Qassim University جامعة القصيم

JEIT

Journal of Emerging and Innovative Technologies

Volume 1, Issue 1
July–December 2026

Received
22 September 2026

Accepted
20 October 2026

Published
05 November 2026

DOI
https://doi.org/10.71422/jeit.0005

License
CC BY-NC-SA 4.0

Cite this article

Almudaymigh AA, Sameem MSI. Evaluating Low-Light Image Enhancement for Object Detection. J Emerg Innov Technol. 2026;1(1):47–62.

Journal website

Research article

Evaluating Low-Light Image Enhancement for Object Detection

Abdulrahman Abdulaziz Almudaymigh[1,*], M. Shujah Islam Sameem[2]

[1] Artificial Intelligence Program, College of Computer Sciences and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia

[2] Department of Computer Science, College of Computer Sciences and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia

*Corresponding author: 226031131@student.kfu.edu.sa

Abstract

Low-light restoration is often placed before an object detector, yet an apparent benefit or loss can reflect both the image transformation and a mismatch between the detector training domain and the transformed test images. We evaluated Original ExDark images, Gamma correction, CLAHE, Zero-DCE, and Retinexformer with YOLO11n. An inference-only comparison used a detector trained on Original images, while matched training trained a separate detector within each image domain. Matched training was repeated with seeds 42, 123, and 2026. Average precision was recomputed from prediction streams retained down to confidence 0.001; precision, recall, and F1 at confidence 0.25 were reported separately. Seed-wise paired image bootstraps, class and subgroup analyses, validation-only sensitivity checks, NIQE/BRISQUE, and an adaptive selector were also examined. In the fixed-threshold inference-only comparison, F1 fell from 0.642 on Original images to 0.619, 0.628, 0.586, and 0.566 for Gamma, CLAHE, Zero-DCE, and Retinexformer. Under matched training, mean mAP@0.50 across three seeds was 0.6232±0.0037 for Original, 0.6161±0.0114 for Gamma, 0.6194±0.0003 for CLAHE, 0.6182±0.0027 for Zero-DCE, and 0.5862±0.0055 for Retinexformer. Differences among Original, Gamma, CLAHE, and Zero-DCE changed with training seed, whereas Retinexformer remained lower in all three runs. For the evaluated ExDark split, YOLO11n detector, and enhancement configurations, domain-matched training recovers much of the fixed-threshold performance loss observed under inference-only preprocessing, but no enhancement provides a stable overall AP gain. Perceptual improvement is therefore not a reliable substitute for task-level evaluation.

Keywords: computer vision; image enhancement; low-light imaging; object detection; image quality assessment; ExDark

1. Introduction

Object detection systems increasingly operate in scenes where illumination is uncontrolled. Dark images can suppress object boundaries, distort color and texture cues, and lower the signal-to-noise ratio. ExDark was designed for this setting using naturally captured low-light images rather than synthetic darkening; it contains 7,363 images, 12 object categories, bounding-box annotations, and ten illumination labels [1].

A common engineering response is to enhance the image before detection. That choice is reasonable for human viewing, but it combines two objectives that need not coincide.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

Gamma correction and CLAHE alter intensity or local contrast, while learned restorers such as Zero-DCE and Retinexformer optimize restoration-oriented objectives [3], [4]. A detector, however, is sensitive to the feature distribution learned during training. A visually improved image may therefore be easier to inspect while becoming less familiar to a detector.

Recent low-light detection research increasingly addresses this mismatch directly. Zero-shot day-night adaptation learns illumination-invariant representations [18]; other systems couple enhancement with detector-specific feature fusion [19], combine image enhancement with domain adaptation [20], or align image- and feature-level domains jointly [21]. These methods are task-aware. They differ from the practical situation considered here, where a generic restoration step is inserted before a detector and the researcher wants to know whether the apparent effect comes from enhancement itself or from train-test mismatch.

The present study does not propose a new enhancement or detection architecture. The contribution lies in the evaluation design rather than in a new model. We compare the same ExDark images and annotations under five image domains, separate inference-only preprocessing from domain-matched detector training, repeat matched training across independent seeds, quantify image-sampling uncertainty separately from training-run variation, and compare no-reference perceptual quality with downstream detection. Together, these controls address a narrower question than task-aware adaptation: how much of the apparent enhancement effect remains once the detector is trained on the transformed image domain?

We therefore ask four questions. First, does generic low-light enhancement improve detection once train-test mismatch is controlled? Second, are any gains stable across training seeds, object classes, illumination conditions, and object sizes? Third, do NIQE and BRISQUE track image-level detection utility? Fourth, can a simple validation-trained selector exploit image-wise complementarity without access to test labels? All conclusions are restricted to ExDark, YOLO11n, and the enhancement settings evaluated here.

2. Related Work

2.1 Low-light enhancement and benchmarks

ExDark provides real low-light scenes with class and illumination labels, making it useful for testing recognition robustness as well as enhancement [1]. Broader enhancement benchmarks have shown that restoration rankings depend strongly on the dataset and evaluation criterion [2]. Classical Gamma correction and CLAHE remain useful controls because their transformations are deterministic and transparent. Zero-DCE learns image-specific enhancement curves without paired references [3], whereas Retinexformer combines Retinex decomposition with illumination-guided Transformer blocks [4]. HVI-CIDNet illustrates the continued progress of restoration methods aimed primarily at perceptual quality [5].

Because ExDark lacks paired normal-exposure references for its full test set, full-reference measures such as PSNR and SSIM are not appropriate as primary quality criteria. We therefore use task metrics for detection and no-reference NIQE and BRISQUE as complementary perceptual indicators [14], [15].

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

2.2 Enhancement, task-aware adaptation, and detection

Previous work shows that enhancement quality and machine-perception performance do not necessarily align. Al Sobbahi and Tekli showed that enhancement rankings change with task and evaluation setting [6]. NOD and GenISP likewise highlight that images preferred by a machine need not be those preferred by a human observer [7], [8]. Selective enhancement has been explored to avoid applying one transformation to every image [9], and other work directly compares restoration models for low-light detection [10].

More recent studies move beyond generic preprocessing. LoLI-Street broadens the evaluation of low-light enhancement and downstream tasks [11]. GEFU explicitly bridges enhancement and high-level understanding through semantically consistent fine-tuning [12], while DBDiff decouples generative restoration from detector-oriented adaptation and then reconnects them through task-alignment modules [13]. DAI-Net learns illumination-invariant representations for zero-shot day-night adaptation [18]. DAMFCN/DarkYOLOv8 couples adaptive enhancement with detector feature fusion [19], domain-adaptive enhancement explicitly reduces normal/low-light domain discrepancy [20], and DLDA aligns both image-level appearance and feature-level representations [21].

