongoing. Architecture selection, loss comparisons, threshold calibration, and source-domain diagnostics were performed only on the training and validation splits.

The validation subset contained 208 Dataset A images and 143 Dataset B images. Dataset B contained no DIE_CRACK validation instances. Its source-domain aggregate therefore includes a zero-valued crack-class contribution and is interpreted as a diagnostic of domain and coverage limitations rather than as a balanced cross-domain benchmark.

3.2. Baseline and training configuration

YOLOv8n was used as the baseline detector. The initial configuration used a 416 x 416-pixel input, 80 epochs, batch size 8, and seed 20260806. A second baseline increased only the input resolution to 640 x 640 pixels. Precision, recall, mAP50, mAP50-95, and class-wise AP50 were calculated on the fixed validation split. The experiments were executed on an NVIDIA RTX 2080 Ti GPU under Ubuntu 22.04.2 using Python 3.10, CUDA 12.3, and the Ultralytics framework.

3.3. Controlled architectural and loss ablations

Three module-level changes were compared with the 640-pixel baseline. First, a stride-4 P2 detection head exposed higher-resolution features to the detector. Second, a CBAM block [4] was placed on the original P3 feature map while the standard three-scale detection head was retained. This P3-CBAM variant contained 3.007 million parameters and required 8.1 GFLOPs, compared with 3.006 million parameters and 8.1 GFLOPs for the 640-pixel baseline. Third, the classification objective of P3-CBAM was replaced by focal loss [5] with gamma = 1.5 and alpha = 0.25. Only the stated component was changed in each comparison.

A deterministic B-style augmentation was evaluated after source-domain analysis revealed a large Dataset A/B gap. Dataset B training images served as appearance references for additional Dataset A training views. Validation images were not transformed and were not used as references. This experiment was retained as a diagnostic negative ablation.

3.4. Validation-only sorting analysis

Detector outputs were mapped to image-level production decisions. Images with no retained defect were assigned PASS; DIE_INK or DIE_CRACK predictions were assigned REWORK; and DIE_BROKEN predictions were assigned REJECT. Confidence thresholds from 0.05 to 0.95 were evaluated on the validation split. Sorting accuracy was the proportion of images receiving the correct three-way decision. The defect false-accept rate was the proportion of defective images incorrectly assigned PASS. The selected threshold was not evaluated on the sealed test set.

3.5. Statistical and reporting boundary

4. Results

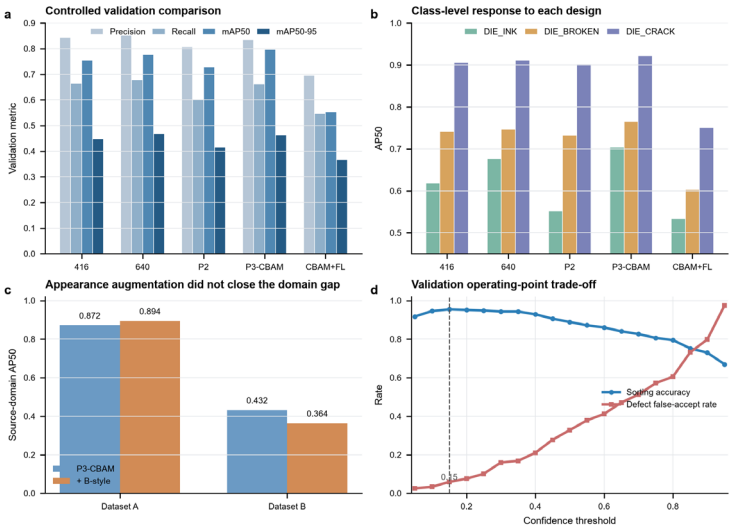

Figure 1. Validation evidence for model selection and image-level sorting. (a) Aggregate metrics for five detector variants. (b) Class-wise AP50. (c) Source-domain AP50 for P3-CBAM and B-style augmentation. (d) Sorting accuracy and defect false-accept rate across confidence thresholds; the dashed line marks 0.15. Values are descriptive results from one fixed validation split and one training seed.

Table 1. Validation results for the five controlled detector variants. Higher is better.

