https://doi.org/10.62836/iaet.v5i1.0008

Global Science Publishing

IAET

Innovations in Applied Engineering and Technology

https://ojs.sgsci.org/journals/iaet

Article

Deep Learning for Industrial Visual Inspection: A Review of Small-Target Detection, Lighting Robustness, and Imbalance-Aware Learning

Shuang Liang * and Jiazheng Xiao

School of Intelligent Manufacturing Engineering, Chongqing University of Arts and Sciences, Chongqing 40216, China

Abstract: Deep learning has substantially increased the automation and accuracy of industrial visual inspection, but performance obtained under curated laboratory conditions often deteriorates once models are deployed on real production lines. The difficulty is especially pronounced when defects occupy only a few pixels, exhibit low contrast, are distorted by uneven illumination, or occur too rarely to support balanced training. Rather than treating these issues separately, this review examines their interaction through 50 representative studies. For small-target inspection, the discussion covers multi-scale pyramids, high-resolution prediction branches, cross-level feature exchange, and detail-preserving designs. For illumination robustness, it reviews image enhancement, illumination normalization, and attention-based feature calibration. For imbalanced learning, it compares resampling, synthetic defect generation, class-aware weighting, logit correction, and gradient rebalancing. Evidence from steel, printed circuit boards, textiles, concrete, and tunnel-lining inspection is used to contrast the strengths and limitations of these approaches. The review also highlights inconsistent definitions of small objects, limited control of lighting conditions, incomplete reporting for minority classes, and insufficient cross-device validation. A more standardized evaluation protocol is therefore advocated, combining scale-, illumination-, and class-distribution-specific metrics with latency and computational-cost reporting.

Keywords: industrial visual inspection; small-target detection; illumination robustness; imbalanced learning; deep learning; intelligent manufacturing

1. Introduction

Machine vision has become an important component of quality assurance and intelligent manufacturing, where it is used to identify surface flaws, structural damage, assembly mistakes, safety-related anomalies, and other abnormal production conditions [1–6]. Its practical appeal over manual inspection lies in rapid throughput, repeatable judgments, and the ability to retain inspection records. Earlier systems depended largely on manually designed rules such as thresholding, edge operators, and texture descriptors. These approaches can work well when the imaging setup is tightly controlled, but their reliability can decline sharply as material texture, lighting, camera settings, or defect morphology changes.

The adoption of deep learning changed this workflow by allowing discriminative representations to be learned directly from inspection images instead of being specified in advance. Convolutional models are now

Received: 1 September 2026; Accepted: 16 September 2026.

* Corresponding: Shuang Liang (liangshuanghaohao@163.com)

used for defect classification, detection, semantic segmentation, and anomaly analysis across steel, PCB, textile, concrete, and tunnel-lining applications. Prior surveys generally report broader adaptability for data-driven representations than for conventional image-processing pipelines. In online inspection, one-stage detectors, particularly members of the YOLO family, are common because they offer a practical compromise between accuracy and inference speed. Nevertheless, strong results on a controlled benchmark do not by themselves establish reliable behavior under the variability of an operating production line.

Despite this progress, three recurring failure modes remain central. First, industrial defects are frequently tiny, slender, irregular, or densely packed, so repeated feature-map downsampling can erase boundary detail before sufficiently discriminative representations are formed. Second, illumination is rarely ideal: low or spatially uneven light reduces contrast, while reflection, shadow, saturation, and colored backlighting can introduce structures that resemble defects. Third, the data distribution is highly skewed because normal samples and common defects dominate training sets, whereas uncommon yet potentially critical defect types may be represented by very few examples. These difficulties are also coupled rather than independent. Small defects are particularly vulnerable to photometric change, and rare categories often contain many small or low-contrast instances.

Existing review articles provide broad taxonomies for defect classification, detection, segmentation, and anomaly-detection methods, but the three issues above are often discussed as separate architectural concerns. Less attention has been paid to how their requirements interact or to cases in which improving one component changes the behavior of another. Evaluation practice creates a related problem. Aggregate accuracy or mAP can remain high even when performance has fallen markedly for small objects, challenging lighting conditions, or infrequent defect classes. A review centered on these coupled failure modes can therefore reveal information that architecture-only summaries tend to hide.

On this basis, the review is organized around three questions. The first concerns how features should be represented and fused when industrial defects are both small and densely distributed. The second examines how image-domain enhancement and feature-domain attention affect detection under illumination interference. The third compares data-level and algorithm-level remedies for imbalanced industrial datasets. Lightweight design and real-time deployment are considered throughout, because memory, latency, and computational constraints determine whether an apparent improvement is useful in practice as well as how effectively the three challenges can be addressed.

A structured literature search was used to construct the review set. Search terms covered industrial inspection, surface-defect detection, small-object detection, multi-scale feature fusion, illumination robustness, attention, class imbalance, long-tailed recognition, defect synthesis, and lightweight deployment. Sources included Google Scholar, IEEE Xplore, ScienceDirect, SpringerLink, and the official repositories of major conferences and publishers. The main search window targeted peer-reviewed journal and conference work published between 2020 and 2026, while earlier foundational studies were retained where necessary. In total, 50 representative papers were selected and organized by application domain, technical challenge, algorithmic approach, reported benefit, limitation, and value for cross-method comparison.

2. Small-Object Detection in Industrial Visual Inspection

2.1. Characteristics and Evaluation of Industrial Small Objects

In industrial inspection, the term small object commonly refers to defects such as scratches, pits, pinholes, welding flaws, broken fibers, fine cracks, missing components, and small printed marks. Because these targets may occupy only a minute fraction of an image, the visual evidence available for recognition is limited from the outset. Dense arrangements further complicate the problem by causing neighboring responses to overlap. Rectangular boxes are also an imperfect representation for defects with irregular outlines or extreme aspect ratios. At the same time, dust, normal workpiece texture, and sensor or imaging noise can produce responses that resemble actual defects, increasing the likelihood of false alarms.

A major source of information loss is the repeated reduction of spatial resolution inside the detector. Deep layers encode stronger semantics but at coarser resolution, so localization and boundary detail are weakened;

early layers preserve spatial structure but offer less class-specific discrimination. Overall mAP is therefore an incomplete summary because medium and large objects can dominate the score and conceal poor behavior on tiny targets. Scale-specific AP, recall, localization error, and the distribution of target sizes should be reported together. The definition of a small object also needs to reflect the imaging setup. Changes in camera distance, sensor resolution, or image resizing can alter the apparent size of the same defect, so applying the generic COCO small-object threshold without scenario-specific adjustment may misrepresent the actual difficulty of an industrial dataset.