Our objective is narrower than that of these task-aware systems. We do not optimize enhancement using detector loss or perform domain adaptation between normal and dark domains. Instead, we hold the detector family, annotations, partitions, and training recipe fixed and ask whether generic restoration remains beneficial after the detector is retrained on the restored domain. The study therefore complements, rather than competes with, task-aware methods.

3. Materials and Methods

3.1 Dataset, annotations, and partitions

We used the official ExDark images and annotations [1]. The dataset comprises 7,363 images from 12 categories (Bicycle, Boat, Bottle, Bus, Car, Cat, Chair, Cup, Dog, Motorbike, People, and Table) and ten illumination conditions. The official imageclasslist partition was retained throughout: 3,000 images for training, 1,800 for validation, and 2,563 for test. The test set contains 7,117 annotated object instances. An integrity audit of the split lists found no duplicate test filenames and no missing illumination labels. All five image domains use the same image identities and bounding-box coordinates.

3.2 Enhancement domains and configuration selection

Five image domains were evaluated. Original contains the unmodified ExDark images. Gamma used γ=0.60. CLAHE was applied to the L channel of CIELAB with clip limit 2.0 and an 8×8 tile grid. Zero-DCE used the authors' Epoch99 checkpoint [3]. Retinexformer used the LOL-v2-real checkpoint with n_feat=40, one stage, and blocks [1,2,2] [4]. Zero-DCE used tiled GPU inference (tile 768, halo 32) with FP16; Retinexformer used tiled inference (tile 256, halo 48) in FP32 after a numerical audit described below.

These settings were fixed before the test-set analysis; they were not selected by optimizing test performance. We also performed a post-hoc validation-only sensitivity check for Gamma (0.4–0.8) and CLAHE clip limits (1.5–3.0). This sensitivity check is descriptive and does not replace the prespecified test results. Because the selected settings were not the validation optimum, the conclusions are

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

explicitly configuration-specific. For the learned enhancers, one published checkpoint was evaluated per method; no checkpoint sweep was performed.

Enhanced outputs retained the source image dimensions and filenames, and labels were copied without geometric modification. The pipeline preserved the original file extension. PNG outputs were lossless; JPG/JPEG outputs were re-encoded through OpenCV using its default encoder settings. The test set contains 1,857 .jpg, 263 .jpeg, and 443 .png images. This storage step is part of the evaluated configuration and is acknowledged as a limitation rather than treated as a property of the enhancement algorithms themselves.

3.3 Detector training and reproducibility

YOLO11n was used as the common detector [16]. Fifteen matched-training runs were completed (five image domains × three seeds). Every run started from yolo11n.pt, used 640×640 inputs, batch size 8, two data-loader workers, and the same augmentation and optimization arguments. Training was capped at 120 epochs with early stopping patience 25. The only intended training argument that varied across the 15 runs was the seed (42, 123, or 2026). Ultralytics 8.4.156 ran with Python 3.11.9 and PyTorch 2.5.1+cu121 on an NVIDIA GeForce GTX 1660 SUPER (6 GB).

The training request used optimizer=auto. Ultralytics resolved this to AdamW with learning rate 6.25×10^{-4} and momentum 0.9. The stored arguments were weight decay 5×10^{-4}, lrf=0.01, warm-up 3 epochs, warm-up momentum 0.8, warm-up bias learning rate 0.1, HSV gains 0.015/0.7/0.4, translation 0.1, scale 0.5, horizontal flip 0.5, mosaic 1.0 with close_mosaic=10, mixup 0, erasing 0.4, and deterministic=True. AMP was requested, but Ultralytics disabled mixed precision after its compatibility check on the GTX 1660 SUPER. The best validation checkpoint was used for each run. Resolved arguments and best epochs for all method-seed pairs are provided in Supplementary Table S5.

Table 1. Core experimental settings used for detector training and evaluation.

Item	Resolved setting
Dataset / split	ExDark; 3000 train / 1800 validation / 2563 test
Detector	YOLO11n initialized from yolo11n.pt; imgsz=640; batch=8; workers=2
Training	max 120 epochs; early stopping patience 25; seeds 42, 123, 2026
Resolved optimizer	AdamW; lr=6.25×10^{-4}; momentum=0.9; weight decay=5×10^{-4}
Augmentation	HSV 0.015/0.7/0.4; translate 0.1; scale 0.5; fliplr 0.5; mosaic 1.0; mixup 0
AP prediction stream	confidence minimum 0.001; NMS IoU=0.70; AP at IoU 0.50 and 0.50:0.95
Fixed operating point	confidence=0.25; class-aware matching at IoU=0.50
Software	Ultralytics 8.4.156; PyTorch 2.5.1+cu121; Python 3.11.9; Windows 11

3.4 Evaluation protocols

Figure 1 separates the two protocols. The inference-only protocol uses the Original-trained seed-42 detector and applies it to Original, Gamma, CLAHE, Zero-DCE, and Retinexformer images. Because those saved prediction streams were generated at confidence 0.25, this comparison is retained only as

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

an illustrative fixed-operating-point analysis of precision, recall, and F1; we no longer interpret its truncated prediction streams as AP. The primary analysis is matched training, in which each detector is trained and tested within the same image domain and repeated for seeds 42, 123, and 2026.

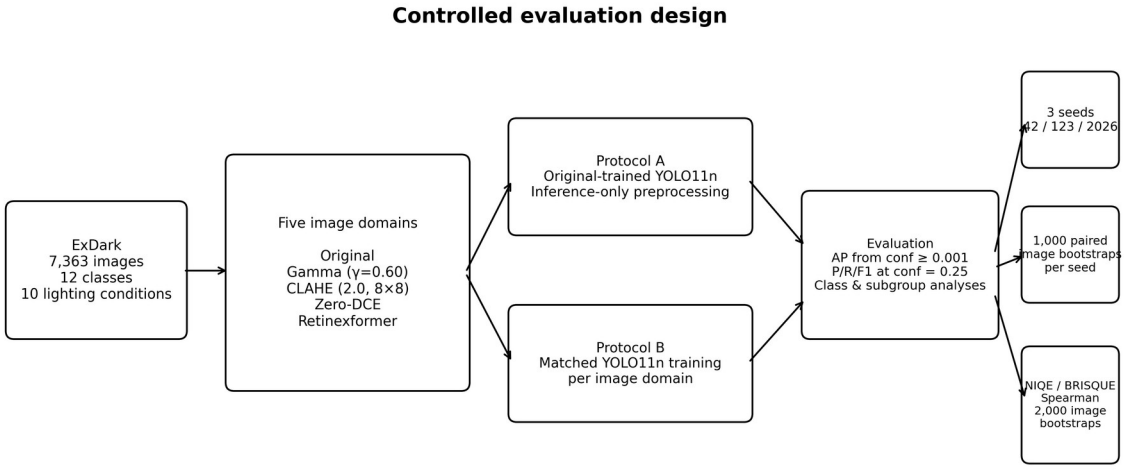


Figure 1. Experimental design. The inference-only arm measures fixed-operating-point behavior when preprocessing is inserted before an Original-trained detector; the matched arm trains and evaluates a separate detector within each image domain.