Variant	Prec.	Rec.	mAP50	mAP50-95	INK AP50	BROKEN AP50	CRACK AP50
YOLOv8n 416 px	0.843	0.665	0.755	0.449	0.618	0.742	0.906
YOLOv8n 640 px	0.853	0.678	0.778	0.468	0.676	0.747	0.911
YOLOv8n-P2 640 px	0.806	0.602	0.729	0.415	0.552	0.732	0.902
P3-CBAM 640 px	0.834	0.663	0.797	0.463	0.704	0.765	0.922
P3-CBAM + focal loss	0.696	0.548	0.553	0.367	0.533	0.603	0.751

The experiments used one deterministic split and one training seed. Results are reported as descriptive validation metrics without inferential tests, confidence intervals, or between-seed variability estimates. Images and bounding boxes were not treated as independent experimental replicates. Positive and negative ablation outcomes were retained to reduce selective reporting.

4.1. Resolution predominantly benefited the smallest defect class

Increasing the input size from 416 to 640 pixels raised validation mAP50 from 0.755 to 0.778 and mAP50-95 from 0.449 to 0.468 (Fig. 1a). Precision increased from 0.843 to 0.853 and recall from 0.665 to 0.678. The largest class-level change occurred for DIE_INK, whose AP50 increased from 0.618 to 0.676 (+0.058; Fig. 1b). DIE_BROKEN and DIE_CRACK AP50 each increased by 0.005. The resolution change therefore predominantly benefited the smallest annotated targets.

4.2. The P2 head increased computation and reduced validation performance

Adding the P2 head reduced mAP50 from 0.778 to 0.729 and mAP50-95 from 0.468 to 0.415. Recall fell from 0.678 to 0.602, and DIE_INK AP50 declined from 0.676 to 0.552. The P2 model required 12.2 GFLOPs, compared with 8.1 GFLOPs for the standard 640-pixel model. Thus, the additional high-resolution path increased computational cost without improving any reported aggregate metric.

4.3. P3-CBAM improved AP50 with a localization trade-off

P3-CBAM achieved the highest mAP50 among the five main variants, increasing the 640-pixel baseline value from 0.778 to 0.797. AP50 increased from 0.676 to 0.704 for DIE_INK, from 0.747 to 0.765 for DIE_BROKEN, and from 0.911 to 0.922 for DIE_CRACK. The parameter count increased by approximately 0.001 million and the reported GFLOPs remained 8.1. However, mAP50-95 decreased slightly from 0.468 to 0.463, while precision and recall were 0.834 and 0.663. P3-CBAM therefore improved validation AP at IoU = 0.50 without improving stricter localization.

4.4. Focal loss degraded all reported validation metrics

Replacing the P3-CBAM classification objective with focal loss reduced mAP50 from 0.797 to 0.553 and mAP50-95 from 0.463 to 0.367. Precision decreased from 0.834 to 0.696 and recall from 0.663 to 0.548. Per-class AP50 declined to 0.533, 0.603, and 0.751 for DIE_INK, DIE_BROKEN, and DIE_CRACK, respectively. Focal reweighting was therefore not selected under the tested gamma and alpha values.

4.5. Source-domain analysis exposed an unresolved Dataset B limitation

P3-CBAM achieved a source-domain mean AP50 of 0.872 on the 208 Dataset A validation images and 0.432 on the 143 Dataset B images, a difference of 0.440 (Fig. 1c). B-style augmentation increased the Dataset A diagnostic to 0.894 but reduced the Dataset B value to 0.364. It therefore failed to close the observed domain gap. Because Dataset B contained no DIE_CRACK validation instances, these values describe the current validation composition rather than balanced domain-generalization performance.

4.6. Threshold selection revealed a safety-accuracy trade-off

At the validation-selected confidence threshold of 0.15, three-way sorting accuracy reached 95.44% (Fig. 1d). PASS, REWORK, and REJECT recall were 97.84%, 94.67%, and 84.09%, respectively, while the defect false-accept rate was 5.88%. A lower threshold of 0.10 reduced false acceptance to 3.36% but reduced overall accuracy to 94.59%. The accuracy-maximizing threshold therefore did not minimize the safety-relevant false-accept rate.