Figure 1 depicts how the response associated with a small defect changes as spatial resolution is progressively reduced. At P2, the response remains compact and well localized; after further downsampling, it becomes weaker and spreads over a less precise spatial region. Semantic evidence from deeper features is important for recognizing the defect, but accurate localization still depends on a pathway that carries fine spatial detail to the prediction head. The strategies in Figure 1 intervene at different points in this process: feature pyramids exchange information across resolutions, a P2 head predicts before excessive spatial reduction occurs, and cross-level fusion injects early detail into deeper semantic representations. Attention can emphasize useful responses and suppress background texture, but it cannot recreate spatial information that has already been discarded.

Evaluation should account for target geometry in addition to total pixel area. Rescaling can make the area of a scratch appear larger while its width remains only one or two pixels; a pinhole, by contrast, may be tiny in both dimensions. An area-only definition therefore fails to distinguish these cases and provides little information about defects with very large aspect ratios. Size statistics should be supplemented with the width and height distributions of bounding boxes. Useful diagnostics also include the proportion of targets assigned to each prediction level, recall stratified by size and aspect ratio, and localization error near the detector's minimum resolvable scale. Together, these measurements help determine whether a reported gain comes from stronger representation, better input-scale matching, or improved sample-assignment rules.

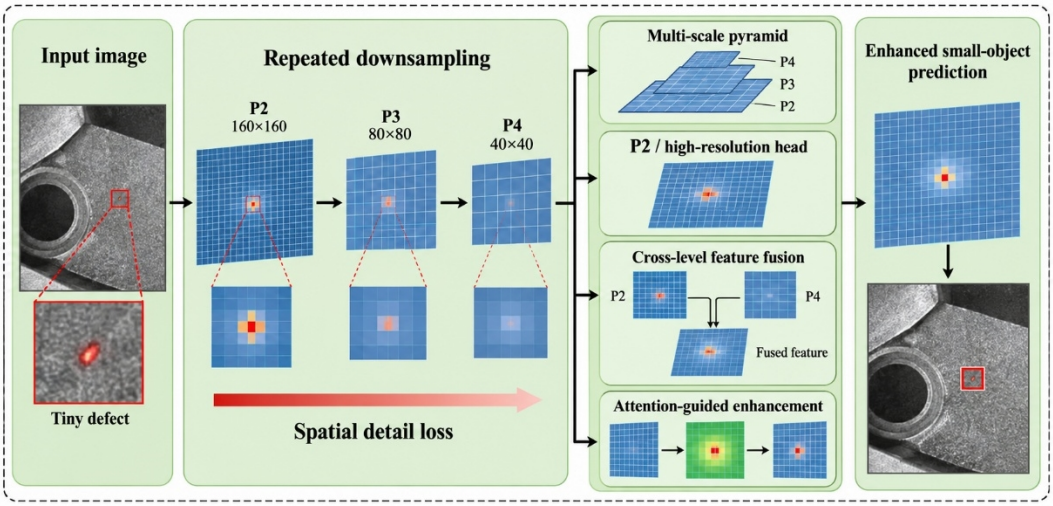

Figure 1. Loss of spatial detail across feature scales and representative remedies for industrial small-target detection.

2.2. Multi-Scale Feature Pyramid Networks

Two-stage detectors provide an important accuracy reference for object detection, whereas the original YOLO formulation showed that classification and localization could be performed in a single network for real-time prediction [7]. Later pyramid designs were motivated largely by the need to preserve information that would otherwise be lost in deep feature hierarchies [8]. FPN combines high-level semantics with higher-resolution detail through lateral connections and a top-down pathway [9]. PANet adds a bottom-up aggregation route, thereby shortening the path through which localization cues travel [10]. EfficientDet employs BiFPN to perform repeated bidirectional fusion and to learn the relative contribution of multi-scale inputs [11]. FaPN addresses a different source of error by aligning upsampled and local feature maps before dense prediction.

Although these designs strengthen communication across scales, repeated rescaling and fusion can increase memory traffic and, consequently, inference latency [12].

More pyramid levels do not automatically translate into better detection. Features from different stages must also be aligned and combined in a way that matches the error characteristics of the task. Element-wise addition is inexpensive, but it implicitly assumes spatial correspondence across feature maps, an assumption that interpolation and downsampling can violate. Concatenation retains the information from both streams more completely, but it expands channel dimensionality and raises the cost of subsequent convolutions. Weighted bidirectional fusion can learn which scales matter most, while explicit alignment modules can correct positional offsets before aggregation. The appropriate choice is therefore problem dependent: missing semantic context calls for stronger contextual exchange, poor boundary localization favors better alignment, and excessive activation by surface texture may require selective reduction of shallow-feature influence.

2.3. High-Resolution Prediction and Detail Preservation

Many standard detectors begin prediction on a feature map with stride eight. Adding a stride-four P2 branch keeps a denser representation and reduces the quantization error that affects very small targets. This change alone is not sufficient, however, because shallow features contain little semantic context and may react strongly to normal texture, which can increase false positives. Information from deeper stages is therefore needed to distinguish genuine defects from background structure. Cross-level fusion provides one way to combine early spatial detail with later semantics, as demonstrated by Zhang et al. for small and extreme-aspect-ratio defects [13]. Other options include space-to-depth operations, dilated or deformable convolution, and increasing the input resolution. Any accuracy gain from these choices should be interpreted together with memory use and batch-one latency.

Whether P2 is worthwhile depends on the distribution of target sizes. Its value is greatest when defect dimensions approach the sampling interval of the conventional P3 prediction level. The additional resolution also has a predictable cost: a stride-four map contains four times as many spatial locations as a stride-eight map, increasing classification and regression work as well as the number of candidate predictions. The overhead can be moderated by using narrower channels, sharing prediction heads, activating the high-resolution path selectively, or removing a large-object branch that contributes little to the application. Thus, the dataset should justify the extra branch rather than P2 being treated as a universally beneficial addition.