3.5 AP, fixed-threshold metrics, and uncertainty

Average precision was recomputed from low-confidence prediction streams retained down to confidence 0.001 after NMS at IoU 0.70. For each class, predictions were ranked by confidence and greedily matched one-to-one to ground-truth boxes of the same class. AP@0.50 used IoU 0.50; AP@0.50:0.95 averaged AP over IoU thresholds 0.50, 0.55, ..., 0.95. mAP is the unweighted mean across the 12 classes. Images without an instance of a given class contribute no positive ground truth for that class but their predictions remain eligible false positives.

Precision, recall, and F1 at confidence 0.25 are reported as a separate deployment-style operating point. At IoU 0.50, predictions are sorted by confidence and matched class-aware, one-to-one, to the unmatched ground-truth box with the highest IoU. The global counts are then micro-aggregated over all images: Precision=TP/(TP+FP), Recall=TP/(TP+FN), and F1=2PR/(P+R). For image-level analyses, the same matching rule is applied within each image; zero denominators evaluate to zero. Validation-selected operating points were obtained by sweeping confidence from 0.05 to 0.75 in steps of 0.05, selecting the threshold with maximum validation F1 separately for each trained model, and applying that threshold unchanged to test.

Two sources of uncertainty are deliberately kept separate. Training-run variability is summarized descriptively as mean±sample SD across the three independent seeds. Image-sampling uncertainty is evaluated conditionally on a fixed trained model using 1,000 paired bootstrap resamples per seed. Each bootstrap resamples 2,563 test image identities with replacement; the same image multiplicities are applied to every method within that seed, AP is recomputed, and 2.5th/97.5th percentiles form the 95% interval. These intervals do not estimate the population distribution of retraining outcomes. With only three training runs, we therefore avoid significance claims about the training-run distribution.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

3.6 Class, illumination, and object-size analyses

Class-wise AP@0.50 and AP@0.50:0.95 were computed for all 12 classes and all three seeds. Illumination analyses use the single ExDark illumination label assigned to each test image; no test image has a missing label. For object size, each ground-truth box is mapped to the 640-pixel evaluation canvas with aspect-ratio-preserving scaling. COCO thresholds are then applied: small area $<32^2$, medium $<96^2$, and large $\geq 96^2$ pixels. The test set contains 570 small (8.0%), 2,529 medium (35.5%), and 4,018 large (56.5%) instances. Full support counts and seed-wise results are supplied in Supplementary Tables S1–S3.

3.7 Perceptual quality and image-level utility

NIQE and BRISQUE were computed with the MATLAB-compatible implementations provided by PyIQA [17]. Images were loaded as RGB tensors and evaluated without geometric resizing beyond each metric's internal preprocessing; lower scores indicate better predicted perceptual quality [14], [15]. Image-level detection utility was computed at confidence 0.25 and IoU 0.50 using the matching rules above. For each method and image, precision, recall, and F1 were first computed for seeds 42, 123, and 2026 and then averaged across seeds. Spearman correlations relate these three-seed mean utilities to NIQE or BRISQUE. Correlation uncertainty was estimated by 2,000 image-level bootstrap resamples with replacement. Method-level IQA-versus-mAP comparisons involve only five methods and are treated descriptively, not as population-level correlation tests.

3.8 Adaptive selector and retrospective oracle

The selector is a secondary inference-only analysis based on the originally saved confidence-0.25 prediction streams. Ten features are extracted from the unmodified image and Original-detector output: mean luminance, luminance SD, mean saturation, entropy, Laplacian variance, dark-pixel ratio, bright-pixel ratio, Original detection count, mean Original detection confidence, and maximum Original detection confidence. Validation labels are the method with the highest image-level F1; ties prefer Original and then the fixed method order. A RandomForestClassifier with 300 trees, max_depth=10, min_samples_leaf=5, class_weight=balanced_subsample, random_state=42, and n_jobs=-1 is fitted on 1,800 validation images and then applied to the 2,563 test images. Classifier prediction alone takes 0.033 ms/image, excluding image-feature extraction and detector inference.

For context, we also compute a retrospective oracle that chooses the method with the highest image-level F1 using test ground truth. The oracle is explicitly non-deployable; it is reported only as an upper bound on image-wise complementarity and is not interpreted as an ordinary deployable detector.

3.9 Retinexformer integrity audit

An early Retinexformer FP16 run generated non-finite values that became all-black outputs after casting. Those outputs were discarded. Retinexformer was regenerated in FP32 and audited before any final evaluation. The audit checked file completeness and all-black outputs in every partition: 3,000/3,000 train, 1,800/1,800 validation, and 2,563/2,563 test images were present, with zero all-black outputs in each partition. The audit passed before the matched-training, multi-seed, IQA, and final analyses were run.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

4. Results

4.1 Primary matched-training results across seeds

The three-seed matched comparison is the primary evidence. Using AP recomputed from confidence 0.001 prediction streams, Original achieved mean mAP@0.50 of 0.6232±0.0037. CLAHE (0.6194±0.0003), Zero-DCE (0.6182±0.0027), and Gamma (0.6161±0.0114) remained close, while Retinexformer was lower at 0.5862±0.0055. The same pattern appears for mAP@0.50:0.95: Original 0.3785±0.0017, Zero-DCE 0.3781±0.0021, CLAHE 0.3754±0.0008, Gamma 0.3751±0.0059, and Retinexformer 0.3523±0.0033.

Table 2. Matched-training test performance across three seeds. AP uses confidence-minimum 0.001; P/R/F1 are fixed at confidence 0.25. Values are mean ± SD across seeds.

