machines

MDPI

Article

Beyond Single-Run Improvements: Controlled Multi-Seed Evaluation of Augmentation and Training Modifications for Maritime Object Detection

Jiashu Liu [1,2] and Ciyong Luo [3,*]

[1] Chongqing University-University of Cincinnati Joint Co-op Institute, Chongqing University, Chongqing 400044, China; liu5ju@mail.uc.edu
[2] Department of Electrical and Computer Engineering, College of Engineering and Applied Science, University of Cincinnati, Cincinnati, OH 45221, USA
[3] State Key Laboratory of Power Transmission Equipment Technology, Chongqing University, Chongqing 400044, China
* Correspondence: luociyong@cqu.edu.cn

Abstract

Small gains in maritime object detection require repeated training and matched controls. We evaluated four detectors on SeaDronesSee (SDS), WSODD, and FloW, and examined training modifications using paired seeds, matched clean copies or views, and five synthetic degradation levels. SDS provides auxiliary validation-based evidence; WSODD and FloW provide held-out test evidence. Across six WSODD seeds, fixed single-view augmentation changed mean average precision (mAP50–95) by −0.10 percentage points (pp), with individual changes from −1.81 to +1.48 pp. Matched comparisons found no statistically supported gain attributable to degraded content under normal inputs, and the complete configuration showed no detectable difference from the scratch reference. Under the strongest synthetic low light, fixed single-view augmentation yielded +4.99 pp, positive in all six seeds. This benefit was not reproduced on the real low-luminance subset. Overall, no tested modification produced a statistically supported overall-accuracy improvement under normal inputs. Reliable assessment requires paired uncertainty estimates, matched controls, and explicit separation of synthetic robustness from evidence of real-world benefit.

Keywords: maritime object detection; data augmentation; training seeds; matched controls; small objects; illumination

Check for updates

Academic Editor: Dan Zhang

Received: 2 October 2026
Revised: 12 November 2026
Accepted: 17 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Water-surface object detection supplies the visual information used in unmanned surface vehicle obstacle avoidance, unmanned aerial vehicle (UAV) search and rescue, and floating waste cleanup. The three public datasets used in this study correspond to these tasks: SeaDronesSee (SDS) records people in the water and related targets in search-and-rescue scenes from the air, WSODD records 14 classes of water-surface objects across different water bodies, weather conditions, and capture times, and FloW records floating waste in inland rivers from an unmanned surface vehicle perspective [1–3]. At long range or from high-altitude viewpoints, key targets may occupy only a few pixels, and surface glare and lighting changes can alter the appearance of both targets and background.

Existing methods improve water-surface detection through changes to feature extraction, localization loss, and training images. WA-YOLO examines attention, localization loss, input resolution, and illumination enhancement, together [4]; Yalçın and Ekenel compare

images containing surface reflections with reflection-inpainted images, and design a filtering method to reduce reflection-related false detections [5]. Because these operations act on different stages, the overall result of a combined scheme cannot show how much each individual change contributes.

Repeated training changes judgments about small improvements. Bouthillier et al. trace fluctuations in performance comparisons to sources such as data sampling, parameter initialization, and hyperparameter selection [6]; Reimers and Gurevych show that reporting score distributions can change the method ranking, based on a single result [7]. The paired WSODD results in this paper illustrate the same risk: the same change can appear as −1.810 percentage points (pp) or +1.480 pp on individual seeds, while the six-seed average change is only −0.099 pp. The first core question is therefore not whether a given run improves performance, but how large the average effect is, relative to seed-to-seed variation.

The augmentation experiment also changes image appearance and the number of exposures of the selected samples. Research by Hoffer et al. on repeated augmented samples within a batch shows that the repetition pattern affects gradient statistics and training outcomes [8]. Accordingly, this paper treats the matched control as a second, parallel core component: the clean copies and the GlareSoft copies use the same images, and are matched in number, labels, and processing pipeline, so as to estimate the increment from the localized-glare transformation relative to additional clean copies; the difference between the complete configuration and the scratch reference retains the total effect of the entire operation chain.

Figure 1 summarizes the research design across SDS, WSODD, and FloW: paired multi-seed experiments quantify the average effect and seed-to-seed variation in augmentation and paired-view training, while matched controls address what multi-seed evaluation alone cannot, namely, which operation in the chain produces an observed change. Seven from-scratch training configurations on WSODD compare operations step by step along that chain, and separate attention from localization loss; clean copies versus GlareSoft copies, which differ only in the content of the transformation, are compared again under pretrained fine-tuning; and, on FloW, the identical paired-view control holds the number of views, update count, and loss normalization constant, while replacing only the content of the second view. Evaluation on normal inputs and under five severity levels of synthetic low light and localized glare yields condition interaction terms that identify the input condition under which a change takes effect, and the inference cost of four reference detectors is measured on an RTX 4090. Training-seed variation can far exceed the average improvement; where the net effect of a complete configuration is near zero, stepwise controls reveal changes that would be hidden in the net mean; and fixed single-view augmentation shows no statistically supported overall-accuracy gain on normal inputs, with a benefit under strong synthetic low light that was not reproduced on the real low-luminance subset.

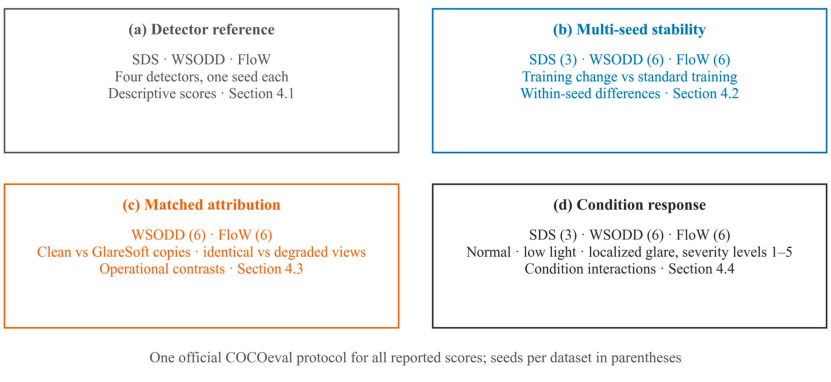


Figure 1. Study design and experimental coverage. Four analyses successively compare detectors, the stability of repeated training, operation contributions under matched controls, and responses to test

conditions. The stability analysis uses paired multi-seed results under normal inputs; the attribution analysis controls the number of copies, source images, and processing pipeline, and, on FloW, it also controls the number of views and the number of updates; the low-light and localized-glare evaluations fix the model, and vary only the input conditions.

2. Related Work

2.1. Maritime Object Detection and Benchmarks

SeaDronesSee records people and related targets in search-and-rescue scenes from the air, covering different flight altitudes and viewpoints [1]; the v2 version used by MaCVi 2023 extends image and category coverage [9]. FloW mounts the camera on a cleaning unmanned surface vehicle facing the water ahead, where small floating waste appears alongside water surface reflections and reflections of shore objects [3]. WSODD records multiple types of surface targets across different water areas, weather conditions, and times of capture [2].

MODS examines object detection and obstacle segmentation together, evaluating detection performance separately in the full region and in the hazardous region near the hull [10]. The MaCVi 2023 public challenge combined a SeaDronesSee-based UAV task with a MODS-based unmanned surface vessel task, setting separate evaluation requirements for detection, tracking, and segmentation [9]. Miri Rekavandi et al. reviewed small-object detection in images and video, and linked general methods to water-surface monitoring tasks [11].

LaRS provides panoramic annotations across diverse water-surface scenes and environmental conditions [12]. SafeSea uses a diffusion model to replace the sea-surface background, then checks whether the generated images retain the sea state and vessels [13]. For maritime obstacle segmentation, Zhang et al. present an augmentation approach constrained jointly by sample quality and coverage [14].

2.2. Detectors and Common Improvement Strategies

RT-DETR processes multi-scale features with a hybrid encoder and supplies the decoder with initial queries through query selection, enabling end-to-end detection without non-maximum suppression [15]. YOLO26, likewise, adopts a non-maximum-suppression-free inference design, and its technical report further describes label assignment and training changes aimed at small objects [16].

CBAM applies channel attention followed by spatial attention to reweight the intermediate features [17]. Focal-EIoU accounts for overlap, center position, and edge-length differences, while up-weighting high-quality bounding-box matches [18]. In water-surface scenes, WA-YOLO evaluates these components, together with illumination enhancement and an input resolution suited to small targets [4]. Work on surface reflection treats reflection-induced false detections as a separate target for improvement [5].

Zhang et al. incorporated the Ghost module and Transformer into YOLOv5 and evaluated ship detection accuracy, model size, and speed together [19]. In image classification of livestock skin diseases, classification accuracy was reported alongside GFLOPs and FPS [20]. Zoph et al. learned transformation policies for object detection, showing that the choice of augmentation can, itself, change detection performance [21]. Li et al. used Feedback Small-Cut&Paste in S3Det to add small-ship objects to training images, deciding whether to apply the operation according to the proportion of small-object loss [22].

2.3. Repeated Training and Analysis of Augmentation Effects

Bouthillier et al. examined initialization, data order, augmentation, sampling, and hyperparameter selection separately, and found that each of these sources of variability

affects performance comparisons [6]. Reimers and Gurevych used sequence-labeling experiments to show that score distributions can change the ranking of methods [7]. Summers and Dinneen further found that even tiny parameter perturbations can amplify during optimization and produce different results [23].

Hoffer et al. found that transformed copies of the same sample produce correlated gradients, and an augmentation benefit remains observable after sample exposure is matched [8]. The relationship between augmentation and the original data can also be described in terms of affinity and diversity, which jointly affect augmentation performance [24]. Work distinguishing data expansion, invariance, and additional training randomness further links image variation to the optimization process [25]. Michaelis et al. evaluated normal and degraded images separately in object detection [26].

AugMix improves classifier performance under common corruptions by mixing image transformations with a consistency constraint [27]. Studies of training randomness further show that software tools and execution environments can affect run outcomes [28]. Dodge et al. link validation performance to the computational budget used in model development [29], and Pineau et al. summarize reproducible practices for sharing code and experimental details [30]. These studies address run variability, augmentation mechanisms, and corruption evaluation separately, whereas improved evaluation for water-surface detection still requires combining multi-seed repetition with matched operation-level attribution.

3. Materials and Methods

3.1. Datasets and Fixed Partitions

Three public datasets anchor the evaluation: SeaDronesSee (SDS) for maritime search and rescue, WSODD for multi-class water-surface detection, and FloW for inland floating-waste detection. SDS defines five target classes [1,9,31]; this study's source-disjoint split contains 8922 training and 1555 validation images, and both model selection and the results reported here use that validation set, so SDS supplies auxiliary validation evidence. WSODD retains 14 target classes [2], and this study's source-disjoint split comprises 3977 training, 1084 validation, and 2406 test images. The complete revised FloW collection used in this study retains all 2000 images and 5272 annotated bounding boxes, divided into 1124 training, 335 validation, and 541 test images with a single floating-waste class [3]. The held-out test sets of WSODD and FloW provide the primary test evidence. Table 1 summarizes the evaluation splits and experimental coverage, and Figure 2 shows annotated examples from SDS and the target-size distributions of the three evaluation sets.

Table 1. Dataset partitions and coverage of the principal experiments. SDS is evaluated on its validation partition; WSODD and FloW use fixed test partitions. Detector comparisons use seed 0. The principal training strategies, matched controls, and condition evaluations use seeds 0, 1, 2, 42, 123, and 3407 on WSODD and FloW, and seeds 0, 1, and 2 on SDS. The auxiliary pretrained single-component comparisons on FloW use three seeds (Section 3.5).

Dataset	Classes	Train/Val/Test	Evaluation	Seeds	Matched Controls
SDS	5	8922/1555/—	Validation	3	—
WSODD	14	3977/1084/2406	Test	6	Seven configurations trained from scratch; clean and GlareSoft copies under pretrained fine-tuning
FloW	1	1124/335/541	Test	6	Identical paired views

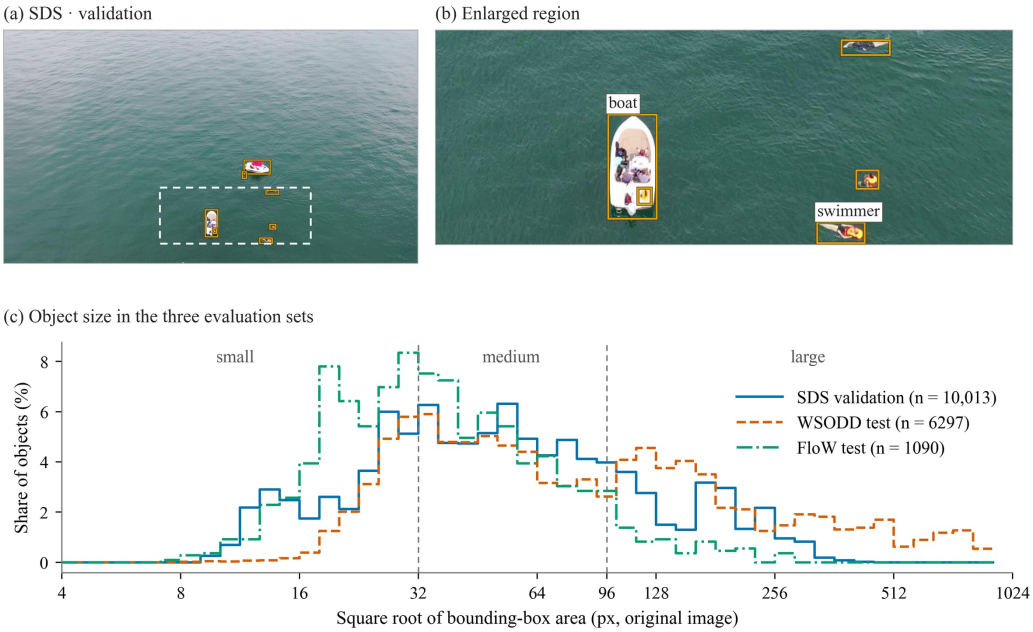


Figure 2. An annotated example from SDS and the object-size distributions of the three evaluation sets. (**a**) An SDS validation image with its ground-truth annotations; (**b**) a magnified view of the dashed region in the same image, where orange boxes mark ground-truth objects and some objects are labeled with their class names; (**c**) distributions of object sizes in the SDS validation set, the WSODD test set, and the FloW test set, with the square root of bounding-box area on the horizontal axis and dashed lines marking the 32 px and 96 px COCO boundaries between small, medium, and large objects.