Finer prediction maps also make annotation errors more consequential. If a defect is only a few pixels wide, a small shift in the label corresponds to a large relative localization error and can destabilize box regression. Extremely elongated defects are similarly sensitive because modest displacement can sharply reduce intersection over union. For narrow scratches or cracks, overlap-based losses may therefore be insufficient on their own. Depending on the inspection objective, geometry-aware regression, more tolerant positive assignment, or boundary-oriented segmentation can be preferable. Bounding boxes are suitable for counting and screening, masks support measurements such as width or area, and anomaly maps provide an alternative for patterns that were not represented during training.

2.4. Cross-Level and Attention-Guided Fusion

Recent industrial detectors often combine three ideas: cross-scale feature exchange, attention, and lightweight computation, with the exact combination chosen for the target being inspected. CFL-YOLO uses a Context-Fusion module to mix local and global information for steel-surface defects, together with large-kernel attention and multi-scale extraction in deeper layers [14]. GS-YOLO targets densely arranged PCB defects by reducing redundant computation in the feature extractor while adding global context during fusion [15]. HDFA-YOLO combines a lightweight backbone, attention, multi-scale fusion, and detail-enhancing convolution for steel inspection [16]. Mobile-YOLO applies a cascaded, attention-based design to online fiberglass-fabric inspection [17], whereas MSAF-YOLO emphasizes attention-guided fusion across scales for steel surfaces [18]. Designs for other geometries include CS-YOLO for fine concrete cracks [19] and Tunnel-YOLO for real-time crack and leakage detection in tunnel linings [20]. SO-YOLOv8 and LYA-YOLO address general and aerial small-object settings; their underlying ideas may transfer to industrial inspection, but such transfer must be

demonstrated empirically [21,22].

Despite differences in material, defect shape, and imaging environment, the studies summarized in Table 1 converge on several recurring design choices. Lightweight backbones are frequently paired with a modified neck, an attention mechanism, or a branch intended to retain fine spatial detail. Direct ranking of the reported accuracy values would be misleading because class definitions, resolutions, datasets, and evaluation protocols differ across studies. The table therefore emphasizes what each modification is intended to do, its deployment motivation, and the limitation that remains. Most methods are well supported on the dataset for which they were developed, while evidence of robustness across materials, production lines, and imaging devices is much more limited.

Table 1. Selected deep-learning approaches for industrial inspection of small targets.

Ref.	Method/Target	Primary Small-Target Issue	Core Idea	Efficiency Focus	Main Caveat
[13]	Zhang et al.; small and extreme-aspect-ratio defects	Very few pixels; elongated shape	Cross-level fusion with focused-CIoU	Deployment cost needs reporting	Evidence remains application specific
[14]	CFL-YOLO; steel defects	Weak responses across scales	Context fusion plus attention	Single-stage detection	Requires scale-controlled ablation
[15]	GS-YOLO; PCB defects	Small targets in dense layouts	Lightweight extraction and contextual fusion	Designed for low computation	Transfer across devices is unclear
[16]	HDFA-YOLO; steel defects	Loss of fine detail	Light backbone with attention and multi-scale fusion	Real-time-oriented design	Several modules make attribution difficult
[17]	Mobile-YOLO; fiberglass fabric	Fine defects on moving material	Three-stage refinement with attention	Online inspection	Cascade latency depends on hardware
[18]	MSAF-YOLO; steel defects	Weak multi-scale feature response	Attention-guided fusion over scales	Efficiency-aware objective	Cross-dataset robustness is uncertain
[19,20]	CS-YOLO and Tunnel-YOLO; cracks/leakage	Thin and irregular structures	Detail retention with multi-scale features	Lightweight/ real-time emphasis	Geometry is strongly task specific
[21,22]	SO-YOLOv8 and LYA-YOLO; general/aerial objects	Dense populations of small targets	High-resolution prediction and efficient fusion	One-stage/ lightweight design	Industrial-domain transfer still needs validation

2.5. Comparative Findings and Limitations

Across the reviewed literature, gains on small defects are usually obtained in one of three ways: preserving a higher-resolution representation, improving semantic exchange between scales, or using attention to reduce irrelevant activation. These mechanisms can complement one another, but combining them can also produce a detector that is unnecessarily expensive. Parameter count and FLOPs describe model size and arithmetic work, yet they do not capture memory access, operator support on a target platform, or measured runtime. A practical accuracy-efficiency comparison should therefore report model size, batch-one latency, frames per second, and the hardware and software used for benchmarking, with accuracy broken down by target scale. Because cross-dataset and cross-device testing remains uncommon, the portability of modules tuned for a particular material, line, or camera setup is still not well established.

Architecture changes are most informative when they are tied to a diagnosed failure rather than accumulated as a collection of available modules. Scale- and geometry-specific error analysis can show where the baseline loses useful information. A P2 head is appropriate when the dominant problem is spatial quantization; feature alignment is better motivated when mismatched positions across levels drive the error; attention is most defensible when visualizations or error analysis reveal persistent responses in irrelevant background regions.

Each modification should then be tested at the intended input resolution and with batch-one inference settings. This makes it easier to separate overlapping functions and to identify both the source of the accuracy improvement and the computation required to obtain it.

3. Illumination Robustness in Industrial Visual Inspection

3.1. Sources and Effects of Illumination Interference

Lighting variation in an inspection system can arise from lamp intensity, viewing angle, surface reflectance, camera exposure, and ambient illumination, and its effect depends strongly on the material and the geometry of the workpiece. Metallic surfaces can generate intense specular highlights, curved parts often exhibit gradual brightness gradients, and colored backlighting may create halos around boundaries. Unlike electronic sensor noise, these effects leave the physical defect unchanged while altering the appearance presented to the detector. Low light reduces defect-to-background contrast, saturation under overexposure removes texture, and shadows or reflections can create edges that resemble defects. Small defects are especially vulnerable because their shape and texture may be represented by only a handful of pixels; a modest photometric disturbance can therefore remove much of the evidence needed for recognition.

Illumination affects both the input image and the internal features produced by the detector. Figure 2 groups existing remedies into two broad routes. Image-level approaches modify the camera image using operations such as brightness adjustment, Retinex-based decomposition, or learned enhancement. Feature-level approaches leave the input unchanged and instead recalibrate channels, spatial locations, or coordinate information inside the network. Representative attention mechanisms include SE, CBAM, ECA, Coordinate Attention, and SimAM [23 – 27]. Both routes aim to provide more useful features for detection, but an image that looks visually improved is not necessarily evidence of greater detection robustness. Evaluation should therefore measure the loss of detection accuracy under lighting disturbance, changes in false-positive behavior, and the additional computational cost introduced by the robustness mechanism.