Method	mAP@0.50	mAP@0.50:0.95	Precision@0.25	Recall@0.25	F1@0.25
Original	0.6232 ± 0.0037	0.3785 ± 0.0017	0.6653 ± 0.0182	0.6321 ± 0.0070	0.6481 ± 0.0081
Gamma	0.6161 ± 0.0114	0.3751 ± 0.0059	0.6479 ± 0.0224	0.6271 ± 0.0085	0.6373 ± 0.0152
CLAHE	0.6194 ± 0.0003	0.3754 ± 0.0008	0.6581 ± 0.0088	0.6269 ± 0.0044	0.6421 ± 0.0024
Zero-DCE	0.6182 ± 0.0027	0.3781 ± 0.0021	0.6661 ± 0.0051	0.6211 ± 0.0043	0.6428 ± 0.0045
Retinexformer	0.5862 ± 0.0055	0.3523 ± 0.0033	0.6467 ± 0.0088	0.5967 ± 0.0108	0.6206 ± 0.0026

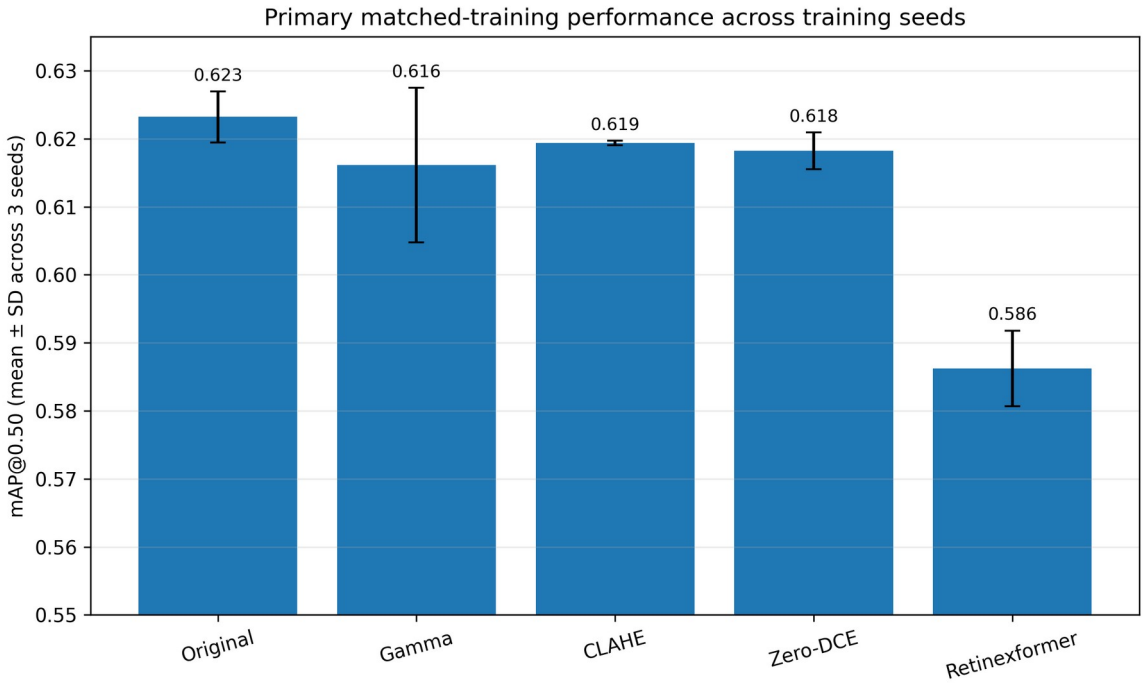


Figure 2. Primary matched-training mAP@0.50 across the three training seeds. Error bars are SD across training runs, not image-bootstrap confidence intervals.

The seed-wise differences show why the runs should not be collapsed into a single inferential claim. Relative to Original, Gamma changed by −0.0138, +0.0042, and −0.0116 mAP@0.50 across seeds 42, 123, and 2026; Zero-DCE changed by +0.0017, −0.0055, and −0.0111. CLAHE remained within about 0.008 of Original in every run. Retinexformer was lower in all three seeds (−0.0353, −0.0313, and −0.0443). Thus, the ranking among Original, Gamma, CLAHE, and Zero-DCE is training-run dependent, whereas the Retinexformer deficit is consistent in this experimental configuration.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

4.2 Conditional paired bootstrap within each trained run

The paired bootstrap addresses test-image sampling only. Table 3 therefore reports seed-specific intervals rather than one pooled "significance" result. For seed 42, the Gamma and Retinexformer mAP@0.50 differences exclude zero; for seed 123 only the Retinexformer difference excludes zero; for seed 2026 the Gamma, Zero-DCE, and Retinexformer differences exclude zero. These results are conditional on the trained models and should not be read as confidence intervals over retraining.

Table 3. Seed-wise paired image-bootstrap differences in mAP@0.50 relative to the Original domain. Each row is conditional on one fixed trained model pair; 1,000 paired image resamples per seed.

Training seed	Method	Δ mAP@0.50	95% image-bootstrap CI	CI excludes 0
42	Gamma	-0.0138	[-0.0255, -0.0029]	Yes
42	CLAHE	+0.0000	[-0.0098, +0.0107]	No
42	Zero-DCE	+0.0017	[-0.0090, +0.0126]	No
42	Retinexformer	-0.0353	[-0.0464, -0.0236]	Yes
123	Gamma	+0.0042	[-0.0068, +0.0151]	No
123	CLAHE	-0.0041	[-0.0150, +0.0061]	No
123	Zero-DCE	-0.0055	[-0.0169, +0.0056]	No
123	Retinexformer	-0.0313	[-0.0433, -0.0206]	Yes
2026	Gamma	-0.0116	[-0.0207, -0.0011]	Yes
2026	CLAHE	-0.0075	[-0.0181, +0.0034]	No
2026	Zero-DCE	-0.0111	[-0.0217, -0.0003]	Yes
2026	Retinexformer	-0.0443	[-0.0552, -0.0328]	Yes

4.3 Inference-only preprocessing is an illustrative fixed-threshold comparison

The original inference-only prediction files were saved after applying confidence 0.25. They are therefore suitable for fixed-operating-point precision, recall, and F1 but not for full AP integration. We have relabeled this analysis accordingly. With the Original-trained seed-42 detector, F1 was 0.642 on Original images, 0.619 on Gamma, 0.628 on CLAHE, 0.586 on Zero-DCE, and 0.566 on Retinexformer. When the detector was trained on the corresponding domain, fixed-threshold F1 recovered to approximately 0.621 for Gamma, 0.643 for CLAHE, 0.648 for Zero-DCE, and 0.618 for Retinexformer in seed 42. The largest recoveries therefore occur for the learned enhancers, consistent with a substantial train-test mismatch component.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