Partitions keep acquisition groups together. For SDS, drone, source-directory, and video identifiers define 39 source groups: 29 are assigned to training and 10 to validation. Still images without video identifiers are grouped by acquisition directory. WSODD uses 147 acquisition-sequence groups, with 89, 26, and 32 assigned to training, validation, and testing. FloW is grouped using traceable directories, original paths, and identified continuous acquisitions, with known timestamps used to order the partitions. Group assignments and image manifests are fixed before training, and every comparison within an experimental block uses the same partition.

The three evaluation sets differ in the size distribution of their ground-truth objects. Under the COCO area ranges, small, medium and large objects account for 29.8%, 47.8% and 22.4% of the 10,013 objects in the SDS validation set, 17.9%, 40.3% and 41.8% of the 6297 objects in the WSODD test set, and 46.3%, 46.8% and 6.9% of the 1090 objects in the FloW test set. Of the 1126 small objects in WSODD, 1101 fall within the 16^2–32^2 px^2 range and concentrate in the ball (523) and boat (362) classes, whereas 2278 of the 2980 small objects in SDS belong to the swimmer class. The same APsmall, therefore, corresponds to different object types across the three datasets. Per-class size composition is given in Table S5.

3.2. Detectors and Descriptive Reference

The detector reference contains one seed-0 run for each model on each dataset. The models are YOLOv8s, YOLO11s, YOLO26s, and RT-DETR-L [15,16,32,33]. We loaded their general-purpose pretrained weights through Ultralytics 8.4.124 and ran 150 epochs with an input-size setting of 960, batches of eight, and patience 150. All four detectors used AdamW (momentum 0.9; weight decay 0.0005), with starting learning rates of 0.001111 for SDS,

0.000556 for WSODD, and 0.002 for FloW. The three YOLO models used automatic mixed precision; RT-DETR-L used FP32.

The detector reference uses each model's best.pt checkpoint, selected by mean average precision (mAP50–95) during native validation. All evaluations use an input-size setting of 960, an actual inference batch size of one, FP32, and have rectangular auto-padding disabled. The confidence threshold is 0.001, with, at most, 300 predictions per image. Detectors that use non-maximum suppression retain an IoU threshold of 0.7; each architecture keeps its native prediction and postprocessing procedure. The official COCOeval scores detections in original-image coordinates with maxDets set to (1, 10, 100), using, at most, 100 detections per image and category for AP. The same scoring and inference settings apply to the other experimental blocks, as described in Section 3.7.

Computational efficiency was measured on an NVIDIA GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) using the four reference models from WSODD, each run in a separate process. Inputs were set to 960 × 960, FP32, batch 1, with Ultralytics 8.4.124 and PyTorch 2.5.1+cu124, and TF32 disabled. Each model was warmed up on 50 images, after which the first 500 decoded images were timed in the fixed order of the test list. CUDA was synchronized before and after each measurement; timing covered preprocessing, host-to-device transfer, forward computation, and postprocessing, but excluded disk reading and image decoding. FPS is the reciprocal of mean latency, and memory is the peak PyTorch allocation for single-image inference after warm-up. Parameter counts and GFLOPs were computed on the unfused training models, with GFLOPs obtained using the Ultralytics estimator.

3.3. Training Strategies and Model Components

Standard training follows the YOLOv8s procedure specified for each experimental block. Fixed single-view augmentation applies maritime degradation generation (MDG) to the resulting augmented input, while retaining one detection view. We fixed the MDG assignment for each SDS and WSODD image so that changing the training seed or batch order, or advancing to another epoch, would reuse its branch, parameter band, and numerical parameters. An independent random-number generator produces this assignment, seeded with the SHA-256 digest of the dataset namespace and image identifier. It selects low light with probability 0.20, glare with 0.20, and their combination with 0.15. The remaining 0.45 preserves the input after standard augmentation. The three parameter bands have equal sampling probability, and the original labels are retained.

MDG produces low-light images through brightness scaling and a gamma transform, and localized glare through a bright elliptical patch with smooth attenuation. For a normalized pixel x, the low-light transform is $b \cdot x^{\gamma}$, with brightness factor b ranging from 0.82 to 0.95 and γ from 1.05 to 1.30. Glare has a target area of 1.75–5% of the image, an aspect ratio of 1.4–3.2, and an additive intensity of 22–48 on a 255-level scale. The mixed branch applies low light before adding glare, and clips pixels to their valid range. The training parameter bands and ellipse-generation rules are specified in Appendix A.

In paired-view training on FloW, each standard-augmented image is concatenated with a corresponding view generated by online stochastic MDG, and the same detector is trained on both views, so a batch of eight source images yields 16 detector inputs with duplicated annotations. The small-object-weighted variant, additionally, assigns a weight of 1.5 to ground-truth objects in the degraded view whose training-augmented area is below 1024 px^2, and the assigned positive anchors inherit this weight in the bounding-box, classification, and distribution focal loss (DFL) terms; other objects, background anchors, and the original normalization denominator keep their existing weights. This training threshold equals the COCO small-object upper bound used in evaluation ($32^2 = 1024$ px^2),

but the two are defined in different coordinate frames: the training threshold is applied to training-augmented area at the 960 input scale, whereas APsmall is computed from the original image pixel area. FloW images are 1280 pixels wide, and rescaling to 960 reduces object area to 0.5625 times its original value, so the training threshold corresponds to about 1820 px^2 in the original image, covering all COCO small objects and the smallest subset of COCO medium objects. FloW consists mostly of small objects: 1559 of the 2665 training-set objects (58.5%) are COCO small objects, and rescaling alone places 1903 (71.4%) below the training threshold; training logs from six seeds show that 75.6–75.7% of object instances in the degraded view were weighted after training augmentation. Of the 1090 test-set objects, 695 fall below the threshold under the same conversion, including all 505 COCO small objects and 190 COCO medium objects. Both configurations use only the detection loss, and contain no cross-view consistency term. Figure 3 compares the single-view and paired-view training paths.

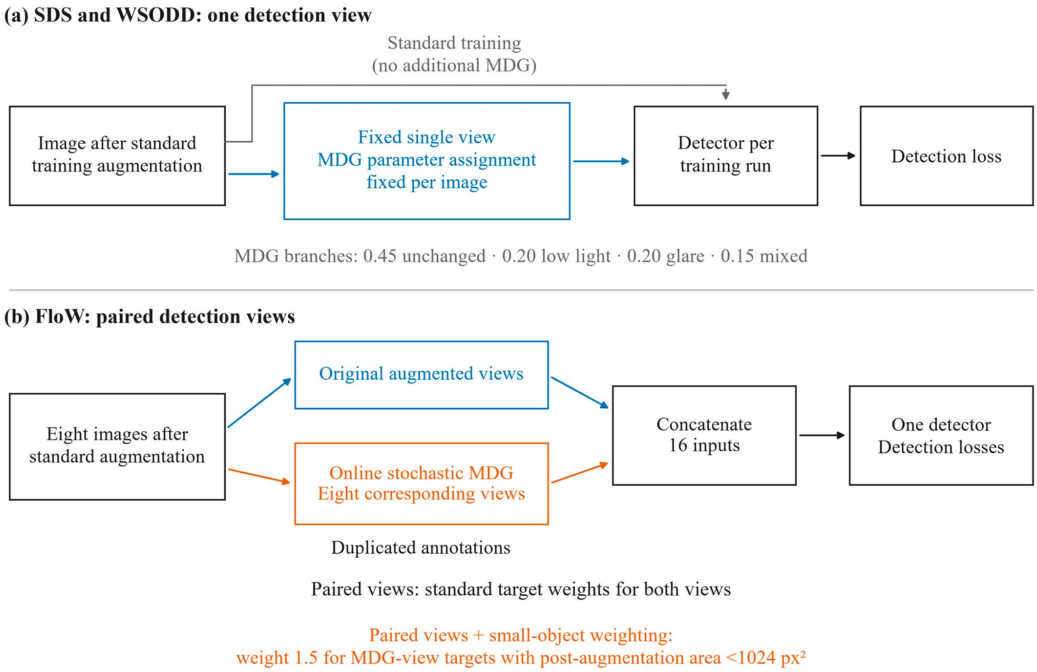


Figure 3. View construction and detection-loss paths. (**a**) Fixed single-view augmentation on SDS and WSODD applies image-specific fixed MDG after standard training augmentation, retaining one detection view; standard training uses the standard augmented input. (**b**) Both paired-view variants on FloW concatenate the original augmented views with online stochastic MDG views for the detector in each run. The small-object-weighted variant weights targets with a post-augmentation training area below 1024 px^2 in the MDG view by 1.5. The illustrated batch expands eight source images to 16 detection inputs.

The structural modification replaces the C2f modules at the P3 and P4 levels of the YOLOv8s backbone with C2f_CBAM modules, applying channel attention followed by spatial attention after each C2f output [17]. The implementation used is Ultralytics' built-in simplified CBAM: channel attention relies on global average pooling alone, passing the result through a 1 × 1 convolution without dimensionality reduction and a sigmoid to obtain channel weights, whereas the original CBAM uses both average-pooling and max-pooling descriptors, each passed through a shared multilayer perceptron with dimensionality reduction and then summed. Spatial attention follows the original: mean and maximum values are taken along the channel dimension and concatenated, then passed through a 7 × 7 convolution and a sigmoid. The remaining backbone layers, neck, and detection

head retain the YOLOv8s structure. This architecture is used both in the attention and localization-loss configuration and in its two variants, the clean copy and the GlareSoft copy.

For the localization loss, a variant of Focal-EIoU [18] replaces the CIoU term while DFL is retained. Like Focal-EIoU, the variant multiplies the EIoU loss by a weight that varies with IoU, though the form of the weight differs from the original. Equation (1) gives the implementation: EIoU subtracts from IoU three penalties for the center distance, width difference, and height difference between the predicted and ground-truth boxes, each normalized by the squared diagonal, width, and height of the smallest enclosing box, respectively; EIoU is clipped to the range −1 to 1, and the loss is 1 minus the clipped EIoU, multiplied by the weight $0.25 + 0.75(1 - \mathrm{IoU})^2$. The numerical stability constant is 10^{-7}. The original weight is IoU raised to the power γ; here, the weight increases as IoU decreases, taking the value 0.25 at IoU = 1 and 1 at IoU = 0, so poorly overlapping samples receive larger weights. The loss form and parameters follow the original implementation described in Section 3.5, and no tuning of them was performed in this study. The attention and localization-loss configuration and its two copy variants use the same components, so the matched localized-glare comparison changes only the content of the copies.

$$\mathrm{L} = \left[0.25 + 0.75(1 - \mathrm{IoU})^2\right] \left[1 - \mathrm{clip}(\mathrm{EIoU}, -1,\ 1)\right] \tag{1}$$

The effect of clipping and the IoU-dependent weight on gradients was checked in two steps, unit tests, and statistics from real training batches. When EIoU falls below −1, clipping makes the EIoU term constant, so gradients pass only through the IoU-dependent weight; if the predicted and ground-truth boxes do not overlap at all, the localization gradient for that sample is zero. In automatic differentiation tests on concentric square boxes, the predicted box entered the clipped region once its side exceeded about 3.73 times that of the ground-truth box; at 4 times, the gradient norm was about 30% of the unclipped form, and at 8 times about 16%. On WSODD training batches, 31.2% of positive samples fell in the clipped region during the first 150 iterations from random initialization, accounting for 43.4% of the localization loss; with the end-of-training checkpoints of three seeds, statistics collected over 50 iterations each at zero learning rate showed the clipped proportion at 0.35–0.51%, accounting for 8.2–8.8% of the localization loss. No positive sample with zero gradient appeared in either stage. At the end of training, the mean IoU of positive samples was 0.82–0.83 and the mean IoU-dependent weight was about 0.29. Clipping, therefore, acts mainly on low-quality predictions early in training and, after convergence, involves less than 1% of positive samples.

3.4. Repeated Training and Paired Comparisons

The stability analyses for WSODD and FloW use six paired seeds: 0, 1, 2, 42, 123, and 3407, and all seeds are included in the summaries. WSODD compares standard training with fixed single-view augmentation; FloW compares standard training, paired-view training, and small-object-weighted paired-view training. For SDS, standard training and fixed single-view augmentation use seeds 0, 1, and 2. Independent training in every case begins from the pretrained weights specified for the experimental block to which the run belongs. Differences are computed only within the same dataset, experimental block, and seed.

The training-strategy experiments use pretrained YOLOv8s and SGD for up to 150 epochs, with patience 100, an input-size setting of 960, and eight source images per batch. The initial learning rate is 0.006, the final learning-rate factor is 0.01, momentum is 0.937, weight decay is 0.0000625, and the nominal batch size is eight. Standard augmentation uses mosaic probability 0.7, mixup probability 0.05, and scale magnitude 0.45; mosaic is disabled for the last 10 epochs. Checkpoints are selected by native validation mAP50–95. Detector comparisons use the preceding AdamW recipe, and the matched-control design uses the

random-initialization recipe below. Strategy effects are calculated against the baseline in their own block.