Different lighting disturbances should not be collapsed into a single category of image degradation. Low illumination mainly reduces signal-to-noise ratio; overexposure clips information that cannot be recovered; cast shadows produce abrupt local intensity changes; and specular highlights create bright regions that can be mistaken for defects. Because the failure mechanisms differ, the same correction is unlikely to work equally well for all of them. Global normalization may correct a general exposure shift while making local glare more severe. Spatial attention can reduce responses to reflective regions, but only when the training data contain enough examples for the network to distinguish reflection-induced pseudo-defects from actual defects.

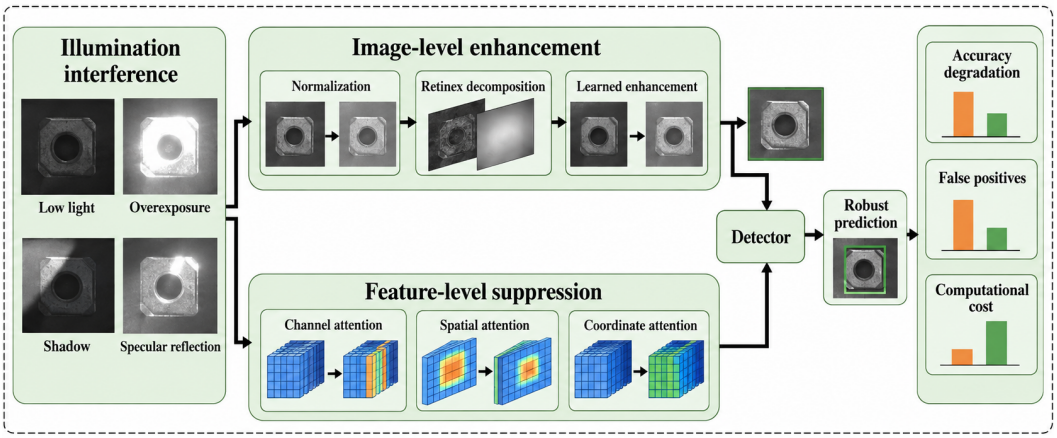

Figure 2. Image-domain enhancement and feature-domain suppression for illumination-robust industrial inspection.

3.2. Image-Level Enhancement and Normalization

Image-domain methods range from conventional histogram and gamma adjustment to learned enhancement.

Deep Retinex separates an observed image into illumination and reflectance components [28]. Zero-DCE estimates pixel-wise enhancement curves without paired low-light and normal-light training images [29], while EnlightenGAN learns an unpaired enhancement mapping through an adversarial objective [30]. Retinex-inspired unrolling incorporates a physical decomposition into a trainable optimization procedure [31], and normalizing-flow approaches represent multiple plausible mappings from a dark input to an enhanced output [32]. These methods can make defects easier to see, but better appearance does not automatically imply better detection. Enhancement may strengthen background texture, shift an apparent boundary, or introduce structures that were not present in the original image. Its value should therefore be judged by downstream performance on real inspection data rather than by visual quality alone.

When lighting changes are predictable and computing resources are limited, classical normalization can still be attractive. Gamma correction can be tuned to the camera response, while local contrast enhancement can reveal weak defects without changing the detector itself. A fixed transformation is not equally suitable for every material, however, and can distort relative contrast or create discontinuities around reflective areas. Learned enhancement adapts more flexibly but introduces a second model whose errors may propagate into detection. Joint training aligns the enhancement stage with the inspection objective, yet it can also exploit dataset-specific correlations. A controlled comparison should keep the original input available, constrain the permitted enhancement range, and evaluate original, separately enhanced, and jointly optimized pipelines under the same detection protocol.

The need for paired images is another important distinction among the methods summarized in Table 2. Supervised enhancement requires aligned captures of the same scene under low and normal illumination, which is difficult on a moving production line because object position, exposure, and reflections can change between frames. Zero-reference and unpaired approaches reduce this acquisition burden, but their objectives usually emphasize exposure, color consistency, or perceptual appearance rather than preservation of defect evidence. Task-aware constraints or downstream detection tests are therefore needed to confirm that fine boundaries and unusual textures survive the transformation. Synthetic darkening can be useful for pretraining, but final validation should rely on real images in which the camera, lighting system, and material surface interact naturally.

Table 2. Representative strategies for handling illumination variation in industrial visual inspection.

Ref.	Approach	Processing Level	Operating Principle	Key Benefit	Main Risk
-	Gamma, histogram, and color normalization	Image	Reshape intensity distribution	Low cost and straightforward deployment	Can amplify noise or reduce useful contrast
[28]	Deep Retinex	Image	Decompose illumination and reflectance	Physically interpretable decomposition	Boundary errors can carry into detection
[29]	Zero-DCE	Image	Estimate zero-reference enhancement curves	Does not require paired images	Visual improvement may not raise detection accuracy
[30]	EnlightenGAN	Image	Unpaired adversarial enhancement	Can train from unpaired data	Potential hallucination and instability
[31,32]	Retinex unrolling and normalizing flow	Image	Model-guided or probabilistic restoration	Flexible mapping for low-light inputs	Higher computation and variable appearance
[23–25]	SE, CBAM, and ECA	Feature	Channel or channel-spatial recalibration	Efficient suppression of interference	Cannot restore information lost by clipping
[26]	Coordinate Attention	Feature	Directional pooling that keeps position cues	Useful for elongated structures	Directional prior may not suit irregular defects
[27]	SimAM	Feature	Parameter-free three-dimensional weighting	No increase in learned parameters	Effectiveness depends on information retained upstream

3.3. Feature-Level Attention and Interference Suppression

SE attention uses global pooling to recalibrate channels, but the resulting summary discards explicit spatial location [23]. CBAM extends channel weighting with a spatial attention stage [24]. ECA captures local interactions between channels with relatively little additional computation [25], Coordinate Attention retains directional positional cues along horizontal and vertical axes [26], and SimAM assigns three-dimensional weights without adding trainable parameters [27]. These mechanisms can reduce responses from irrelevant regions, yet none can restore information that has been clipped by saturation or eliminated through repeated downsampling. Consequently, an improvement in average accuracy alone is not sufficient evidence of lighting robustness; disturbed-light subsets must be evaluated separately.