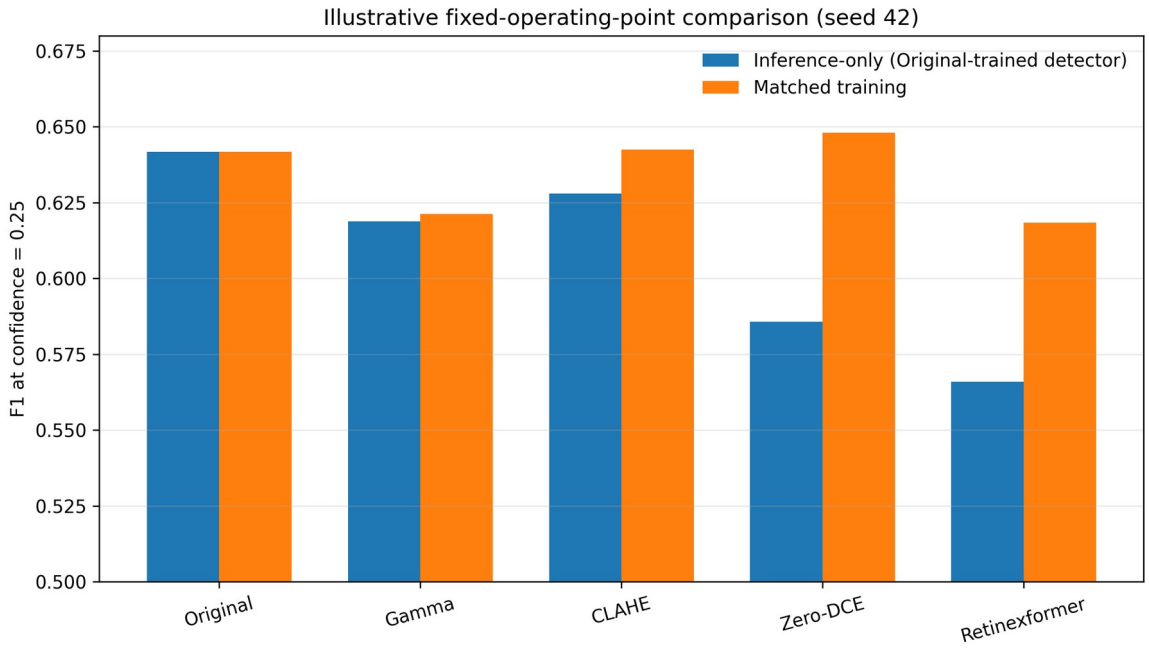


Figure 3. Illustrative seed-42 comparison at the fixed confidence threshold 0.25. This figure is not used for AP inference; it isolates the operating-point effect of inference-only preprocessing versus domain-matched training.

4.4 Operating-point sensitivity

Validation-selected thresholds were 0.35 or 0.40 for most trained models, rather than the common deployment threshold 0.25. Figure 4 shows the mean validation F1 curves across seeds. The selected thresholds were frozen before application to test. This analysis confirms that precision and recall at a fixed confidence can move because of score calibration even when rank-based AP changes less. We therefore keep fixed-threshold metrics and AP conceptually separate in the reported analysis.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

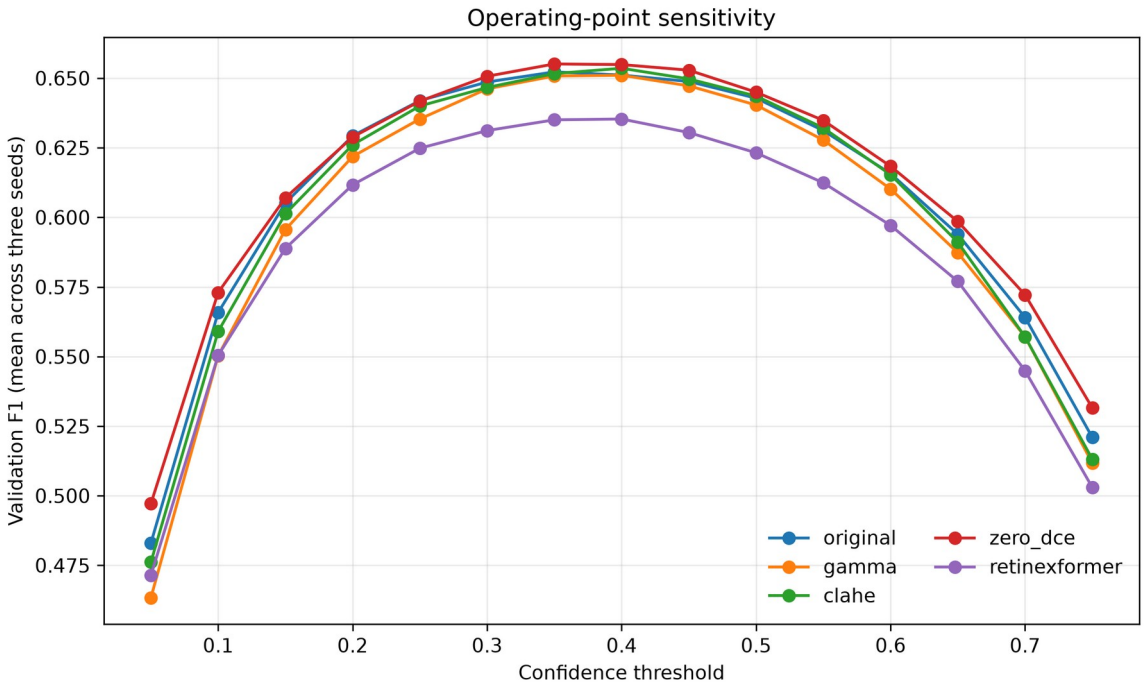


Figure 4. Validation F1 as a function of confidence threshold, averaged across the three matched-training seeds. Thresholds used for the validation-selected operating-point analysis were selected per trained model and then applied unchanged to test.

4.5 Class-wise and subgroup behavior

No enhancement method dominates the 12 classes. In three-seed mean AP@0.50, the largest positive class-specific changes relative to Original are modest: CLAHE is +0.0070 for Cat and +0.0068 for Dog; Zero-DCE is +0.0043 for Table and +0.0034 for Chair; Gamma is +0.0038 for Boat and +0.0022 for Car. Retinexformer is lower than Original in all 12 classes. Full AP@0.50 and AP@0.50:0.95 values, class supports, and run-to-run SDs are given in Supplementary Table S1.