SDS does not reserve a held-out test split, so the same validation set is used for checkpoint selection and for reporting results. To estimate how this selection affects the paired comparisons, we additionally evaluated the last.pt checkpoint saved at the end of training for the six SDS runs, and report its results alongside those for best.pt. Both checkpoints pass through the same unified evaluation pipeline. The stopping epoch for last.pt is still determined by validation-based early stopping, so this comparison serves only as a sensitivity analysis.

3.5. Matched Controls for Augmentation Attribution

The WSODD matched controls retain the random initialization and training recipe of the original implementation, allowing the seven configurations to be compared along a common sequence of changes. The scratch reference uses standard YOLOv8s with CIoU and box/DFL weights of 7.5/1.5. The adjusted loss weights change this pair to 4.0/1.3, values carried over from the original implementation and not separately tuned in this study, and the attention + localization-loss configuration then adds the CBAM and Focal-EIoU variant combination described above. Two further configurations apply the adjusted loss weights with attention only or localization loss only, separating the two components. The classification weight remains 0.5. All seven configurations start from YAML with seeds 0, 1, 2, 42, 123, and 3407, using Ultralytics 8.3.182, batch size four, and the principal SGD settings given above. Training allows up to 150 epochs with patience 100; best.pt is selected by validation fitness $0.1 \cdot \text{mAP50} + 0.9 \cdot \text{mAP50–95}$.

Clean-copy controls and localized-glare augmentation with GlareSoft together used 397 training images selected under seed 42, the floor of 10% of the original training set. The same selection was used across six training seeds, and one copy was appended to each selected image, giving 4374 training images per group. Both groups shared the same RGB read/write pipeline and a save-quality setting of 95, with matched labels, filename suffixes, and order. The control group retained clean content; GlareSoft copies added smooth elliptical bright spots covering 1.225–3.5% of the image area, with aspect ratios of 1.4–3.2, added brightness of 16–34, and attenuation parameters of 1.6–2.4. Selection and transformation were applied to the training set only.

Each contrast captures a distinct operational contribution. The first two measure the effects of adjusting box/DFL weights and of further adding CBAM with the Focal-EIoU variant; the two single-component configurations each yield their individual effect as the difference from the adjusted loss-weights configuration, and the combined configuration yields the effect of adding the other component as the difference from them. Adding clean copies to the component configuration captures the combined effects of repeated exposure, re-encoding, and additional iterations per epoch, as the training set grows. GlareSoft copies versus clean copies estimate the glare-transformation increment with copy count, source images, and processing matched. Two further contrasts measure the addition of GlareSoft copies to the component configuration and the total change from the scratch reference to the complete configuration. Figure 4 shows these relationships. Every contrast is first calculated between runs with the same training seed.

Two further controls apply the same matching logic to other settings. The first repeats the comparison of clean copies with GlareSoft copies under pretrained fine-tuning: following the standard training recipe of Section 3.4, clean copies or GlareSoft copies of the same 397 images are added to the WSODD training set, with six seeds trained for each, the difference between the two estimating the glare-transformation increment under pretrained conditions and the difference of the clean-copy configuration from standard training giving

the effect of appending copies, as such. The second is the identical paired-view control on FloW: the MDG parameters are still sampled as usual, but, after preprocessing, the second view is replaced by a copy of the first, so that the number of views, the number of updates per epoch, annotation copying and loss normalization all match paired-view training, with only the content of the second view differing; this configuration is, likewise, trained with six seeds. In addition, FloW has runs of pretrained single-component training using seeds 0, 1 and 2, adding only CBAM or only the Focal-EIoU variant, with all other settings the same as FloW standard training.

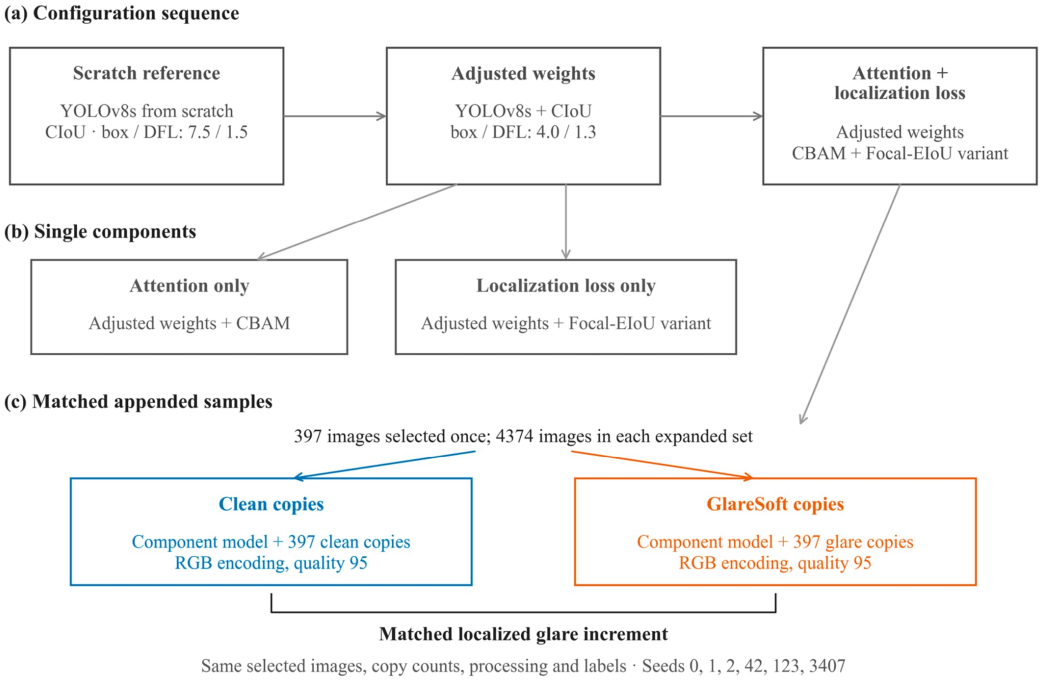


Figure 4. Seven matched WSODD configurations trained from scratch. Starting from the scratch reference, the design adjusts loss weights and then adds the attention and localization-loss combination; the two single-component configurations add only one of these components to the adjusted loss weights. Clean copies and GlareSoft copies append the same selected images to the component configuration; only the latter receive a localized elliptical bright patch. Arrows identify adjacent operational comparisons. Comparing the GlareSoft-copy configuration with the scratch reference gives the total change from the complete configuration.

3.6. Evaluation Under Normal and Degraded Conditions

The condition evaluation fixed the model weights, image identities, and annotations while varying only input appearance. The standard training checkpoints and the fixed single-view augmentation checkpoints for SDS and WSODD were evaluated on normal inputs, five severity levels of synthetic low light, and five severity levels of localized glare; WSODD used all six seeds and SDS three, giving 198 records in total. Each model and seed received the same transformed version of the same image. Normal evaluation used the original inputs, and transformations were generated at the original resolution before entering a common inference pipeline. We report the change in the standard training models relative to normal inputs, the paired augmentation difference within each condition, and the condition interaction term, defined as the augmentation difference of a seed under a degraded condition minus its augmentation difference under normal inputs. We, additionally, evaluated the weights of the matched controls, the pretrained copy configurations, and FloW's standard training, two paired-view configurations, and identical paired-view control under normal inputs, low-light severity levels 3–5, and localized-glare severity level

3. Beyond the synthetic conditions, we used image statistics and filename timestamps to delineate four overlapping real-condition subsets of the WSODD test set (low luminance images with mean gray level below 110, high saturation images in which pixels with gray level of at least 250 account for more than 0.3%, and images captured at 18:00–18:59 and 19:00–19:59) and rescored them with saved normal-input predictions as a descriptive check; the definitions and sizes of the subsets are given in the Supplementary Material.

Each degradation condition has five severity levels (Table A5). Levels 1–3 take the midpoints of three training-parameter bands, while levels 4 and 5 extend beyond the training range. Low light follows the brightness scaling and gamma transformation in Section 3.3: the brightness coefficient and gamma exponent run from 0.925 and 1.085 at level 1 to 0.84 and 1.255 at level 3, then 0.70 and 1.50 at level 4 and 0.50 and 2.00 at level 5, with results rounded to 8 bits. For localized glare, the elliptical target area as a share of the image and the added intensity at the center rise from 2.275% and 26.0 at level 1 to 4.45% and 43.5 at level 3, then 8% and 80 at level 4 and 15% and 120 at level 5; the aspect ratio is fixed at 2.3, the added intensity inside the ellipse decays as $\exp(-2d)$, where d is the sum of squares of the normalized elliptical coordinates and 2 is the attenuation coefficient, pixels outside the ellipse remain unchanged, and results are rounded and clipped to 0–255. The glare center is determined by the dataset namespace and file name, and is identical for all models and seeds. All levels were fixed before any model was evaluated under these conditions. Figure 5 shows input examples for levels 3 and 5.

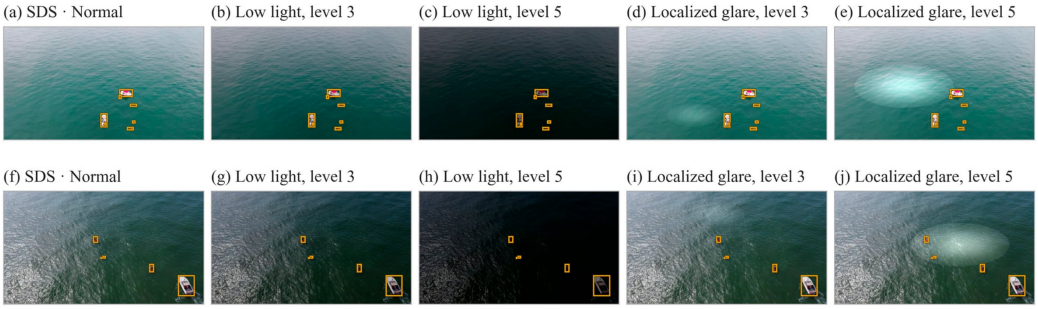


Figure 5. Examples of controlled synthetic test inputs. Each of the two rows shows one SDS image as its normal input, alongside its low-light and localized-glare versions at severity level 3 and severity level 5, with ground-truth bounding boxes unchanged. Severity level 3 is the midpoint of the highest training parameter band, while severity level 5 lies beyond the training range; parameters for each level are given in Table A5. Bright-spot positions are fixed by the dataset namespace and file names.

3.7. Metrics and Statistical Summaries

This study adopts the metric system of the COCO object-detection benchmark [34]. In the official pycocotools 2.0.10 COCOeval, the primary metric mAP50–95 averages over IoU thresholds from 0.50 to 0.95 in steps of 0.05; APsmall uses the COCO small-object area range in original-image coordinates, with mAP50 as a supplementary metric. The results reported here come from 123 best.pt checkpoints and 6 SDS last.pt checkpoints, yielding 621 evaluation records, all using 101 recall points and maxDets of 1, 10, and 100. Inference uniformly uses a 960 input, an actual batch size of 1, and FP32, with rectangular auto-padding disabled.

Repeated-training analyses used same-seed paired differences as the unit. Model metrics are expressed as percentages, and differences between configurations in percentage points (pp); Equation (2) defines the same-seed change as the difference between the strategy and baseline AP multiplied by 100. For each comparison, we report the per-seed differences, their mean, the sample standard deviation (denominator $n - 1$), the number of seeds with a positive difference, and a 95% confidence interval (CI) for the mean, computed

from a t distribution with n−1 degrees of freedom and critical values of 4.30 at n = 3 and 2.57 at n = 6. The plots retain the individual seed difference points, and the error bars represent the 95% confidence intervals. We also computed *p* values from two-sided paired t tests and from an exact sign-flip test that enumerates all 2-to-the-power-n sign combinations, for which the smallest attainable *p* value is 0.25 at n = 3 and 0.031 at n = 6; these *p* values are listed in Tables S2 and S3. In the text, statistical support refers to an unadjusted, paired t-based 95% CI excluding zero; the exact sign-flip test is a sensitivity analysis, and disagreement is stated explicitly. The t-based inference assumes independent seed-level differences with an approximately normal sampling distribution; the sign-flip analysis assumes sign symmetry under its null. With three or six seeds, these assumptions and limited resolution warrant caution. Intervals and *p* values are not adjusted for multiple comparisons. Secondary analyses are exploratory, and failure to detect a difference does not establish equivalence or absence of an effect. The primary comparisons are the paired mAP50–95 and APsmall differences defined in Sections 3.4–3.6; comparisons broken down by scale, class, and severity level are secondary analyses.

$$d_i = 100\left(AP_{strategy,i} - AP_{baseline,i}\right),\ \bar{d} = \frac{\sum_i d_i}{n}$$

$$s_d = \sqrt{\frac{\sum_i \left(d_i - \bar{d}\right)^2}{n-1}} \quad (2)$$

For each seed and degradation condition c, the condition interaction equals the augmented-minus-standard-training difference under that condition, minus the corresponding difference for the same seed under normal input, as defined in Equation (3). A positive interaction indicates a larger relative advantage for the augmented model under that condition, even if the absolute AP of both models declines. This double difference is first computed within matched seeds, model pairs, and images, then pooled across all seeds for that dataset (six for WSODD, three for SDS) to obtain the mean, sample standard deviation, and 95% confidence interval. By definition, the interaction under normal input is zero.

$$I_{i,c} = d_{i,c} - d_{i,normal} \quad (3)$$

Scale and class analyses rescore saved predictions on the normal inputs without rerunning inference. In addition to the three COCO-size tiers, small (area below 32^2 px^2), medium (32^2–96^2 px^2), and large (no smaller than 96^2 px^2), areas below 96^2 px^2 are further summarized in four intervals: below 16^2, 16^2–32^2, 32^2–64^2, and 64^2–96^2 px^2. All areas use original-image pixel coordinates; larger objects are reported through APlarge, and APmedium is also provided. Per-class results are the average precision of that class across all IoU thresholds; when the evaluation set contains no objects of a given class or scale interval, the corresponding term is excluded from the computation.