Where attention is inserted determines what information it can modify. Early placement operates on high-resolution texture and may suppress lighting artifacts before they propagate, but weak defects at this stage can resemble normal texture and may be suppressed as well. Deeper attention acts on more semantic features, although it cannot reconstruct detail already lost at the input or during feature reduction. Channel attention is efficient when illumination changes activate particular groups of channels, whereas spatial or coordinate mechanisms may be better suited to localized glare, directional shadows, or elongated defects. A single module placed at a technically motivated location and tested on explicit lighting subsets provides clearer evidence than repeatedly inserting the same mechanism throughout the network.

3.4. Robust Training and Evaluation

Photometric augmentation can expose the detector to changes in brightness, contrast, gamma, shadow, blur, and color during training, but the perturbation range should reflect conditions observed on the production line. Unrealistically strong random transformations may create images that the real system will never encounter and can therefore shift the learned decision boundary in an unhelpful direction. Robustness testing should keep normal light, low light, overexposure, shadow, reflection, and mixed disturbances as separate subsets instead of averaging them from the outset. Cross-camera and cross-line evaluation is also valuable when corresponding data are available. Along with absolute accuracy, reports should include degradation from the normal-light baseline, class-wise recall, false-positive rate, calibration, and any preprocessing latency.

A controlled lighting subset can be created by imaging the same product at different lamp intensities and angles while holding camera position and focus constant. Synthetic exposure changes can enlarge such a set, but they cannot reproduce every optical interaction between source, sensor, and material. Aggregating all lighting conditions can also conceal asymmetric failures: a model may improve in low light while producing more false alarms on reflective surfaces. Relative degradation, worst-condition results, and shifts in confidence calibration provide a more informative picture. Repeated trials are particularly important for rare lighting conditions because estimates from small subsets can change substantially with sample selection.

3.5. Comparative Findings and Limitations

Claims of illumination robustness are highly dependent on how the test protocol is constructed. Capturing the same item under multiple lamp intensities and angles while fixing camera position and focus isolates photometric variation more effectively than changing several factors at once. Synthetic exposure transforms can increase sample volume, but they do not fully reproduce the interaction among the light source, imaging sensor, and material surface. A single mean score over all lighting conditions is therefore difficult to interpret. Higher accuracy in dark scenes may coexist with more reflection-driven false positives. Reporting performance relative to a normal-light baseline, the worst lighting condition, and calibration changes gives a clearer account, and uncommon conditions should be repeated when sample counts are small.

Table 2 also shows that no single strategy dominates every form of lighting disturbance. Image-level processing is easy to observe and can be placed in front of different detectors, but it adds preprocessing time and may modify evidence that is relevant to the defect. Feature-level methods are trained jointly with the detector and often require less extra computation, although their behavior is less transparent and more dependent on the

training distribution. A hybrid pipeline is justified when the two components address different errors: enhancement can restore measurable contrast, while attention suppresses residual background responses. When both components improve the same subset in essentially the same way, the simpler configuration is easier to justify from an accuracy-efficiency perspective.

4. Imbalanced Learning in Industrial Visual Inspection

4.1. Forms of Imbalance

Imbalance in industrial inspection appears at several levels rather than as a single class-count problem. Normal products can vastly outnumber defective ones, and some defect categories may be orders of magnitude more frequent than others. Dense detectors introduce a second imbalance because negative locations greatly exceed positive targets within each image. Target scale adds another asymmetry: large or high-contrast defects usually generate stronger learning signals than tiny defects. These cases require different remedies. A severe normal-versus-defect imbalance can favor anomaly-detection formulations, unequal defect frequencies mainly bias classification, and foreground-background imbalance is created by the detection objective itself. For this reason, a high overall accuracy can coexist with poor recognition of rare, safety-relevant defects. Per-class recall, macro-F1, balanced accuracy, and separate head-, medium-, and tail-group results are needed to reveal that behavior.

Raw sample counts provide only a partial description of the training distribution. For each class, it is also useful to record how often it appears across images, the number of instances, typical object size, acquisition line or source, and associated lighting conditions. These views can lead to different conclusions: a category may appear in many images but contain few total instances, while another can dominate optimization by producing large numbers of positive anchors or pixels. Reviews of imbalanced industrial diagnosis distinguish sampling and data expansion from representation learning and classifier correction because they act on different mechanisms [33]. Reducing gradients from easy negatives, for example, cannot create the visual diversity that is missing from an underrepresented class.

Operational cost should also influence how imbalance is evaluated. Missing a safety-critical crack can be more serious than producing several false alarms on a cosmetic defect even if the two classes occur with similar frequency. Frequency alone therefore does not encode the consequence of an error. When the relative costs of missed detections and false alarms can be estimated, they can inform class-specific thresholds and model comparison. When such costs are uncertain, precision-recall curves at multiple operating points are more informative than one global threshold. Class-wise calibration also matters because correctly localized tail-class instances may receive systematically lower confidence and be removed before the final prediction stage.

4.2. Data-Level Strategies

At the data level, oversampling and class-aware mini-batches increase the number of times minority examples are presented during training. Repeating the same samples too often, however, can encourage overfitting. Undersampling avoids that repetition by reducing the contribution of majority classes, but it discards some of their variation. Augmentation instead expands minority diversity through geometric, photometric, or copy-paste transformations, provided that the resulting defects remain physically plausible. Generative models can extend this idea further. Fu et al. applied image generation to slab-defect recognition under small-sample and imbalanced conditions [34], while reviews of industrial imbalance treat resampling, data expansion, and feature learning as complementary tools [33]. Synthetic images should not be accepted on appearance alone; their value should be verified on a real-only test set, with duplicate checks, distributional comparison, and domain-expert review.

When the missing variability can be identified and collected, additional real data remain the most direct solution. Active learning can prioritize low-confidence outputs or production samples that differ from the current training distribution, allowing annotation to focus on unusual combinations of defect type, scale, and lighting. Synthetic data are more useful when collection is expensive or when critical failures occur too infrequently to provide enough examples. Regardless of the generation method, dataset separation must happen first: training, validation, and test partitions should be fixed before augmentation or synthesis. Otherwise, information from the test set can leak into the generation process and make the reported performance overly optimistic.