Illumination subgroups also show localized gains, but the supports matter. Zero-DCE is +0.083 mAP@0.50 over Original for the Shadow subgroup, yet Shadow contains only 14 images and 55 instances. Zero-DCE is +0.039 under Twilight (174 images, 372 instances), while CLAHE is +0.031 under Twilight. These subgroup effects are descriptive because only three training runs are available and some conditions are small. Figure 5 shows the mean differences; Supplementary Table S2 reports the image/instance supports and SD across seeds.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

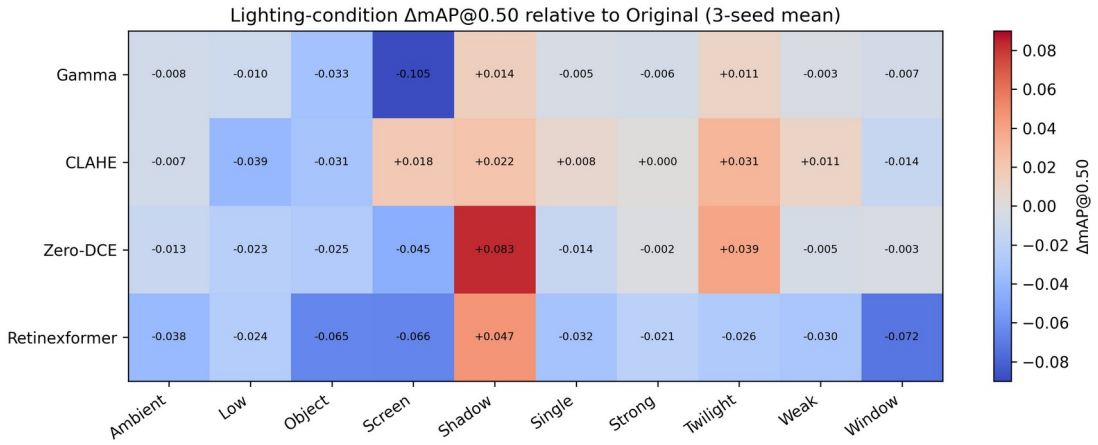


Figure 5. Three-seed mean change in mAP@0.50 relative to Original for each ExDark illumination condition. Small subgroups such as Shadow (14 images) and Screen (28 images) are interpreted cautiously; full supports and SDs are provided in Supplementary Table S2.

Object-size recall follows a similarly mixed pattern. Original recall is 0.1877±0.0228 for small, 0.5441±0.0115 for medium, and 0.7505±0.0074 for large objects. Zero-DCE is slightly higher for small objects (0.1977±0.0041) but lower for medium and large objects. The test-set size composition is 570 small, 2,529 medium, and 4,018 large instances. Supplementary Table S3 gives all method-seed values.

4.6 Validation-only sensitivity of classical enhancement settings

The post-hoc sensitivity analysis is intentionally restricted to the validation split. For Gamma, mAP@0.50 increases from 0.5529 at γ=0.4 to 0.6287 at γ=0.8; the prespecified γ=0.6 yields 0.6018. For CLAHE, clip limit 1.5 yields 0.6183, compared with 0.6124 for the prespecified clip limit 2.0, and performance decreases as the clip limit rises further. Because these checks were run after the main experiment, we do not substitute the better validation settings into the reported test results. Instead, the sensitivity results demonstrate that enhancement conclusions depend on the operating configuration and should not be generalized to the algorithm family as a whole.

Table 4. Post-hoc validation-only sensitivity analysis. Main-test settings are identified but were not retuned after this analysis.

Method	Parameter	Role	Val mAP@0.50	Val mAP@0.50:0.95	Val F1@0.25
Gamma	0.4	Sensitivity	0.5529	0.3296	0.5846
Gamma	0.5	Sensitivity	0.5814	0.3485	0.6033
Gamma	0.6	Main	0.6018	0.3647	0.6130
Gamma	0.7	Sensitivity	0.6203	0.3779	0.6229
Gamma	0.8	Sensitivity	0.6287	0.3830	0.6301
CLAHE	1.5	Sensitivity	0.6183	0.3781	0.6197
CLAHE	2.0	Main	0.6124	0.3717	0.6171
CLAHE	2.5	Sensitivity	0.6065	0.3685	0.6158
CLAHE	3.0	Sensitivity	0.5997	0.3642	0.6139

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

4.7 Perceptual quality versus detection utility

Retinexformer produced the lowest (best) mean NIQE (3.366) and BRISQUE (27.819), yet it also produced the lowest mean matched-training mAP@0.50 (0.5862). Across images within each method, better perceptual quality is weakly associated with better image-level detection utility, but the associations are small. For example, the NIQE–F1 Spearman ρ is −0.148 for Original and −0.101 for Retinexformer; BRISQUE–F1 is −0.102 and −0.055, respectively. Although the 2,000-resample image-bootstrap intervals exclude zero for these coefficients, the associations remain weak in magnitude. The five method-level quality-versus-mAP points are shown only descriptively.

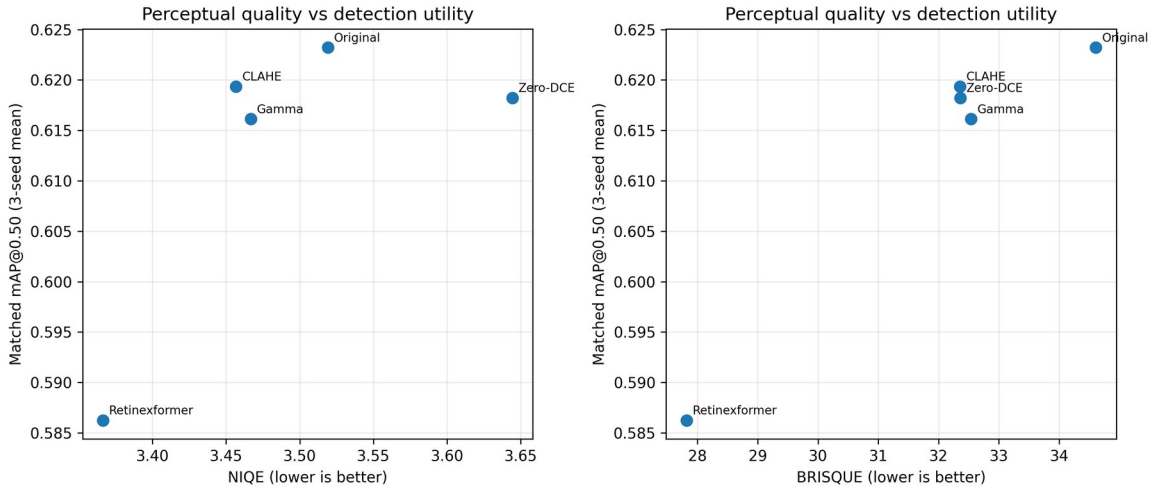


Figure 6. Perceptual image quality and matched detection performance. Lower NIQE/BRISQUE indicates better predicted perceptual quality; error bars on mAP are SD across the three training seeds. The five method-level points are descriptive, not a population-level correlation test.