4. Results

4.1. Seed-0 Detector Reference Across the Three Datasets

In the single-seed reference, the detector with the highest mAP50–95 value varies by dataset (Table 2). RT-DETR-L reaches 24.00% on WSODD and 46.42% on FloW, the latter 2.96 pp above the highest-scoring YOLO model on that dataset; on SDS, YOLO26s reaches 47.34%, 0.42 pp above RT-DETR-L. As a scale reference, in the training-strategy experiment block of Section 4.2, the sample standard deviation of mAP50–95 across seeds for YOLOv8s under standard training was 0.35 pp (SDS, three seeds), 0.86 pp (WSODD, six seeds) and 1.01 pp (FloW, six seeds); the 0.42 pp difference on SDS is of the same order of magnitude.

Table 2. Reference results for four detectors on three datasets. Each detector–dataset combination was trained once. Metrics are reported as percentages, with mean average precision (mAP50–95) as the primary metric, and comparisons are made separately, within each dataset.

Dataset	Detector	mAP50–95 (%)	AP_{small} (%)	mAP50 (%)
SDS	YOLOv8s	46.472	28.572	71.011
SDS	YOLO11s	45.420	23.715	69.372
SDS	YOLO26s	47.341	34.385	71.642
SDS	RT-DETR-L	46.920	28.464	70.190
WSODD	YOLOv8s	21.816	11.717	43.520
WSODD	YOLO11s	21.704	8.205	43.705
WSODD	YOLO26s	22.812	12.986	45.270
WSODD	RT-DETR-L	23.995	12.920	46.296
FloW	YOLOv8s	41.618	25.291	83.426
FloW	YOLO11s	43.463	29.073	87.331
FloW	YOLO26s	41.093	26.661	83.563
FloW	RT-DETR-L	46.419	31.766	87.216

Small-object metrics changed the gaps between some models. On SDS, YOLO26s reached an APsmall of 34.39%, 5.81 pp above YOLOv8s, while their overall mAP50–95 differed by only 0.87 pp. On WSODD, YOLO26s and RT-DETR-L had nearly identical APsmall values, 12.99% and 12.92%, respectively, reversing the order seen for overall mAP. Figure 6 shows the relationship between the two metrics.

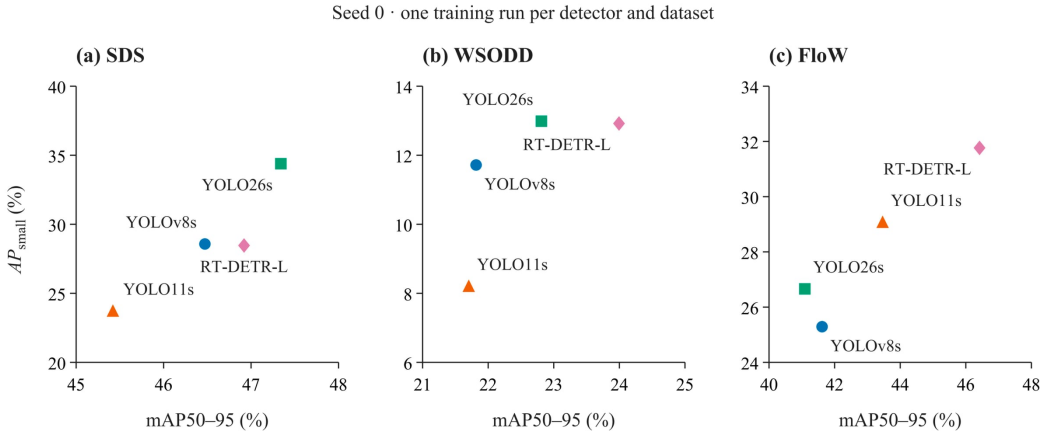


Figure 6. Overall and small-object performance of four detectors. The horizontal axis is mAP50–95 and the vertical axis is APsmall, both in percent; colors and point shapes remain consistent across panels. Each point represents one trained detector–dataset combination, and the coordinate ranges are adjusted by dataset to show model differences.

RT-DETR-L's higher overall AP on WSODD comes at a greater computational cost (Table 3). Its average pipeline latency is 39.17 ms (25.53 FPS), compared with 14.62 ms (68.42 FPS) for YOLOv8s. RT-DETR-L has 32.84 M parameters and 241.67 GFLOPs, whereas the three YOLO models have 9.43–11.14 M parameters and 49.16–64.51 GFLOPs. YOLOv8s has the lowest latency and peak memory allocation, and RT-DETR-L has the highest overall mAP.

4.2. Stability Across Training Seeds

Under normal input, the average effect of fixed single-view augmentation on WSODD is far smaller than the variation between seeds (Table 4, Figure 7). Across the six seeds, the paired change in mAP50–95 was −0.099 ± 1.104 pp on average (95% CI −1.26 to +1.06), with a sample standard deviation about 11 times the absolute mean change; individual

seed changes ranged from −1.81 to +1.48 pp, with three seeds positive and three negative, so any single run could present a near-zero average effect as a clear gain or loss. APsmall changed by +0.52 ± 1.99 pp on average (95% CI −1.57 to +2.60), rising in five seeds, with a larger decrease in the remaining seed pulling the mean down.

Table 3. Computational cost and inference efficiency of the four seed-0 WSODD models. Measurements use an RTX 4090, FP32, 960 × 960 inputs, batch size one, 50 warm-up images, and the same 500 test images. Latency is mean pipeline time, including preprocessing and postprocessing; memory is peak PyTorch allocation. AP values come from unified COCOeval on the full 2406-image test set. GFLOPs are estimates for the unfused models.

Detector	Params (M)	GFLOPs	Latency (ms)	FPS	Peak Memory (MiB)	mAP50–95 (%)	AP_{small} (%)
YOLOv8s	11.141	64.51	14.62	68.42	124.79	21.816	11.717
YOLO11s	9.433	49.16	15.10	66.23	163.41	21.704	8.205
YOLO26s	9.959	51.98	16.43	60.85	163.61	22.812	12.986
RT-DETR-L	32.835	241.67	39.17	25.53	330.16	23.995	12.920

Table 4. Results of multi-seed training strategies under normal conditions. Values are the mean ± sample standard deviation across training seeds; paired changes are computed between the strategy and standard training for the same seed, in percentage points (pp), and are given with 95% confidence intervals. Seeds 0, 1, 2 were used for SDS, and 0, 1, 2, 42, 123, 3407 for WSODD and FloW.

(a) Performance of Each Configuration

Dataset	Configuration	n	mAP50–95 (%)	AP_{small} (%)
SDS	Standard training	3	46.856 ± 0.355	29.511 ± 0.746
SDS	Fixed single view	3	46.542 ± 0.363	29.045 ± 0.569
WSODD	Standard training	6	22.184 ± 0.856	12.588 ± 1.423
WSODD	Fixed single view	6	22.085 ± 0.702	13.106 ± 1.384
FloW	Standard training	6	45.070 ± 1.011	27.268 ± 0.983
FloW	Paired views	6	45.073 ± 0.804	27.411 ± 0.958
FloW	Paired views + small-object weighting	6	44.267 ± 1.126	27.101 ± 1.529

(b) Paired Changes from Standard Training

Dataset	Configuration	n	ΔmAP50–95 (pp)	95% CI (pp)	ΔAP_{small} (pp)	95% CI (pp)
SDS	Fixed single view	3	−0.314 ± 0.146	−0.678 to +0.050	−0.466 ± 1.281	−3.648 to +2.716
WSODD	Fixed single view	6	−0.099 ± 1.104	−1.257 to +1.060	+0.518 ± 1.985	−1.565 to +2.602
FloW	Paired views	6	+0.003 ± 1.153	−1.207 to +1.214	+0.143 ± 1.609	−1.545 to +1.832
FloW	Paired views + small-object weighting	6	−0.803 ± 1.767	−2.657 to +1.052	−0.167 ± 1.297	−1.528 to +1.194

On WSODD, the near-zero overall effect decomposes into components that point in different directions, once scale and class are separated. Fixed single-view augmentation lowers APmedium by 0.59 pp on average (95% CI −1.16 to −0.01), with a decrease in five of six seeds; the mean change for APlarge is +0.47 pp, with an interval that includes zero, and the APsmall change is reported in the preceding paragraph. Among the 14 classes, none has an interval excluding zero; boat is the only class that rises across all six seeds (+0.33 pp, 95% CI −0.01 to +0.66), while ball, the class with the most small objects, shows −0.50 pp (95% CI −1.60 to +0.60). Full scale-bin and per-class results are given in Table S5.

On the SDS validation set, the direction of the fixed single-view augmentation effect depended on the checkpoint used. Evaluated with best.pt, selected by validation mAP50–95, the paired change in mAP50–95 was −0.31 ± 0.15 pp (95% CI −0.68 to +0.05), decreasing for all three seeds; APlarge fell by 0.79 pp (95% CI −1.29 to −0.29) and APsmall changed by −0.47 ± 1.28 pp, rising for one seed and falling for two. With last.pt, taken at the end

of training, the paired change in mAP50–95 was +0.08 ± 0.28 pp (95% CI −0.60 to +0.77) and that in APlarge was +0.43 pp (95% CI −3.47 to +4.34), both positive for two seeds. Across the six runs, best.pt occurred at epochs 13–27 and training stopped early at epochs 113–127; last.pt scored 2.50 pp lower than best.pt on mAP50–95 under standard training and 2.11 pp lower under augmentation. The small decrease on SDS thus appeared only when the checkpoint was selected on the validation set, and this study does not treat it as independent evidence of generalization.

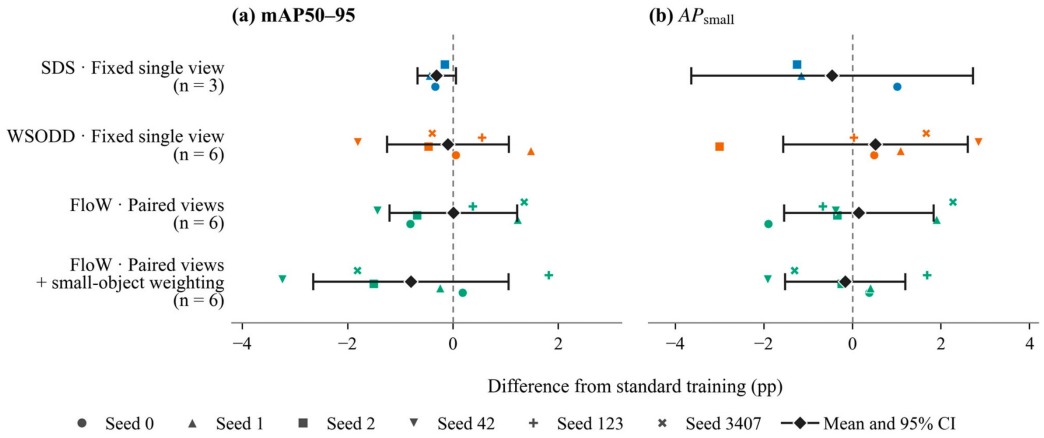


Figure 7. Paired effects of training strategies across seeds under normal evaluation conditions, shown separately for mAP50–95 and APsmall. SDS and WSODD assess fixed single-view augmentation, while FloW assesses paired-view training with and without small-object weighting, each relative to standard training in the same block. Each colored point represents one seed's difference. Black diamonds and error bars denote the mean and 95% confidence interval of the paired differences, measured in pp. SDS uses three seeds, and WSODD and FloW each use six seeds. Colored points are slightly offset vertically, to distinguish nearby values.

Paired-view training in FloW showed no statistically detectable change in mAP50–95: the mean paired change across six seeds was 0.00 ± 1.15 pp (95% CI −1.21 to +1.21), with three seeds improving and three declining, and individual seeds ranging from −1.44 to +1.35 pp. With small-object weighting, the mean change in mAP50–95 was −0.80 ± 1.77 pp (95% CI −2.66 to +1.05), and four seeds declined. This decrease fell on medium and large objects: APmedium and APlarge changed by −0.74 pp (95% CI −2.39 to +0.91) and −3.67 pp (95% CI −8.79 to +1.46), respectively, each declining in five seeds, while APsmall changed by −0.17 ± 1.30 pp (95% CI −1.53 to +1.19), with three seeds improving and three declining. Under normal input, weighting for small objects did not yield higher small-object accuracy; its behavior under low light is reported in Section 4.4.

4.3. Attributing Gains Through Matched Controls

The mean net effect of the four steps from the scratch reference to the complete configuration is close to zero. Among these four steps, only loss-weight adjustment has a paired t-based 95% CI excluding zero, in the negative direction (Table 5, Figure 8). Its paired t-test p value is 0.028, whereas the exact sign-flip p value is 0.0625; thus, the evidence is sensitive to the inferential method. Relative to the scratch reference, that adjustment lowered mAP50–95 by 0.65 ± 0.52 pp (95% CI −1.19 to −0.10), with a decrease across five seeds. Combining attention + localization loss then yielded +0.39 ± 0.56 pp (95% CI −0.20 to +0.98), improving four seeds; clean copies contributed +0.09 ± 0.71 pp and GlareSoft copies +0.07 ± 0.68 pp, each with three increases and three decreases. At the level of mean changes, the last three steps sum to +0.56 pp and numerically offset much of the loss-weight adjustment; this does not establish separate benefits of those steps. Thus, the complete

configuration stands at −0.09 ± 0.53 pp, relative to the scratch reference (95% CI −0.64 to +0.46).