4.3. Loss Reweighting and Effective-Number Methods

Focal Loss is widely used in dense detection because it reduces the contribution of easy examples when background locations dominate the objective [35]. Category imbalance can also be handled through frequency-dependent weighting, although direct inverse-frequency weights can become unstable when a very rare class receives an excessively large gradient. Class-Balanced Loss moderates this effect by basing the weight on the effective number of samples rather than the raw count [36]. LDAM follows a margin-based approach, assigning larger classification margins to infrequent classes [37]. All of these methods redistribute the influence of samples that are already present. They do not supply visual patterns that the training set never contained and are therefore most appropriate when minority classes are reasonably diverse but underrepresented during optimization.

The selected loss should correspond to the gradient imbalance observed during training. Focal Loss is useful when a large population of easy background examples overwhelms the objective, but an overly strong focusing term can also suppress moderately difficult samples that still contain useful information. Class-balanced weights directly reflect category frequency, yet aggressive weights may amplify annotation noise in the rarest classes. Margin-based objectives enforce stronger separation around minority categories and may require a stable underlying representation. These behaviors are not easily diagnosed from one final aggregate metric. Class-specific learning curves and gradient statistics more directly reveal which samples dominate optimization and when different classes begin to diverge.

4.4. Logit Adjustment and Gradient Rebalancing

Decoupled training addresses long-tailed recognition by separating representation learning from the later rebalancing of the classifier [38]. Frequency information can also be inserted directly into the prediction objective. Balanced Meta-Softmax introduces class priors during classification [39], while logit adjustment modifies output scores at training or inference time [40]. Detection creates an additional imbalance because frequent categories and background regions can generate large negative gradients. Equalization Loss selectively weakens negative gradients that suppress rare classes [41], whereas Seesaw Loss changes the relative positive and negative contribution dynamically according to their accumulated effect [42]. Strong suppression can improve tail recall, but it can also label more background regions as rare classes. Recall should therefore be examined together with precision and the practical cost of additional false alarms.

Prior-based correction can be applied without changing the feature extractor, but it assumes that the class distribution used to estimate the priors remains relevant. A curated training set may not reflect defect prevalence on the production line, and those frequencies can change across lines or over time, causing a fixed correction to become miscalibrated. Decoupled training offers greater control by learning a general representation before classifier balancing, although the two-stage procedure introduces extra validation and model-selection decisions. Dynamic gradient methods are particularly relevant when each image produces many negative candidates. To interpret a higher tail-class recall correctly, results should also be broken down by category and target scale so that improved recognition can be distinguished from a broad rise in low-confidence detections.

4.5. Evaluation of Head, Medium, and Tail Categories

Before comparing models, the thresholds used to divide classes into head, medium, and tail groups should be stated explicitly and kept fixed across experiments. Overall mAP remains useful as a summary, but it does not show how performance is distributed across frequency groups. It should therefore be accompanied by group-specific metrics, macro-F1, balanced accuracy, and class-level precision-recall curves. A smaller difference between head and tail performance is meaningful only when the change comes from better tail recognition rather than reduced performance on common classes. Confidence scores also deserve inspection, because rare defects can be localized correctly but assigned lower scores and then removed by a single threshold applied to every class.

Class frequency is often entangled with target scale and lighting. If most examples in a tail category are small and low contrast, adding a P2 head may improve that class even when no reweighting is used. Conversely, balancing the classifier may have little effect if the feature extractor has already removed the spatial detail

required to recognize the defect. These effects can be separated by reporting head-, medium-, and tail-group results for small, medium, and large targets. When enough data exist, the same analysis should be repeated under disturbed illumination. Because tail subsets are typically small, bootstrap confidence intervals or repeated train-test splits provide more stable estimates than a single partition.

4.6. Comparative Findings and Limitations

Data-oriented and algorithm-oriented remedies target different sources of imbalance. Reweighting and logit correction can reduce classifier bias when rare classes already contain adequate visual diversity but exert too little influence during optimization. They cannot recover defect appearances that never occur in the training data; in that situation, targeted acquisition or physically constrained synthesis is required. Standard multi-class detection can also become unreliable when only a handful of examples are available. Depending on whether the problem concerns rare known classes or previously unseen defects, anomaly detection, few-shot learning, or open-set recognition may be more appropriate. Evaluation should also disentangle frequency from target scale and lighting, because a performance drop attributed to the long tail may actually be caused by the concentration of small or poorly illuminated instances in minority classes.

The strategies in Table 3 can therefore be matched to the mechanism that produces the imbalance. Class-aware sampling, loss weighting, margin adjustment, and logit correction are suitable when minority examples are varied but too infrequent to shape learning sufficiently. When intra-class diversity is itself limited, new real observations or physically plausible synthesis are needed instead. Focal Loss and gradient-balancing approaches are more directly relevant when the objective is dominated by large numbers of negative candidates. In every case, class-specific error analysis remains necessary. If several configurations raise rare-class recall, the simplest configuration that preserves precision, calibration, and head-class performance on a real-only test set offers the clearest trade-off.

Table 3. Representative approaches to imbalance in industrial visual inspection.

Ref.	Approach	Imbalance Source	Additional Data	Primary Benefit	Suggested Evaluation
-	Oversampling and class-aware batching	Unequal class frequency	No	Raises exposure of minority examples	Macro-F1 plus per-class recall
-	Geometric, photometric, and copy-paste augmentation	Minority diversity and scale	No external data	Expands data in a task-specific way	Tail recall on a real-only test set
[34]	Synthetic defect generation	Small-sample imbalance	Generated samples	Broadens appearance diversity	Real-only macro-F1 and expert assessment
[35]	Focal Loss	Foreground-background and easy-hard imbalance	No	Limits domination by easy negatives	mAP, PR curves, and false-positive rate
[36,37]	Class-Balanced Loss and LDAM	Class frequency and margin imbalance	No	Frequency-aware optimization	Balanced accuracy and head-tail results
[38]	Decoupled training	Classifier bias	No	Separates feature learning from rebalancing	Head / medium / tail performance
[39,40]	Balanced Softmax and logit adjustment	Class-prior imbalance	No	Simple prior-aware correction	Calibration and class-wise precision
[41,42]	Equalization and Seesaw losses	Dominant negative gradients	No	Reduces suppression of rare classes	Tail AP, head-tail gap, and alarm burden

5. Conclusions and Future Directions

5.1. Main Findings

This review treats small-target representation, illumination variability, and data imbalance as interconnected

causes of failure rather than isolated design problems. For small defects, robust detection depends on retaining sufficiently fine spatial features and moving detail across network levels through bidirectional or cross-level fusion. Attention can reduce background activation, but its benefit depends on where it is inserted and on how much useful information has survived earlier processing. Lighting robustness requires a balance between correcting the input and suppressing interference inside the learned representation. In imbalanced datasets, expanding the visual diversity of minority classes addresses a different limitation from correcting optimization bias, so the two approaches can be complementary. Reported accuracy improvements should be interpreted together with latency, memory demand, false-alarm cost, and the degree to which the evaluation setup reflects production conditions.