5. Discussion

The ablations support a narrow design conclusion: for this fixed validation split, P3-level attention was a more efficient route to higher mAP50 than adding a P2 detection head. The resolution experiment first established that the smallest class benefited from more input pixels. However, exposing an additional stride-4 prediction path did not preserve that gain and increased computation. This result is consistent with the broader observation that shallow spatial features can assist small-object detection [3], while showing that a complete new head can also introduce optimization or false-positive costs in a compact architecture.

The P3-CBAM result should not be interpreted as an across-the-board accuracy gain. Its mAP50 improved by 0.019 and every class AP50 increased, but mAP50-95, precision, and recall were slightly lower than those of the 640-pixel baseline. The evidence therefore suggests improved moderate-overlap detection rather than more precise localization. This distinction matters because a production system may require stable box placement for defect measurement, not only successful class detection at IoU = 0.50.

The negative focal-loss and B-style results are equally informative. Focal loss was motivated by hard and imbalanced dense detection [5], but one gamma-alpha setting sharply degraded this detector. The result does not refute focal loss in general; it shows that the tested reweighting was incompatible with the current data and optimization conditions. Similarly, deterministic appearance transfer did not reduce the Dataset A/B gap. The lower Dataset B result

may reflect imperfect style matching, annotation ambiguity, missing class coverage, or a shift in defect geometry rather than appearance alone. These alternatives cannot be separated by the current experiment.

The sorting analysis also reveals that the best overall decision threshold is not necessarily the safest operating point. A threshold of 0.15 maximized validation accuracy, whereas 0.10 produced fewer false acceptances. Deployment would therefore require a cost-sensitive threshold chosen according to the relative consequences of unnecessary rework and missed defects. The present three-way mapping is transparent, but it has not yet been validated prospectively on production data.

Several limitations constrain the conclusions. All comparisons used one split and one seed, so between-run variability is unknown. Dataset B lacked DIE_CRACK validation examples, weakening source-domain interpretation. The dataset was made publicly accessible by its authors through the original article and a public repository, but the repository did not state a standard open-data license when checked on 12 August 2026. Accordingly, the present study cites the original sources and does not redistribute the image archive. Most importantly, the held-out test set was intentionally not evaluated because model development was still active. The results therefore support architecture selection and a short development-stage report, but not a claim of final generalization or deployment readiness.

6. Conclusion

Controlled validation ablations identified P3-CBAM as the most promising compact variant for three-class chip-surface defect detection. It improved mAP50 from 0.778 to 0.797 with negligible added model cost, whereas a P2 head, focal loss, and B-style augmentation were not beneficial under the tested conditions. The same experiments exposed a small localization trade-off, a large source-domain gap, and a threshold-dependent false-accept risk. These findings justify advancing P3-CBAM to later independent evaluation, while the sealed test set, single-seed design, and unresolved Dataset B limitation define the current boundary of the evidence.

References

[1] Wang, S., Wang, H., Yang, F., Liu, F., & Zeng, L. (2022). Attention-based deep learning for chip-surface-defect detection. International Journal of Advanced Manufacturing Technology, 121, 1957-1971. https://doi.org/10.1007/s00170-022-09425-4

[2] Zhu, X., Wang, S., Su, J., Liu, F., & Zeng, L. (2024). High-speed and accurate cascade detection method for chip surface defects. IEEE Transactions on Instrumentation and Measurement, 73, 1-12. https://doi.org/10.1109/TIM.2024.3351238

[3] Huang, H., Tang, X., Wen, F., & Jin, X. (2022). Small object detection method with shallow feature fusion network for chip surface defect detection. Scientific Reports, 12, Article 3914. https://doi.org/10.1038/s41598-022-07654-x

[4] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), Computer Vision – ECCV 2018 (pp. 3-19). Springer. https://doi.org/10.1007/978-3-030-01234-2_1

[5] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 2980-2988). IEEE. https://doi.org/10.1109/ICCV.2017.324

[6] Dehaerne, E., Dey, B., Halder, S., & De Gendt, S. (2023). Optimizing YOLOv7 for semiconductor defect detection. In Metrology, Inspection, and Process Control XXXVII (Vol. 12496, Article 1249604). SPIE. https://doi.org/10.1117/12.2657564