Table 5. Matched comparisons. Panel (**a**) reports means ± sample standard deviations over seven WSODD configurations trained from scratch, in percent. Panel (**b**) reports means ± sample standard deviations of same-seed paired differences, 95% confidence intervals, and the number of seeds with positive differences, in percentage points. Panel (**c**) presents the comparison of clean copies with GlareSoft copies under pretrained fine-tuning and the FloW comparison of paired-view training with identical paired views. All comparisons use seeds 0, 1, 2, 42, 123 and 3407, and are made within their respective experimental blocks. Clean copies and GlareSoft copies match source images, counts, and processing pipeline.

(**a**) **Performance of the Seven Configurations**

Configuration	**mAP50–95 (%)**	AP_{small} **(%)**	**mAP50 (%)**
Scratch reference	19.365 ± 0.127	12.114 ± 2.302	38.583 ± 0.565
Adjusted loss weights	18.716 ± 0.518	10.099 ± 1.522	38.213 ± 0.825
Attention only	18.753 ± 0.735	9.793 ± 1.751	38.056 ± 0.923
Localization loss only	18.984 ± 0.521	11.662 ± 0.995	38.414 ± 0.607
Attention + localization loss	19.108 ± 0.411	11.642 ± 2.438	39.194 ± 0.724
Clean copies	19.201 ± 0.346	10.720 ± 1.508	39.109 ± 0.630
GlareSoft copies	19.275 ± 0.489	12.567 ± 1.248	39.335 ± 0.851

(**b**) **Paired Operational Changes**

Comparison	**ΔmAP50–95 (pp)**	**95% CI (pp)**	**Positive Seeds**	ΔAP_{small} **(pp)**	**95% CI (pp)**	**Positive Seeds**
Loss-weight adjustment	−0.649 ± 0.519	−1.194 to −0.105	1/6	−2.015 ± 3.317	−5.496 to +1.466	1/6
Attention + localization-loss addition	+0.392 ± 0.561	−0.196 to +0.981	4/6	+1.542 ± 3.233	−1.851 to +4.935	4/6
Clean-copy addition	+0.093 ± 0.708	−0.651 to +0.836	3/6	−0.921 ± 2.588	−3.637 to +1.794	2/6
Matched localized-glare increment	+0.074 ± 0.675	−0.635 to +0.783	3/6	+1.847 ± 1.907	−0.154 to +3.848	5/6
GlareSoft vs. component configuration	+0.167 ± 0.630	−0.494 to +0.828	4/6	+0.926 ± 3.590	−2.841 to +4.693	4/6
Complete configuration vs. scratch reference	−0.090 ± 0.525	−0.641 to +0.462	3/6	+0.453 ± 2.712	−2.393 to +3.299	3/6
Attention only vs. adjusted loss weights	+0.037 ± 0.886	−0.893 to +0.968	2/6	−0.307 ± 2.817	−3.263 to +2.650	2/6
Localization loss only vs. adjusted loss weights	+0.268 ± 0.965	−0.744 to +1.281	4/6	+1.563 ± 1.718	−0.240 to +3.366	5/6
Localization loss added to attention only	+0.355 ± 0.884	−0.573 to +1.283	4/6	+1.849 ± 3.233	−1.544 to +5.242	5/6
Attention added to localization loss only	+0.124 ± 0.807	−0.723 to +0.971	2/6	−0.020 ± 2.305	−2.439 to +2.398	3/6

(**c**) **Copy Comparisons Under Pretrained Fine-Tuning and the Identical Paired-View Control**

Comparison	**ΔmAP50–95 (pp)**	**95% CI (pp)**	**Positive Seeds**	ΔAP_{small} **(pp)**	**95% CI (pp)**	**Positive Seeds**
WSODD, pretrained: GlareSoft copies vs. clean copies	−0.567 ± 2.350	−3.034 to +1.899	3/6	+0.815 ± 1.671	−0.938 to +2.569	4/6
WSODD, pretrained: clean copies vs. standard training	−0.540 ± 0.240	−0.792 to −0.288	0/6	−0.583 ± 1.027	−1.661 to +0.496	2/6
FloW: paired views vs. identical paired views	+0.541 ± 1.158	−0.674 to +1.756	4/6	+0.174 ± 1.350	−1.243 to +1.591	3/6
FloW: identical paired views vs. standard training	−0.537 ± 1.803	−2.429 to +1.355	3/6	−0.031 ± 1.757	−1.874 to +1.813	3/6

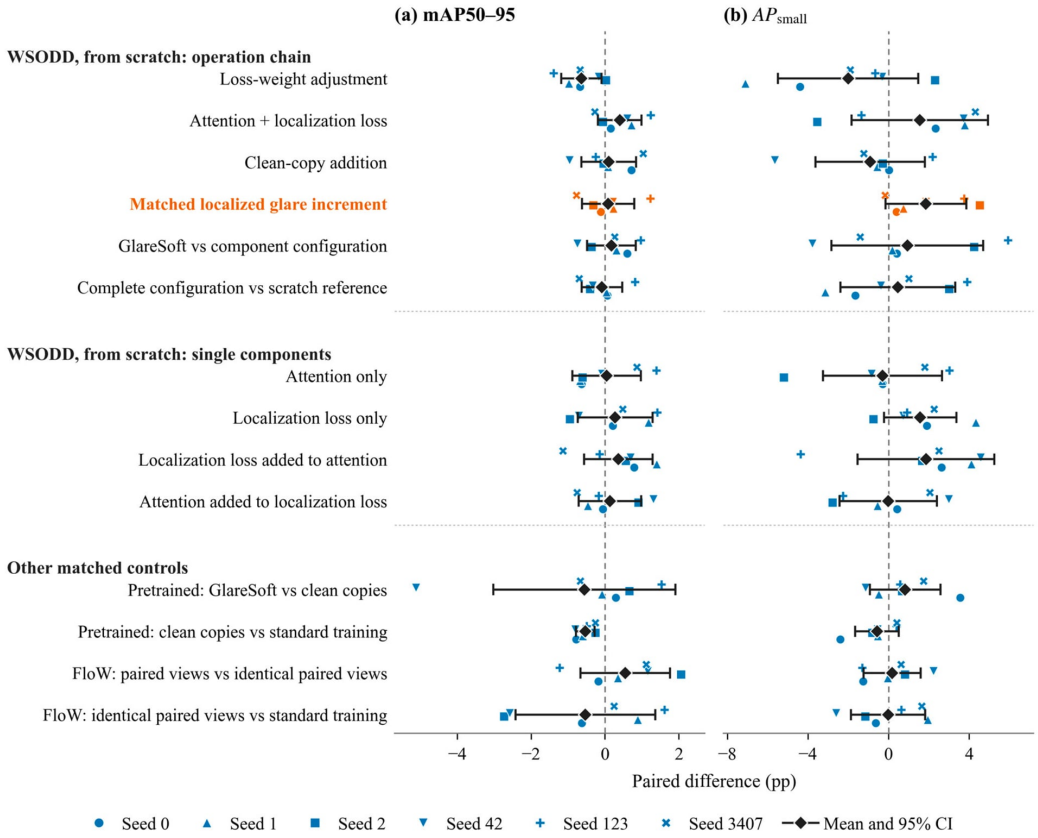


Figure 8. Paired differences in mAP50–95 and APsmall for matched comparisons; each point corresponds to one seed, and diamonds and horizontal error bars show means and 95% confidence intervals in pp. The **upper** section shows the operational chain for WSODD training from scratch, with orange highlighting the localized-glare increment of GlareSoft, relative to clean copies; the **middle** section shows single-component comparisons; the **lower** section shows copy comparisons under pretrained fine-tuning and the comparison of identical paired views for FloW.

Small-object metrics show larger differences and greater seed-to-seed variation, and no single-step interval excludes zero. Against clean copies, GlareSoft gives an APsmall increment of +1.85 ± 1.91 pp (95% CI −0.15 to +3.85), positive for five of the six seeds; adjusting the weights gives −2.02 ± 3.32 pp (95% CI −5.50 to +1.47), with five of the six seeds declining; adding CBAM together with the Focal-EIoU variant gives +1.54 ± 3.23 pp (95% CI −1.85 to +4.94), and adding clean copies gives −0.92 ± 2.59 pp. The complete configuration against the reference gives +0.45 ± 2.71 pp (95% CI −2.39 to +3.30), with three seeds up and three down. The WSODD test set contains only 1126 small objects, concentrated in a few classes (Section 3.1); across the seven configurations, seed-to-seed standard deviations are 1.0–2.4 pp for APsmall but only 0.1–0.7 pp for mAP50–95, which is why the intervals on small objects are markedly wider.

Decomposing the combination on WSODD shows that neither component alone produced a statistically supported change. With attention only, mAP50–95 was +0.04 ± 0.89 pp (95% CI −0.89 to +0.97) and APsmall was −0.31 ± 2.82 pp; with localization loss only, mAP50–95 was +0.27 ± 0.96 pp (95% CI −0.74 to +1.28) and APsmall was +1.56 ± 1.72 pp (95% CI −0.24 to +3.37), positive in five of the six seeds. Relative to attention only, the attention + localization-loss configuration changed APsmall by +1.85 ± 3.23 pp, and, relative to localization loss only, by −0.02 ± 2.30 pp. These descriptive shifts do not establish a benefit of either component. The within-seed interaction, defined as combined − attention only − localization loss only + adjusted loss weights, was +0.09 pp for mAP50–95 (95% CI −1.37 to +1.54) and +0.29 pp for APsmall (95% CI −2.99 to +3.56); neither interval excludes zero

(Table S2 component-interaction supplement). Pretrained single-component training on FloW pointed in a different direction: localization loss only reduced APsmall by 1.20 ± 0.26 pp (95% CI −1.84 to −0.56), with a decrease across all three seeds, and gave mAP50–95 of −1.30 ± 0.86 pp, while attention only gave mAP50–95 of −1.79 ± 2.50 pp.

The effect of appending copies depends on the training setting. Under training from scratch, the clean copies were +0.09 ± 0.71 pp, relative to the attention + localization-loss configuration; under pretrained fine-tuning, they reduced mAP50–95 by 0.54 ± 0.24 pp, relative to standard training (95% CI −0.79 to −0.29), with all six seeds showing a decrease. Under pretrained fine-tuning, the increment from the glare transformation also lacked statistical support: GlareSoft copies scored −0.57 ± 2.35 pp on mAP50–95, relative to clean copies (95% CI −3.03 to +1.90), and +0.82 ± 1.67 pp on APsmall (95% CI −0.94 to +2.57), with four seeds positive, a direction consistent with training from scratch, but smaller in magnitude. The wide mAP50–95 interval stems mainly from one seed, whose best.pt was selected at epoch 12 and which reached a test mAP50–95 of 16.67%, whereas the other five seeds ranged from 20.03% to 23.02%.

FloW's identical paired-view control applies the same question to the second view: paired-view training against identical paired views yields +0.54 ± 1.16 pp mAP50–95 (95% CI −0.67 to +1.76), positive in four of the six seeds, and +0.17 ± 1.35 pp APsmall, while identical paired views against standard training yields −0.54 ± 1.80 pp (95% CI −2.43 to +1.35). With the number of views, number of updates, and loss normalization held equal, the paired comparison does not establish a benefit of the degraded content in the second view; its interval remains compatible with both negative and positive effects. The WSODD copy comparison and the FloW view comparison, therefore, converge on the same conclusion: once the effect of additional samples or additional views is matched out, the degradation transform itself has no statistically supported gain on either dataset.

4.4. Performance Under Low Light and Glare

Synthetic low light progressively reduces the accuracy of the standard training model as severity level increases, whereas localized glare leaves accuracy almost unchanged across all five levels (Table 6, Figure 9). Under normal input, the standard training model reaches 22.18% mAP50–95 on WSODD and 46.86% on SDS; at low-light severity level 3, these fall to 21.72% and 46.72%, at level 4 to 20.44% and 46.13%, and at level 5 to 13.98% and 42.88%. The strongest glare level, with an ellipse area of 15% and a center intensity of 120, lowers WSODD and SDS by only 0.19 and 0.28 pp, respectively. With these parameters, synthetic localized glare does not stress the detectors, so these results provide limited evidence about glare robustness. They show no statistically detectable difference between the approaches under this particular input variation, and indicate neither equivalence nor robustness to real glare.

The advantage of fixed single-view augmentation over standard training grows as low light worsens. On WSODD, the paired difference in mAP50–95 is −0.10 pp under normal input and −0.06, +0.07, and +0.22 pp at severity levels 1–3 within the training range, with every interval including zero; beyond that range, it reaches +0.97 pp at severity level 4 (95% CI −0.49 to +2.43) and +4.99 pp at severity level 5 (95% CI +2.68 to +7.30), positive across all six seeds at severity level 5. The condition interaction is +0.32 pp at severity level 3 (95% CI −0.03 to +0.68, positive for five seeds), +1.07 pp at severity level 4 (95% CI +0.34 to +1.80), and +5.09 pp at severity level 5 (95% CI +3.62 to +6.56), positive across all six seeds at the latter two levels. For SDS, the paired difference stays between −0.34 and −0.09 pp at severity levels 1–4, and turns to +1.37 pp at severity level 5 (95% CI +0.35 to +2.38), positive for all three seeds, with an interaction of +1.68 pp (95% CI +0.41 to +2.95); with only three seeds, an exact sign-flip test yields a minimum p value of 0.25. The

four real-condition subsets of the WSODD test set do not reproduce this direction: on the 160 images with mean gray level below 110, the paired difference in mAP50–95 for fixed single-view augmentation is −2.26 pp (95% CI −4.70 to +0.18), positive for only one of six seeds, and the intervals for the other three subsets also include zero (Table S2).