5.2. Relationships among the Three Challenges

Object scale, illumination, and class frequency interact throughout the inspection pipeline. Tiny defects are particularly sensitive to lighting because only a few pixels carry their structure, and tail categories can become disproportionately difficult when they contain more small or low-contrast examples than common classes. A method introduced for one issue can also alter another: image enhancement may shift the feature distribution of minority classes, while high-resolution prediction increases memory and computation. These interactions make ablation results harder to attribute. An apparent gain from class reweighting may partly depend on the scale profile of the affected class, and a gain credited to attention may actually reflect a change in cross-scale fusion. Reporting results jointly by target size, lighting condition, and class frequency is therefore important for identifying the mechanism behind an improvement.

5.3. Future Research Directions

Future benchmarks should annotate target size, lighting condition, and class frequency together so that the three factors can be analyzed instead of averaged away. Evaluation protocols should include controlled lighting subsets, cross-camera and cross-line tests, and separate head-, medium-, and tail-class metrics. Synthetic defects should obey physical constraints and should be judged only on real test data. Open-set detection is also needed when the deployment environment can produce defect types that were absent from training. Practical deployment studies should report hardware-specific batch-one latency, memory use, energy consumption, and runtime stability rather than relying only on parameter counts or FLOPs. Lightweight backbones such as MobileNetV2, ShuffleNetV2, and GhostNet remain useful starting points [43–45], while real-time transformer detectors provide another region of the accuracy-efficiency design space [46].

5.4. Conclusions

Industrial visual inspection is increasingly being evaluated as a system-level robustness problem rather than as an exercise in maximizing one aggregate accuracy score. Meaningful progress will depend on coordinated treatment of target scale, photometric variation, category imbalance, and deployment cost. Transparent protocols, controlled comparisons, and architectures chosen for diagnosed failure modes are likely to provide more reliable advances than simply stacking additional modules.

Funding

This research was funded by Yongchuan District Science and Technology Projects, grant number (2024yc-jbgs20012).

Author Contributions

Writing—original draft, S.L. and J.X.; writing—review and editing, S.L. and J.X. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1 Czimmermann T, Ciuti G, Milazzo M, *et al.* Visual-Based Defect Detection and Classification Approaches for Industrial Applications—A Survey. *Sensors* 2020; **20**: 1459. https://doi.org/10.3390/s20051459.

2 Yang J, Li S, Wang Z, *et al.* Using Deep Learning to Detect Defects in Manufacturing: A Comprehensive Survey and Current Challenges. *Materials* 2020; **13**: 5755. https://doi.org/10.3390/ma13245755.

3 Gao Y, Li X, Wang XV, *et al.* A Review on Recent Advances in Vision-Based Defect Recognition towards Industrial Intelligence. *Journal of Manufacturing Systems* 2022; **62**: 753–766. https://doi.org/10.1016/j.jmsy.2021.05.008.

4 Jha SB, Babiceanu RF. Deep CNN-Based Visual Defect Detection: Survey of Current Literature. *Computers in Industry* 2023; **148**: 103911. https://doi.org/10.1016/j.compind.2023.103911.

5 Liu Y, Zhang C, Dong X. A Survey of Real-Time Surface Defect Inspection Methods Based on Deep Learning. *Artificial Intelligence Review* 2023; **56**: 12131–12170. https://doi.org/10.1007/s10462-023-10475-7.

6 Ma Y, Yin J, Huang F, *et al.* Surface Defect Inspection of Industrial Products with Object Detection Deep Networks: A Systematic Review. *Artificial Intelligence Review* 2024; **57**: 333. https://doi. org/10.1007/s10462-024-10956-3.

7 Ren S, He K, Girshick R, *et al.* Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2017; **39**: 1137–1149. https://doi.org/10.1109/tpami.2016.2577031.

8 Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. https://doi.org/10.1109/cvpr.2016.91.

9 Lin TY, Dollar P, Girshick R, et al. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. https://doi.org/10.1109/cvpr.2017.106.

10 Liu S, Qi L, Qin H, et al. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. https://doi.org/10.1109/cvpr.2018.00913.

11 Tan M, Pang R, Le QV. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. https://doi.org/10.1109/cvpr42600.2020.01079.

12 Huang S, Lu Z, Cheng R, *et al.* FaPN: Feature-Aligned Pyramid Network for Dense Image Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. https://doi.org/10.1109/iccv48922.2021.00090.

13 Zhang Y, Sun Y, Huang Q, *et al.* Cross-Level Multiscale Feature Fusion Enhanced YOLO with Focused-CIoU for Detection of Small and Extreme Aspect Ratio Defects. *Measurement Science and Technology* 2026; **37**: 025401. https://doi.org/10.1088/1361-6501/ae26a5.

14 Ren J, Li Z, Zhou Z, *et al.* CFL-YOLO: A Modified YOLOv8 for Detection of Steel Surface Defects. *Measurement* 2026; **266**: 120509. https://doi.org/10.1016/j.measurement.2026.120509.

15 Li G, Gan Y, Zhang W, *et al.* GS-YOLO: A Lightweight and High-Performance Method for PCB Surface Defect

Detection. *Expert Systems with Applications* 2026; **303**: 130583. https://doi.org/10.1016/j.eswa.2025.130583.

16 Li J, He X, Chen X, *et al.* HDFA-YOLO: A Real-Time Steel Surface Defect Detection Model Based on Backbone Lightweight Design and Multi-Scale Feature Fusion. *Measurement* 2026; **258**: 119390. https://doi.org/10.1016/j.measurement.2025.119390.