4.8 Adaptive selection and retrospective upper bound

The Random Forest selector chose Original for 1511 of 2,563 test images, Gamma for 553, CLAHE for 198, Zero-DCE for 154, and Retinexformer for 147. At confidence 0.25 it achieved precision 0.676, recall 0.586, and F1 0.628, below the Original fixed-threshold baseline. The retrospective oracle, which uses test ground truth and is not deployable, reaches F1 0.719. The gap between the learned selector and oracle indicates that image-wise complementarity exists, but the hand-crafted feature vector does not predict it reliably enough.

4.9 Representative detections

Figure 7 illustrates three matched-domain examples at seed 42 and confidence 0.25. Green ground-truth boxes are detected, orange ground-truth boxes are missed, and red boxes are false positives. The examples show that enhancement can recover a missed object in one image and suppress or introduce detections in another. They are qualitative illustrations only; we do not infer a causal mechanism from edge, texture, or confidence changes without a dedicated feature-level analysis.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

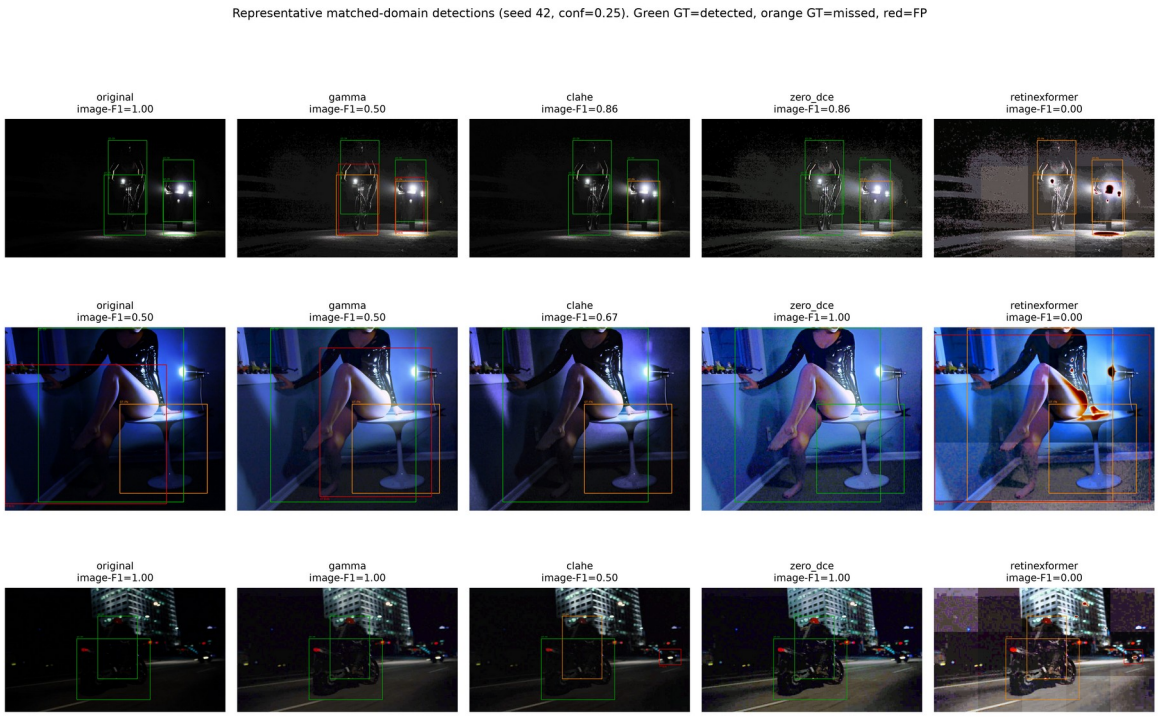


Figure 7. Representative matched-domain detections (seed 42, confidence 0.25). Green: detected ground truth; orange: missed ground truth; red: false positive. Examples illustrate image-wise variation and are not used as quantitative evidence of a causal mechanism.

5. Discussion

The results support a more specific interpretation. When AP is computed from low-confidence prediction streams and matched training is treated as the primary comparison, Original, Gamma, CLAHE, and Zero-DCE occupy a narrow performance band. Their ordering changes across seeds. This variation matters: a single seed can make Zero-DCE appear slightly better or Gamma appear slightly better, but neither advantage persists across all runs. Retinexformer is the one clear exception in this configuration because it remains lower across every seed and every class despite strong perceptual-quality scores.

The fixed-threshold inference-only experiment still has practical value, but its role is narrower than the matched-training analysis. It shows what happens if an enhancement stage is inserted ahead of a detector without retraining: F1 decreases for all four transformations, especially the learned methods. Matched training recovers much of this loss. This is consistent with distribution mismatch, but the study does not prove that mismatch is the sole cause. Storage encoding, enhancement artifacts, score calibration, and architecture-specific feature responses may also contribute.

The subgroup results do not support a claim that one enhancer is broadly superior under a particular form of darkness. Some conditions show positive mean differences, but several have limited support and the direction can vary by seed. Likewise, Zero-DCE shows a modest small-object recall advantage while giving lower medium- and large-object recall. These findings are best viewed as hypotheses for targeted follow-up rather than as deployment rules.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

The perceptual-quality analysis reinforces the separation between restoration and detection objectives. Retinexformer improves NIQE and BRISQUE more than the other domains but does not improve detector AP. This observation is compatible with recent task-aware work that explicitly optimizes semantic alignment or domain adaptation [12], [13], [18]-[21]. It does not imply that Retinexformer is ineffective as a restoration method; it means that this pretrained LOL-v2-real configuration is not an automatic substitute for task-aware adaptation on ExDark with YOLO11n.

The selector analysis points in the same direction from a different angle. There is enough image-wise complementarity for a ground-truth oracle to improve F1 substantially, yet a Random Forest based on simple luminance, texture, saturation, and Original-detector confidence features does not capture that complementarity well enough. More capable task-aware selectors may succeed, but they would constitute a different method and should be validated independently.

5.1 Limitations

The scope of the evidence is intentionally narrow. All detector experiments use YOLO11n, one official ExDark split, and the stated enhancement configurations. The results therefore do not establish a general property of all detectors, low-light datasets, or enhancement methods. Architecturally distinct detectors and additional low-light benchmarks would be valuable replications.

Three training seeds expose meaningful run-to-run variation but are too few for strong inference about the distribution of retraining outcomes. We therefore report seed SDs descriptively and use bootstrap intervals only for conditional image-sampling uncertainty within fixed trained models.

Gamma and CLAHE were not exhaustively tuned, and only one pretrained checkpoint was evaluated for each learned enhancer. The post-hoc validation sensitivity analysis shows that classical settings matter. Accordingly, the conclusions apply to the evaluated configurations rather than to the algorithm families as a whole.