Table 6. mAP50–95 for panel (**a**) and APsmall for panel (**b**) of the standard training models and the fixed single-view augmentation models on normal input and at each severity level of synthetic low light and localized glare. Each row uses six seeds for WSODD and three seeds for SDS. Absolute metrics are means ± sample standard deviation in percent; paired differences are augmentation minus standard training, and the condition interaction is that difference minus the same-seed difference under normal input, both reported as mean ± sample standard deviation with 95% confidence intervals in percentage points. Corresponding results for APmedium and APlarge are given in Table S3.

(a) Overall mAP50–95

Dataset	Condition	Level	Standard (%)	Fixed Single View (%)	Change (pp)	95% CI (pp)	Interaction (pp)	95% CI (pp)
SDS	Normal	—	46.856 ± 0.355	46.542 ± 0.363	−0.314 ± 0.146	−0.678 to +0.050	0.000 ± 0.000	—
SDS	Low light	1	46.822 ± 0.232	46.506 ± 0.332	−0.316 ± 0.130	−0.639 to +0.007	−0.002 ± 0.090	−0.227 to +0.223
SDS	Low light	2	46.824 ± 0.191	46.483 ± 0.309	−0.341 ± 0.138	−0.684 to +0.002	−0.027 ± 0.132	−0.356 to +0.302
SDS	Low light	3	46.721 ± 0.147	46.389 ± 0.334	−0.332 ± 0.212	−0.859 to +0.194	−0.018 ± 0.197	−0.508 to +0.471
SDS	Low light	4	46.132 ± 0.193	46.038 ± 0.278	−0.094 ± 0.121	−0.395 to +0.208	+0.220 ± 0.233	−0.358 to +0.799
SDS	Low light	5	42.879 ± 0.445	44.245 ± 0.236	+1.366 ± 0.409	+0.351 to +2.382	+1.680 ± 0.510	+0.413 to +2.947
SDS	Localized glare	1	46.860 ± 0.348	46.553 ± 0.347	−0.307 ± 0.146	−0.671 to +0.057	+0.007 ± 0.008	−0.012 to +0.026
SDS	Localized glare	2	46.858 ± 0.342	46.546 ± 0.354	−0.312 ± 0.164	−0.720 to +0.096	+0.002 ± 0.018	−0.043 to +0.046
SDS	Localized glare	3	46.861 ± 0.345	46.541 ± 0.361	−0.320 ± 0.161	−0.720 to +0.080	−0.006 ± 0.016	−0.045 to +0.033
SDS	Localized glare	4	46.786 ± 0.336	46.464 ± 0.406	−0.322 ± 0.130	−0.644 to −0.000	−0.008 ± 0.075	−0.195 to +0.178
SDS	Localized glare	5	46.573 ± 0.360	46.230 ± 0.381	−0.342 ± 0.080	−0.540 to −0.144	−0.028 ± 0.076	−0.218 to +0.162
WSODD	Normal	—	22.184 ± 0.856	22.085 ± 0.702	−0.099 ± 1.104	−1.257 to +1.060	0.000 ± 0.000	—
WSODD	Low light	1	22.123 ± 0.805	22.068 ± 0.689	−0.055 ± 1.136	−1.247 to +1.137	+0.043 ± 0.141	−0.105 to +0.192
WSODD	Low light	2	21.939 ± 0.731	22.008 ± 0.696	+0.069 ± 1.106	−1.091 to +1.230	+0.168 ± 0.260	−0.105 to +0.441
WSODD	Low light	3	21.721 ± 0.677	21.944 ± 0.730	+0.223 ± 1.121	−0.954 to +1.399	+0.321 ± 0.337	−0.033 to +0.675
WSODD	Low light	4	20.436 ± 0.772	21.405 ± 0.993	+0.969 ± 1.389	−0.489 to +2.427	+1.068 ± 0.696	+0.338 to +1.798
WSODD	Low light	5	13.980 ± 1.408	18.970 ± 1.432	+4.990 ± 2.198	+2.683 to +7.298	+5.089 ± 1.404	+3.616 to +6.562
WSODD	Localized glare	1	22.149 ± 0.886	22.078 ± 0.704	−0.071 ± 1.123	−1.250 to +1.108	+0.028 ± 0.048	−0.022 to +0.078
WSODD	Localized glare	2	22.146 ± 0.900	22.072 ± 0.709	−0.074 ± 1.140	−1.271 to +1.122	+0.024 ± 0.080	−0.060 to +0.108
WSODD	Localized glare	3	22.152 ± 0.904	22.064 ± 0.713	−0.088 ± 1.149	−1.294 to +1.118	+0.010 ± 0.083	−0.077 to +0.098
WSODD	Localized glare	4	22.132 ± 0.934	22.057 ± 0.674	−0.075 ± 1.102	−1.231 to +1.081	+0.023 ± 0.105	−0.087 to +0.134
WSODD	Localized glare	5	21.992 ± 0.982	21.894 ± 0.757	−0.098 ± 1.126	−1.280 to +1.084	+0.000 ± 0.217	−0.228 to +0.228

Table 6. *Cont.*

(**b**) AP_{small}								
Dataset	**Condition**	**Level**	**Standard (%)**	**Fixed Single View (%)**	**Change (pp)**	**95% CI (pp)**	**Interaction (pp)**	**95% CI (pp)**
SDS	Normal	—	29.511 ± 0.746	29.045 ± 0.569	−0.466 ± 1.281	−3.648 to +2.716	0.000 ± 0.000	—
SDS	Low light	1	29.069 ± 0.713	28.529 ± 0.664	−0.540 ± 0.963	−2.932 to +1.853	−0.074 ± 0.556	−1.455 to +1.307
SDS	Low light	2	28.823 ± 0.631	28.445 ± 0.519	−0.377 ± 0.867	−2.532 to +1.777	+0.088 ± 0.587	−1.371 to +1.548
SDS	Low light	3	28.557 ± 0.480	28.176 ± 0.345	−0.381 ± 0.408	−1.394 to +0.631	+0.085 ± 0.898	−2.147 to +2.316
SDS	Low light	4	28.076 ± 1.073	27.425 ± 0.579	−0.651 ± 0.538	−1.987 to +0.684	−0.186 ± 1.348	−3.534 to +3.162
SDS	Low light	5	26.752 ± 0.477	26.355 ± 0.553	−0.397 ± 0.459	−1.538 to +0.743	+0.068 ± 1.665	−4.068 to +4.205
SDS	Localized glare	1	29.255 ± 0.486	29.089 ± 0.657	−0.166 ± 1.121	−2.951 to +2.619	+0.300 ± 0.177	−0.139 to +0.739
SDS	Localized glare	2	29.048 ± 0.549	29.058 ± 0.656	+0.010 ± 1.158	−2.866 to +2.886	+0.476 ± 0.206	−0.035 to +0.987
SDS	Localized glare	3	29.228 ± 0.591	29.099 ± 0.642	−0.128 ± 1.231	−3.185 to +2.928	+0.337 ± 0.247	−0.276 to +0.951
SDS	Localized glare	4	29.251 ± 0.873	28.737 ± 0.339	−0.514 ± 1.211	−3.522 to +2.495	−0.048 ± 0.081	−0.248 to +0.152
SDS	Localized glare	5	28.621 ± 0.944	28.309 ± 0.353	−0.311 ± 1.274	−3.476 to +2.853	+0.155 ± 0.684	−1.544 to +1.853
WSODD	Normal	—	12.588 ± 1.423	13.106 ± 1.384	+0.518 ± 1.985	−1.565 to +2.602	0.000 ± 0.000	—
WSODD	Low light	1	12.471 ± 1.290	13.469 ± 1.417	+0.998 ± 1.436	−0.509 to +2.505	+0.480 ± 1.112	−0.688 to +1.647
WSODD	Low light	2	12.437 ± 1.723	13.274 ± 1.878	+0.837 ± 1.972	−1.232 to +2.906	+0.319 ± 1.569	−1.327 to +1.965
WSODD	Low light	3	11.858 ± 1.831	13.281 ± 1.864	+1.423 ± 1.294	+0.065 to +2.781	+0.905 ± 1.618	−0.793 to +2.603
WSODD	Low light	4	12.162 ± 1.672	13.019 ± 1.956	+0.857 ± 2.110	−1.358 to +3.071	+0.338 ± 3.018	−2.829 to +3.505
WSODD	Low light	5	11.088 ± 1.295	11.993 ± 2.335	+0.905 ± 2.590	−1.812 to +3.623	+0.387 ± 4.068	−3.882 to +4.656
WSODD	Localized glare	1	11.874 ± 1.594	12.390 ± 1.778	+0.517 ± 2.485	−2.092 to +3.125	−0.002 ± 0.639	−0.673 to +0.669
WSODD	Localized glare	2	12.221 ± 1.490	12.998 ± 1.277	+0.778 ± 1.867	−1.182 to +2.737	+0.259 ± 0.563	−0.331 to +0.850
WSODD	Localized glare	3	12.434 ± 1.376	13.372 ± 1.045	+0.937 ± 2.028	−1.191 to +3.066	+0.419 ± 0.577	−0.187 to +1.025
WSODD	Localized glare	4	12.693 ± 1.187	12.836 ± 1.392	+0.143 ± 2.043	−2.001 to +2.288	−0.375 ± 1.106	−1.536 to +0.786
WSODD	Localized glare	5	10.788 ± 1.192	10.582 ± 1.105	−0.206 ± 0.620	−0.857 to +0.444	−0.725 ± 1.686	−2.494 to +1.044

Under low light, the gain from augmentation falls on different object sizes as severity increases: at severity level 3, the strongest level within the training range, the gain appears on small objects, while beyond the training range it comes mainly from medium and large objects. At severity level 3, the paired difference in APsmall for WSODD was +1.42 pp (95% CI +0.07 to +2.78), positive across all six seeds, and the condition interaction relative to normal input was +0.90 pp (95% CI −0.79 to +2.60), positive in five seeds. At severity levels 4 and 5, the APsmall differences were +0.86 and +0.91 pp, with intervals that include zero. By severity level 5, the standard training WSODD model showed APsmall 1.50 pp below normal input, with APmedium and APlarge lower by 5.28 and 10.88 pp; the augmented model fell by 1.11, 2.12, and 4.86 pp, respectively. At this level, the paired differences

between the two training approaches were +2.57 pp for APmedium (95% CI +0.96 to +4.18) and +6.50 pp for APlarge (95% CI +2.87 to +10.13), positive across all six seeds. For SDS at severity level 5, the APmedium and APlarge differences were +1.99 and +3.35 pp, positive across all three seeds, with intervals that include zero.

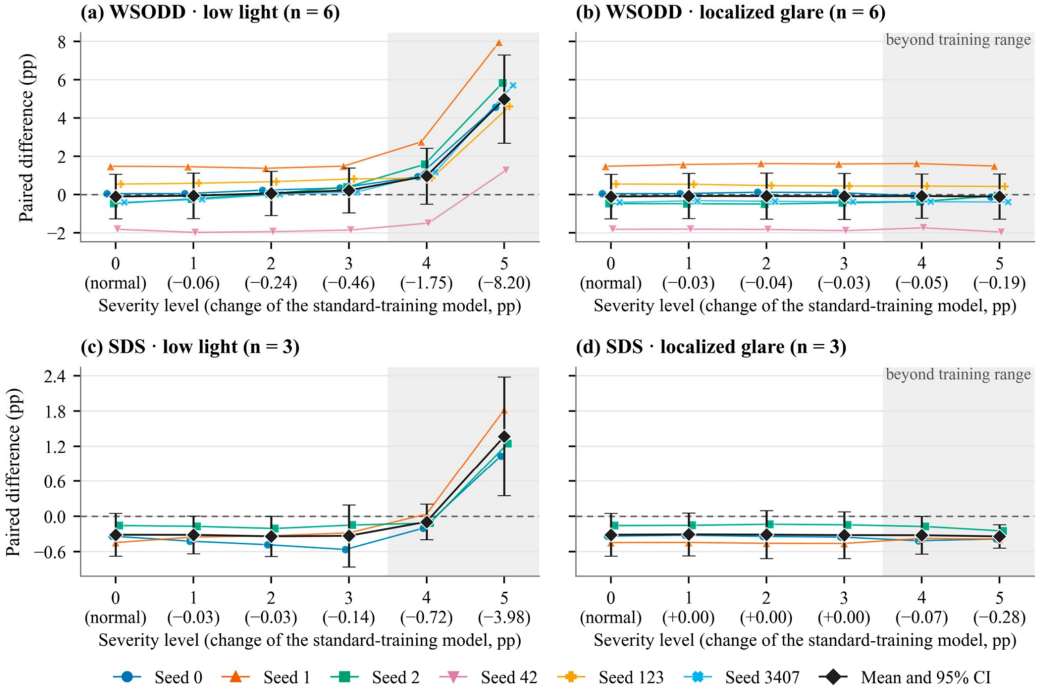


Figure 9. Paired differences in mAP50–95 for fixed single-view augmentation relative to standard training, as test degradation-severity increases. The top row shows WSODD (six seeds) and the bottom row SDS (three seeds); the left column shows synthetic low light and the right column localized glare. Severity level 0 is the normal input, levels 1–3 fall within the training parameter range, and levels 4 and 5 in the shaded region lie beyond that range. Thin lines and colored markers denote individual seeds, while black diamonds and error bars denote means and 95% CIs in pp. Values in parentheses below the horizontal axis are the mean change in mAP50–95 of the standard training model at that level, relative to the normal input.