17 Li J, Kang X. Mobile-YOLO: An Accurate and Efficient Three-Stage Cascaded Network for Online Fiberglass Fabric Defect Detection. *Engineering Applications of Artificial Intelligence* 2024; **134**: 108690. https://doi.org/10.1016/j.engappai.2024.108690.

18 Wang Z, Zhou W, Li Y. MSAF-YOLO: An Efficient Multi-Scale Attention Fusion Network for High-Precision Steel Surface Defect Detection. *Measurement* 2026; **257**: 118640. https://doi. org/10.1016/j.measurement.2025.118640.

19 Qingyi W, Lingyun G, Bo C. The Lightweight CS-YOLO Model Applied for Concrete Crack Segmentation. *Expert Systems with Applications* 2025; **277**: 127298. https://doi.org/10.1016/j.eswa.2025.127298.

20 Yang R, Zhang H. Tunnel-YOLO: An Improved You Only Look Once Algorithm for Real-Time Shield Tunnel Lining Leakage Detection. *Engineering Applications of Artificial Intelligence* 2025; **162**: 112403. https://doi.org/10.1016/j.engappai.2025.112403.

21 Wu P, Xu Y, Ma Y, *et al.* LYA-YOLO: A Lightweight and Accurate YOLO Model in Drone Aerial Image Scenes. *Expert Systems with Applications* 2026; **321**: 132166. https://doi.org/10.1016/j.eswa.2026.132166.

22 Iqra, Giri KJ. SO-YOLOv8: A Novel Deep Learning-Based Approach for Small Object Detection with YOLO beyond COCO. *Expert Systems with Applications* 2025; **280**: 127447. https://doi.org/10.1016/j.eswa.2025.127447.

23 Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. https://doi.org/10.1109/cvpr.2018.00745.

24 Woo S, Park J, Lee JY, *et al.* CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018, Proceedings the 15th European Conference, Munich, Germany, 8 – 14 September 2018*; Springer International Publishing: Cham, Switzerland, 2018. https://doi.org/10.1007/978-3-030-01234-2_1.

25 Wang Q, Wu B, Zhu P, *et al.* ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. https://doi.org/10.1109/cvpr42600.2020.01155.

26 Hou Q, Zhou D, Feng J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20 –25 June 2021. https://doi.org/10.1109/cvpr46437.2021.01350.

27 Yang L, Zhang RY, Li L, *et al.* SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 11863–11874.

28 Wei C, Wang W, Yang W, *et al.* Deep Retinex Decomposition for Low-Light Enhancement. In Proceedings of the British Machine Vision Conference (BMVC 2018), Tyne, UK, 3–6 September 2018.

29 Guo C, Li C, Guo J, *et al.* Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. https://doi.org/10.1109/cvpr42600.2020.00185.

30 Jiang Y, Gong X, Liu D, *et al.* EnlightenGAN: Deep Light Enhancement without Paired Supervision. *IEEE Transactions on Image Processing* 2021; **30**: 2340–2349. https://doi.org/10.1109/tip.2021.3051462.

31 Liu R, Ma L, Zhang J, *et al.* Retinex-Inspired Unrolling with Cooperative Prior Architecture Search for Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. https://doi.org/10.1109/cvpr46437.2021.01042.

32 Wang Y, Wan R, Yang W, *et al.* Low-Light Image Enhancement with Normalizing Flow. P*roceedings of the AAAI Conference on Artificial Intelligence* 2022; 36: 2604–2612. https://doi.org/10.1609/aaai.v36i3.20162.

33 Ren Z, Lin T, Feng K, *et al.* A Systematic Review on Imbalanced Learning Methods in Intelligent Fault Diagnosis. *IEEE Transactions on Instrumentation and Measurement* 2023; **72**: 1–35. https://doi.org/10.1109/tim.2023.3246470.

34 Fu T, Li P, Liu S. An Imbalanced Small Sample Slab Defect Recognition Method Based on Image Generation. *Journal of Manufacturing Processes* 2024; **118**: 376–388. https://doi.org/10.1016/j.jmapro.2024.03.028.

35 Lin TY, Goyal P, Girshick R, *et al.* Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. https://doi.org/10.1109/iccv.2017.324.

36 Cui Y, Jia M, Lin TY, *et al.* Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. https://doi.org/10.1109/cvpr.2019.00949.

37 Cao K, Wei C, Gaidon A, *et al.* Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.

38 Kang B, Xie S, Rohrbach M, *et al.* Decoupling Representation and Classifier for Long-Tailed Recognition. In Proceedings of the International Conference on Learning Representations (ICLR 2020), Virtual, 26 April–1 May 2020.

39 Ren J, Yu C, Sheng S, *et al.* Balanced Meta-Softmax for Long-Tailed Visual Recognition. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2020), Virtual, 6 – 12 December 2020; pp. 4175–4186.

40 Menon AK, Jayasinghe S, Rawat AS, *et al.* Long-Tail Learning via Logit Adjustment. In Proceedings of the International Conference on Learning Representations (ICLR 2021), Virtual, 3–7 May 2021.

41 Tan J, Wang C, Li B, *et al.* Equalization Loss for Long-Tailed Object Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13 – 19 June 2020. https://doi.org/10.1109/cvpr42600.2020.01168.

42 Wang J, Zhang W, Zang Y, *et al.* Seesaw Loss for Long-Tailed Instance Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20 –25 June 2021. https://doi.org/10.1109/cvpr46437.2021.00957.

43 Sandler M, Howard A, Zhu M, *et al.* MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. https://doi.org/10.1109/cvpr.2018.00474.

44 Ma N, Zhang X, Zheng HT, *et al.* ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Computer Vision—ECCV 2018, Proceedings of the 15th European Conference, Munich, Germany, 8 – 14 September 2018*; Springer International Publishing: Cham, Switzerland, 2018. https://doi.org/10.1007/978-3-030-01264-9_8.

45 Han K, Wang Y, Tian Q, et al. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13 – 19 June 2020. https://doi.org/10.1109/cvpr42600.2020.00165.

46 Zhao Y, Lv W, Xu S, *et al.* DETRs Beat YOLOs on Real-Time Object Detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 16 – 22 June 2024. https://doi.org/10.1109/cvpr52733.2024.01605.