Under localized glare, the differences between the two training methods were essentially the same as under normal input. The condition interaction for mAP50–95 ranged from 0.00 to +0.03 pp across the five severity levels on WSODD and from −0.03 to +0.01 pp on SDS, with all intervals containing zero. For APsmall, the condition interaction ranged from −0.72 to +0.42 pp on WSODD and from −0.05 to +0.48 pp on SDS, and these intervals also all contained zero; on SDS, the APsmall condition interaction at severity levels 1–3 was +0.30 to +0.48 pp, positive for all three seeds, yet the paired differences within these three conditions still ranged from −0.17 to +0.01 pp.

Condition evaluation yielded three results for the matched-control and FloW configurations. For GlareSoft copies relative to clean copies, the condition interaction on mAP50–95 was −0.01 pp (95% CI −0.09 to +0.08) under localized glare at severity level 3 and +0.01 pp under pretrained fine-tuning, providing no statistically supported glare-specific benefit. The complete configuration showed no statistically detectable difference from the scratch reference under normal input, and exceeded it by 0.83 pp (95% CI +0.35 to +1.32) at low-light severity level 4 and by 1.23 pp (95% CI +0.54 to +1.93) at level 5, with all six seeds positive. On FloW, paired-view training gave a condition interaction of +0.61 pp (95% CI −0.23 to +1.46) at low-light severity level 5, with five seeds positive; small-object-weighted paired-view training gave a condition interaction of +1.43 pp (95% CI +0.12 to +2.74) on

mAP50–95, and +1.49 pp (95% CI +0.57 to +2.40) on APsmall, each with five seeds positive. The effect of small-object weighting, therefore, appears on small objects under the strongest low light, rather than under normal input. Results for all configurations under all conditions are given in Table S3.

5. Discussion

5.1. Deciding Whether to Adopt an Improvement

The mean alone understates how much training randomness affects method selection, consistent with Bouthillier et al.'s finding that benchmark variability can change performance comparisons [6]; the repeated training in this study also quantifies that effect. On WSODD, fixed single-view augmentation and standard training shared the same seed, but did not move together: the sample standard deviation of their paired differences was 1.10 pp, the same as the 1.11 pp expected when the two configurations are treated as independently trained. For this specific comparison, the observed 95% CI half-width is about 1.16 pp. Assuming similar dispersion, approximately 22 seeds would be needed for a t-based CI half-width of 0.5 pp; this is an illustrative precision calculation, not a power analysis or a universal detection threshold. Against the same standard training reference, the case of clean copies added under pretrained fine-tuning is the opposite: two runs with the same seed were highly synchronized, with a sample standard deviation of paired differences of only 0.24 pp versus the 1.24 pp expected under independent training, so the 0.54 pp decrease is statistically supported with six seeds. How much training noise same-seed pairing cancels varies with the modification, and cannot be assumed in advance; the number of repetitions needed to judge a modification should rest on the dispersion of the paired differences themselves. The target metric also changes the choice: on WSODD, the near-zero overall effect comprises scale components of differing direction (Section 4.2), and whether to adopt a modification depends on the target scale of the application. Practically, under normal input, the mean mAP50–95 changes in all comparisons in Tables 4 and 5 fall within ±1 pp, the same order as seed-to-seed variation, and none represents a statistically supported improvement; larger differences appear under strong synthetic low light (Section 5.3).

5.2. What Matched Controls Add to Augmentation Studies

The complete configuration shows no statistically detectable difference from the scratch reference, but this does not establish equivalence or show that every step had no effect. At the level of estimated means, the negative loss-weight change and the positive sum of the following three changes approximately cancel. The choice of baseline, therefore, determines the conclusion a reader sees. Starting from the configuration with adjusted loss weights, the next three steps total +0.56 pp; starting from the scratch reference, the same model is −0.09 pp, the former numerically offsetting most of the first-step decrease, without establishing separate benefits of the later steps. The step that appends samples, likewise, needs its own reference: sample duplication alters the optimization process, and augmentation can also affect results through data volume and training randomness [8,25]. Under pretrained fine-tuning, GlareSoft copies differ from standard training by an average of −1.11 pp, about half of which (−0.54 pp) is already present when clean copies are appended, and only that half is statistically supported; comparing only the training set with GlareSoft copies added against the original training set would attribute this part of the drop to the glare transformation. Once copies or views are matched, what remains is the increment of the transformation itself. In this study, that increment is not statistically supported on either WSODD or FloW; for GlareSoft on small objects the increment is positive, with a lower interval bound slightly below zero, leaving substantial uncertainty that merits additional seeds, rather than a claim of benefit.

5.3. Connecting Detection Performance to Application Conditions

The value of illumination augmentation depends on input conditions, and the condition interaction separates this from the average effect under normal input. Fixed single-view augmentation shows no statistically supported overall-accuracy gain under normal input; its observed benefit appears as a reduction in the loss caused by synthetic low light. At the strongest severity level, 5, the standard training model's mAP50–95 on WSODD is 8.20 pp lower than under normal input, while the augmented model is only 3.11 pp lower, avoiding about three-fifths of the loss (40–70% across the six seeds); on SDS, the two are 3.98 and 2.30 pp lower, avoiding about two-fifths. Severity levels 1–3 fall within the training parameter range, where the standard training model's loss is under 0.5 pp and there is little room to recover overall accuracy, with gains in this range appearing only on small objects (Section 4.4); the augmentation's advantage emerges at levels 4 and 5, where the model begins to fail, and these levels lie beyond the parameter range seen during training, indicating that its effect extends to stronger degradation than that used in training. Evaluating normal inputs alone would miss the observed synthetic-condition benefit, whereas evaluating only strong synthetic low light would overstate the evidence for practical deployment. The real low-luminance subset did not reproduce the benefit, so its value under naturally occurring low light remains unresolved. Detection degradation benchmarks, likewise, show that normal and degraded input provide different performance information [26]. Synthetic localized glare hardly reduces accuracy at any of the five levels, so this evaluation cannot determine the effect of glare augmentation; Yalçın and Ekenel identified false detections associated with mirrored targets by removing mirrored reflections from water-surface images [5], and the impact of real reflections requires such recorded data to be tested.

5.4. Scope and Further Validation

Follow-up work should repeat the paired comparison and matched control on independently collected sources and on real low-light, glare and mirror-reflection images. Gains under synthetic low light remain unverified on real dark images; the four small subsets of the WSODD test set, defined by brightness and acquisition time, did not reproduce this direction (Table S2). The effect of GlareSoft on real glare remains untested. The current detector reference contains one training seed; the principal training strategies and matched controls used six seeds on WSODD and FloW and three on SDS, while the auxiliary pretrained component comparisons on FloW used three seeds, with no resampling of the data splits. The SDS evaluation in this study uses the validation set involved in checkpoint selection, and includes no independent test evaluation, so its results change direction with the checkpoint; WSODD and FloW use held-out test sets. Source grouping is limited by the available acquisition identifiers. Latency and memory results correspond to an RTX 4090 and this study's software environment, and require remeasurement for deployment on other hardware.

6. Conclusions

A unified evaluation across three water-surface datasets yields four findings. First, the observed ranking of four detectors changes with dataset and metric, but these single-seed comparisons are descriptive. Second, the average change from fixed single-view augmentation on WSODD is close to zero, with a paired 95% CI half-width of about 1.16 pp for this six-seed comparison. SDS results are auxiliary validation-based evidence, and change direction with checkpoint selection. Third, matched controls show how the estimated overall change depends on the reference: a mean 0.65 pp decrease from loss-weight adjustment is numerically offset by later steps, leaving a complete-configuration change of

−0.09 pp (95% CI −0.64 to +0.46). The loss-weight decrease has a t-based interval excluding zero, but a sign-flip p value of 0.0625. Matching appended samples or views reveals no statistically supported normal-input gain attributable to degraded content on WSODD or FloW; adding clean copies under pretrained fine-tuning, instead, lowers overall accuracy. Fourth, fixed single-view augmentation benefits WSODD under the strongest synthetic low light (+4.99 pp, positive in all six seeds), but this direction is not reproduced on the 160-image real low-luminance subset (−2.26 pp; 95% CI −4.70 to +0.18). These synthetic results, therefore, do not establish a benefit under real low-light deployment.

These results support a directly usable evaluation practice. When reporting an improvement, give paired per-seed differences with their confidence intervals and judge, from the spread of those differences, whether the number of seeds is sufficient, rather than citing a single training score; when evaluating an augmentation, include a clean control matched in sample count and processing pipeline, separate the effect of the transformation from that of the additional samples, and state the baseline used for comparison; when comparing training schemes, use input conditions and target scales that match the deployment environment, because differences invisible under normal input can reach several pp under degraded conditions. Under this practice, no training modification in the principal comparisons achieved a statistically supported overall-accuracy improvement under normal input, whereas fixed single-view augmentation showed a consistent benefit across seeds, under the strongest synthetic low light. Failure to detect an improvement is not proof of no effect, and real-world benefit remains to be established.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/machines14111373/s1, Table S1 contains all 621 evaluations by dataset, configuration, seed, and condition, together with the results for the real-condition subsets of WSODD. Tables S2 and S3 provide paired differences and condition interactions; the S2 folder also includes the within-seed attention-by-localization-loss interaction. Table S4 reports computational measurements and individual image timings. Table S5 reports target size composition, scale binning results, and per-class results. The S1 folder defines the real-condition subsets, and the S4 folder also includes the settings used for the efficiency measurements. Table S6 lists the training configuration of each experimental block, and the S6 folder provides the image lists for the training, validation, and test splits of each dataset, as well as the image lists used to generate the copies. S7 supplies the retained training and evaluation records. A supplementary verification script recomputes paired summaries from S1 and S2; it does not rerun training or inference.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, and project administration, J.L.; supervision and funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, Ministry of Science and Technology (sub-project), grant number 2024YFB3410605-1.

Institutional Review Board Statement: Not applicable. This study used publicly available benchmark datasets and involved no direct recruitment of human participants or collection of new identifiable personal data.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study used the publicly available SeaDronesSee, WSODD and FloW datasets, which can be obtained from their respective distributors. The Supplementary Materials report per-seed evaluation results, paired differences and condition interaction terms (Tables S1–S3), along with the training configuration for all experiments (Table S6), the image lists for the training, validation and test splits of each dataset (S6 folder), and the retained training and evaluation records (S7 folder); these configurations and lists allow the data splits and training settings of this study to be reconstructed. Evaluation used the official pycocotools COCOeval, with settings given in Table A4.

Model checkpoints, saved prediction files and implementation code are retained by the authors, and can be provided, upon reasonable request, to verify the reported results.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMP	Automatic mixed precision
AP	Average precision
CBAM	Convolutional block attention module
CI	Confidence interval
CIoU	Complete intersection over union
COCO	Common Objects in Context
DFL	Distribution focal loss
EIoU	Efficient intersection over union
FP32	32-bit floating-point precision
FPS	Frames per second
GFLOPs	Giga floating-point operations
HSV	Hue, saturation, and value
IoU	Intersection over union
LR	Learning rate
mAP	Mean average precision
MDG	Maritime degradation generation
NMS	Non-maximum suppression
pp	Percentage points
RGB	Red, green, and blue
RNG	Random number generator
RT-DETR	Real-time detection transformer
SD	Sample standard deviation
SDS	SeaDronesSee
SGD	Stochastic gradient descent
UAV	Unmanned aerial vehicle
WSODD	Water Surface Object Detection Dataset
YOLO	You Only Look Once

Appendix A. Fixed Training and Evaluation Parameters

Table A1 lists training settings by experimental block, Tables A2 and A3 specify the MDG and GlareSoft transformations, Table A4 records the fixed evaluation settings, and Table A5 lists the severity levels for controlled-condition evaluation. Training MDG acts after standard augmentation, whereas controlled test degradation is constructed at the original image resolution. Fixed single-view augmentation holds each image's MDG assignment constant; both paired-view variants use stochastic online MDG. The detector environment and complete FloW runtime records identify an NVIDIA GeForce RTX 4090, PyTorch 2.5.1+cu124, and torchvision 0.20.1+cu124.

Table A1. Training settings for the three experimental blocks. Each column describes a separate training recipe; the group-specific loss weights are given in Section 3.5. Batch size counts source images before paired-view processing. Epochs indicate the training limit, and the native validation procedure selects the saved checkpoint.

Setting	Detector Comparison	Training Strategies	WSODD Matched Controls from Scratch
Initialization	General-purpose pretrained weights	Pretrained YOLOv8s	Random initialization from YAML
Ultralytics version	8.4.124	8.4.124	8.3.182
Input size/source batch	960/8	960/8	960/4

Table A1. *Cont.*

Setting	Detector Comparison	Training Strategies	WSODD Matched Controls from Scratch
Maximum epochs/patience	150/150	150/100	150/100
Optimizer	AdamW	SGD	SGD
Initial learning rate	SDS: 0.001111; WSODD: 0.000556; FloW: 0.002	0.006	0.006
Final learning-rate factor	0.01	0.01	0.01
Momentum	0.9	0.937	0.937
Weight decay	0.0005	0.0000625	0.0000625
Nominal batch size	64	8	8
Warmup epochs/momentum/bias LR	3/0.8/0.1	3/0.8/0.1	3/0.8/0.1
HSV gains (h, s, v)	0.015, 0.7, 0.4	0.015, 0.6, 0.2	0.015, 0.6, 0.2
Translate/scale magnitude	0.1/0.5	0.1/0.45	0.1/0.45
Horizontal/vertical flip probability	0.5/0	0.5/0	0.5/0
Mosaic/mixup probability	1.0/0	0.7/0.05	0.7/0.05
Mosaic disabled	Last 10 epochs	Last 10 epochs	Last 10 epochs
Rotation/shear/perspective	0/0/0	0/0/0	0/0/0
Training precision	YOLO: AMP; RT-DETR-L: FP32	AMP	AMP
Checkpoint selection	Native validation mAP50–95	Native validation mAP50–95	0.1·mAP50 + 0.9·mAP50–95 validation fitness

Table A2. MDG training parameter ranges. Parameters are sampled uniformly within each range, and the three ranges are selected with equal probability. The branch probability determines whether an image receives low light, localized glare, both transformations, or no additional MDG; for the fixed single-view augmentation, this assignment is fixed for each image, whereas the FloW paired-view experiments sample it online.

Parameter Band	Brightness b	Gamma γ	Glare Target-Area Ratio	Intensity (0–255 Scale)
1 (lower)	0.90–0.95	1.05–1.12	0.0175–0.028	22–30
2 (middle)	0.86–0.90	1.12–1.21	0.028–0.039	30–39
3 (upper)	0.82–0.86	1.21–1.30	0.039–0.050	39–48
Common glare geometry	—	—	Aspect ratio: 1.4–3.2	Attenuation parameter: 1.6–2.4
Branch probabilities	Low light: 0.20	Glare: 0.20	Mixed: 0.15	No additional MDG: 0.45

Table A3. Transformation and matching settings for training and controlled evaluation. Area ratios specify the target area used to construct each ellipse; integer radii and image bounds determine the resulting patch. The clean-copy control and GlareSoft augmentation share the selected images and image-processing steps, with the glare transformation applied only to the GlareSoft copies.

Operation	Setting
Training low light	$x' = b \cdot x^{\gamma}$ for normalized pixels x
Training glare mask	$m = \exp(-k \cdot d) \cdot 1[d \leq 1]$, with $d = ((x-c_x)/r_x)^2 + ((y-c_y)/r_y)^2$
MDG target geometry	$A = r \cdot H \cdot W$; $r_y = \max(2, \text{floor}(\text{sqrt}(A/(\pi a))))$; $r_x = \max(3, \text{floor}(a \cdot r_y))$; radii limited to image bounds
MDG application	Add m·intensity/255, clip to 0–1; mixed branch applies low light before glare

Table A3. *Cont.*

Operation	Setting
Fixed single-view assignment	SHA-256(dataset namespace + null separator + image identifier); independent local Python RNG
FloW paired-view inputs	Eight source images concatenated with eight stochastic MDG views; copied labels; shared detector
FloW small-object weighting	1.5 for degraded-view targets below 1024 px^2 after augmentation; 1.0 otherwise; box/cls/DFL weighted, normalization unchanged
FloW identical paired views	MDG parameters sampled as in paired views; after preprocessing, the second view is replaced by a copy of the first; number of views, updates per epoch, label copying, and loss normalization unchanged
GlareSoft sample count	397 = floor(0.10 × 3977); both copy configurations contain 4374 training images
GlareSoft selection	NumPy default_rng(42), sampling without replacement; shared selection for all six training seeds
GlareSoft per-image RNG	NumPy default_rng(42 + source index); same fixed generation for all training seeds
GlareSoft transform probabilities	Low light: 0; localized glare: 1; combined transform disabled
GlareSoft target geometry	A_{max} = max(0.035·H·W,64); A uniform in 0.35·A_{max} to A_{max}; a uniform in 1.4–3.2
GlareSoft radii	r_y = max(10,floor(sqrt(A/(πa)))); r_x = max(14,floor(a·r_y))
GlareSoft light parameters	Intensity uniform in 16–34; k uniform in 1.6–2.4; same elliptical mask form
Clean/GlareSoft processing match	Same copied labels, image identities, filename suffixes, and ordering; PIL RGB input; saving quality 95
Controlled-test low light	rint(255 × b × $(v/255)^{\gamma}$), clipped to 0–255; b and γ for each severity level in Table A5
Controlled-test localized glare	Target-area ratio and central intensity for each severity level in Table A5; aspect ratio 2.3; k = 2.0; rint and clip to 0–255
Test glare position	Fixed per image from a hash of dataset namespace and filename; shared across models and seeds

Table A4. Scoring and inference settings used for the unified re-evaluation. All three experimental blocks share the same COCOeval implementation, area ranges, prediction parameters, input batch, precision, and padding settings; model-loading versions are listed by experimental block. The table itemizes each setting and its value.

Setting	Unified Scoring and Inference
Evaluator	Official pycocotools 2.0.10 COCOeval
Input size/actual batch	960 × 960/1
Precision/rectangular auto-padding	FP32/disabled
Confidence threshold	0.001
NMS IoU, where applicable	0.7
Prediction limit	300
Evaluation detection limit	COCO maxDets = (1, 10, 100)
IoU thresholds/recall points	0.50:0.05:0.95/101
Area coordinates	Original-image pixels
Small/medium area boundaries	32^2/96^2 px^2

Table A4. *Cont.*

Setting	Unified Scoring and Inference
Model-loading implementation	Ultralytics 8.4.124; matched controls: 8.3.182
Checkpoint selection	Unchanged saved best.pt files; last.pt additionally for the six SDS runs
Condition construction	Fixed original-resolution transforms
Repeated-run dispersion	Sample SD, denominator $n - 1$
Paired-difference interval	95% CI of the mean from a t distribution with $n - 1$ degrees of freedom

Table A5. Severity levels for the evaluation of controlled conditions. Levels 1–3 are the midpoints of the three training-parameter bands, while levels 4 and 5 lie beyond the training range. For localized glare, the aspect ratio is 2.3 and the attenuation coefficient is 2.0 at every level.

Severity Level	Brightness b	Gamma γ	Glare Target-Area Ratio	Intensity (0–255 Scale)	Relation to the Training Range
1	0.925	1.085	0.02275	26.0	Midpoint of training band 1
2	0.88	1.165	0.0335	34.5	Midpoint of training band 2
3	0.84	1.255	0.0445	43.5	Midpoint of training band 3
4	0.70	1.50	0.08	80.0	Beyond the training range
5	0.50	2.00	0.15	120.0	Beyond the training range

References

1. Varga, L.A.; Kiefer, B.; Messmer, M.; Zell, A. SeaDronesSee: A Maritime Benchmark for Detecting Humans in Open Water. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Waikoloa, HI, USA, 2022; pp. 3686–3696. [CrossRef]
2. Zhou, Z.; Sun, J.; Yu, J.; Liu, K.; Duan, J.; Chen, L.; Chen, C.L.P. An Image-Based Benchmark Dataset and a Novel Object Detector for Water Surface Object Detection. *Front. Neurorobot.* **2021**, *15*, 723336. [CrossRef] [PubMed]
3. Cheng, Y.; Zhu, J.; Jiang, M.; Fu, J.; Pang, C.; Wang, P.; Sankaran, K.; Onabola, O.; Liu, Y.; Liu, D.; et al. FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Montreal, QC, Canada, 2021; pp. 10933–10942. [CrossRef]
4. Sun, H.; Zhao, H.; Liu, Z.; Jiang, G.; Zhao, J. WA-YOLO: Water-Aware Improvements for Maritime Small-Object Detection Under Glare and Low-Light. *J. Mar. Sci. Eng.* **2026**, *14*, 37. [CrossRef]
5. Yalçın, S.; Ekenel, H.K. Impact of Surface Reflections in Maritime Obstacle Detection. *arXiv* **2024**, arXiv:2410.08713. [CrossRef]
6. Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Mohammadi Sepahvand, N.; Raff, E.; Madan, K.; Voleti, V.; et al. Accounting for Variance in Machine Learning Benchmarks. *Proc. Conf. Mach. Learn. Syst.* **2021**, *3*, 747–769.
7. Reimers, N.; Gurevych, I. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-Networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 338–348. [CrossRef]
8. Hoffer, E.; Ben-Nun, T.; Hubara, I.; Giladi, N.; Hoefler, T.; Soudry, D. Augment Your Batch: Improving Generalization Through Instance Repetition. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 8126–8135. [CrossRef]
9. Kiefer, B.; Kristan, M.; Pers, J.; Zust, L.; Poiesi, F.; De Alcantara Andrade, F.A.; Bernardino, A.; Dawkins, M.; Raitoharju, J.; Quan, Y.; et al. 1st Workshop on Maritime Computer Vision (MaCVi) 2023: Challenge Results. In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*; IEEE: New York, NY, USA, 2023; pp. 265–302. [CrossRef]
10. Bovcon, B.; Muhovic, J.; Vranac, D.; Mozetic, D.; Pers, J.; Kristan, M. MODS—A USV-Oriented Object Detection and Obstacle Segmentation Benchmark. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 13403–13418. [CrossRef]
11. Miri Rekavandi, A.; Xu, L.; Boussaid, F.; Seghouane, A.-K.; Hoefs, S.; Bennamoun, M. A Guide to Image- and Video-Based Small Object Detection Using Deep Learning: Case Study of Maritime Surveillance. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 2851–2879. [CrossRef]

12. Žust, L.; Perš, J.; Kristan, M. LaRS: A Diverse Panoptic Maritime Obstacle Detection Dataset and Benchmark. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 20247–20257. [CrossRef]
13. Tran, M.; Shipard, J.; Mulyono, H.; Wiliem, A.; Fookes, C. SafeSea: Synthetic Data Generation for Adverse & Low Probability Maritime Conditions. In *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*; IEEE: Waikoloa, HI, USA, 2024; pp. 821–829. [CrossRef]
14. Zhang, M.; Armin, M.A.; Li, X.; Liang, S.; Petersson, L.; Sun, C.; Ahmedt-Aristizabal, D.; Hayder, Z. Quality-Driven and Diversity-Aware Sample Expansion for Robust Marine Obstacle Segmentation. In *Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*; IEEE: New York, NY, USA, 2026; pp. 1136–1145. [CrossRef]
15. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 16965–16974. [CrossRef]
16. Jocher, G.; Qiu, J.; Liu, M.; Lyu, S.; Akyon, F.C.; Kalfaoglu, M.E. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. *arXiv* **2026**, arXiv:2606.03748. [CrossRef]
17. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 11211, pp. 3–19. [CrossRef]
18. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and Efficient IOU Loss for Accurate Bounding Box Regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
19. Zhang, J.; Jin, J.; Ma, Y.; Ren, P. Lightweight Object Detection Algorithm Based on YOLOv5 for Unmanned Surface Vehicles. *Front. Mar. Sci.* **2023**, *9*, 1058401. [CrossRef]
20. Senussi, M.F.; Heikal, A.S.; Abdelbasset, A.G.; Kang, H.-S. Robust Multi-Class Lumpy Skin Disease Diagnosis for Practical Livestock Applications. *Sci. Rep.* **2026**, *16*, 26662. [CrossRef] [PubMed]
21. Zoph, B.; Cubuk, E.D.; Ghiasi, G.; Lin, T.-Y.; Shlens, J.; Le, Q.V. Learning Data Augmentation Strategies for Object Detection. In *Proceedings of the Computer Vision—ECCV 2020*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 566–583. [CrossRef]
22. Li, L.; Hou, Z.; Ma, M.; Xiang, J.; Yuan, C.; Xia, G. Spotlight on Small-Scale Ship Detection: Empowering YOLO with Advanced Techniques and a Novel Dataset. In *Proceedings of the Computer Vision—ACCV 2024*; Cho, M., Laptev, I., Tran, D., Yao, A., Zha, H., Eds.; Springer Nature Singapore: Singapore, 2025; Volume 15477, pp. 3–17. [CrossRef]
23. Summers, C.; Dinneen, M.J. Nondeterminism and Instability in Neural Network Optimization. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 9913–9922.
24. Gontijo-Lopes, R.; Smullin, S.J.; Cubuk, E.D.; Dyer, E. Tradeoffs in Data Augmentation: An Empirical Study. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
25. Geiping, J.; Goldblum, M.; Somepalli, G.; Shwartz-Ziv, R.; Goldstein, T.; Wilson, A.G. How Much Data Are Augmentations Worth? An Investigation into Scaling Laws, Invariance, and Implicit Regularization. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
26. Michaelis, C.; Mitzkus, B.; Geirhos, R.; Rusak, E.; Bringmann, O.; Ecker, A.S.; Bethge, M.; Brendel, W. Benchmarking Robustness in Object Detection: Autonomous Driving When Winter Is Coming. *arXiv* **2019**, arXiv:1907.07484. [CrossRef]
27. Hendrycks, D.; Mu, N.; Cubuk, E.D.; Zoph, B.; Gilmer, J.; Lakshminarayanan, B. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
28. Zhuang, D.; Zhang, X.; Song, S.; Hooker, S. Randomness in Neural Network Training: Characterizing the Impact of Tooling. In Proceedings of the Conference on Machine Learning and Systems (MLSys 2022), Santa Clara, CA, USA, 29 August–1 September 2022; Volume 4, pp. 316–336.
29. Dodge, J.; Gururangan, S.; Card, D.; Schwartz, R.; Smith, N.A. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 2185–2194. [CrossRef]
30. Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Lariviere, V.; Beygelzimer, A.; d'Alche-Buc, F.; Fox, E.; Larochelle, H. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *J. Mach. Learn. Res.* **2021**, *22*, 1–20.
31. MaCVi. SeaDronesSee Object Detection v2—Official Dataset Documentation. Available online: https://www.macvi.org/dataset (accessed on 20 September 2026).
32. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8; Ultralytics, 2023. Available online: https://docs.ultralytics.com/models/yolov8/ (accessed on 20 September 2026).

33. Jocher, G.; Qiu, J. Ultralytics YOLO11; Ultralytics, 2024. Available online: https://docs.ultralytics.com/models/yolo11/ (accessed on 20 September 2026).
34. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Proceedings of the Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; Volume 8693, pp. 740–755. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.