Enhanced images preserve dimensions and annotations but are stored using the original file extension. JPG/JPEG inputs are therefore re-encoded after enhancement, while PNG outputs remain lossless. This implementation detail is shared by the enhanced domains but not by Original and may contribute small codec-related differences.

Finally, NIQE and BRISQUE are no-reference models rather than direct human judgments, and the representative detections are qualitative. Mechanistic explanations involving edges, texture, noise, or confidence calibration are presented as plausible hypotheses, not established causes.

6. Conclusion

On the evaluated ExDark split with YOLO11n, inserting a generic low-light enhancer ahead of an Original-trained detector lowers fixed-threshold F1 for all four tested transformations. Retraining the detector within the transformed domain recovers much of that loss. Once matched training is repeated across three seeds and AP is recomputed from low-confidence prediction streams, Original, Gamma, CLAHE, and Zero-DCE remain close and their ranking changes between runs; none provides a stable overall AP advantage. Retinexformer remains lower in this configuration even though it achieves the best NIQE and BRISQUE scores. Within this experimental scope, the practical implication is straightforward: perceptual restoration should not be assumed to improve object detection. Any

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

claimed benefit should be evaluated alongside detector adaptation, training-run variability, and the exact enhancement operating point.

Declarations

Funding

This research received no external project funding.

Conflict of Interest

The authors declare that they have no competing interests.

Ethics Statement

No new study involving human participants or animals was conducted. The work is a secondary computational analysis of the publicly released ExDark dataset and its existing annotations.

Data and Code Availability

The ExDark dataset is publicly available from the official dataset repository. Source code, configuration files, split lists, and revision-analysis scripts are available in a public repository. To preserve double-blind review, the repository URL is withheld from this blinded manuscript and has been supplied to the editorial office. Machine-readable tables and split lists are also supplied as supplementary material.

Author Contributions (CRediT)

Author 1: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, and writing—original draft. Author 2: Supervision, conceptualization, methodology, validation, and writing—review and editing. Both authors reviewed and approved the revised manuscript.

AI-Assisted Tools Disclosure

ChatGPT was used as an assistive tool for language editing, code troubleshooting, and organization of the experimental workflow. All experiments were run locally by the authors, and all numerical results, tables, and figures were generated from and checked against saved experimental outputs. Generative AI was not used to create or alter dataset images or to generate experimental data. The authors take responsibility for the final text, analyses, and interpretations.

Acknowledgment

The authors acknowledge the developers and maintainers of ExDark, Zero-DCE, Retinexformer, Ultralytics YOLO11, and PyIQA for making the resources used in this study publicly available.

References

1. Loh YP, Chan CS. Getting to know low-light images with the Exclusively Dark dataset. Comput Vis Image Underst 2019;*178*:30–42. doi: 10.1016/j.cviu.2018.10.010.
2. Liu J, Xu D, Yang W, Fan M, Huang H. Benchmarking low-light image enhancement and beyond. Int J Comput Vis 2021;*129*(4):1153–84. doi: 10.1007/s11263-020-01418-8.
3. Guo C, et al. Zero-reference deep curve estimation for low-light image enhancement. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR); 2020. p. 1780–9.
4. Cai Y, Bian H, Lin J, Wang H, Timofte R, Zhang Y. Retinexformer: One-stage Retinex-based transformer for low-light image enhancement. In: Proc IEEE/CVF Int Conf Comput Vis (ICCV); 2023. p. 12504–13.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.

5. Yan Q, et al. HVI: A new color space for low-light image enhancement. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR); 2025. p. 5678–87.
6. Al Sobbahi R, Tekli J. Comparing deep learning models for low-light natural scene image enhancement and their impact on object detection and classification: Overview, empirical evaluation, and challenges. Signal Process Image Commun 2022;*109*:116848. doi: 10.1016/j.image.2022.116848.
7. Morawski I, Chen YA, Lin YS, Hsu WH. NOD: Taking a closer look at detection under extreme low-light conditions with Night Object Detection dataset. In: Proc Br Mach Vis Conf (BMVC); 2021.
8. Morawski I, Chen YA, Lin YS, Dangi S, He K, Hsu WH. GenISP: Neural ISP for low-light machine cognition. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit Workshops (CVPRW); 2022. p. 630–9.
9. Kim D, Kim W. Improving object detection performance through selective low-light enhancement. In: Proc IEEE 13th Int Conf Consum Electron-Berlin (ICCE-Berlin); 2023. p. 29–32. doi: 10.1109/ICCE-Berlin58801.2023.10375676.
10. Bhatti MOF, Dailey MN. Improving low light object detection using image enhancement models. In: Data Science and Artificial Intelligence. Commun Comput Inf Sci 2023;*1942*:198–212. doi: 10.1007/978-981-99-7969-1_15.
11. Islam MT, Alam I, Woo SS, Anwar S, Lee IH, Muhammad K. LoLI-Street: Benchmarking low-light image enhancement and beyond. In: Computer Vision – ACCV 2024. Lect Notes Comput Sci 2025;*15476*:348–65. doi: 10.1007/978-981-96-0917-8_20.
12. Wang S, et al. From enhancement to understanding: Build a generalized bridge for low-light vision via semantically consistent unsupervised fine-tuning. In: Proc IEEE/CVF Int Conf Comput Vis (ICCV); 2025. p. 13804–14. doi: 10.1109/ICCV51701.2025.01281.
13. Cheng J, Xu B, Pan W. Generative low-light enhancement for computer vision: Revisiting the gap between perceptual quality and detection utility. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit Workshops (CVPRW); 2026. p. 4936–45.
14. Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 2012;*21*(12):4695–708. doi: 10.1109/TIP.2012.2214050.
15. Mittal A, Soundararajan R, Bovik AC. Making a “completely blind” image quality analyzer. IEEE Signal Process Lett 2013;*20*(3):209–12. doi: 10.1109/LSP.2012.2227726.
16. Jocher G, Qiu J. Ultralytics YOLO11 [software]. 2024.
17. Chen C, et al. IQA-PyTorch: PyTorch toolbox for image quality assessment [software].
18. Du Z, Shi M, Deng J. Boosting object detection with zero-shot day-night domain adaptation. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR); 2024. p. 12666–76.
19. Fu S, Zhao Q, Liu H, Tao Q, Liu D. Low-light object detection via adaptive enhancement and dynamic feature fusion. Alex Eng J 2025;*126*:60–9. doi: 10.1016/j.aej.2025.04.047.
20. Wang G, Guo Y, Hu T. Enhancing low-light object detection through domain adaptation and image enhancement. Vis Comput 2025;*41*(14):11947–58. doi: 10.1007/s00371-025-04136-9.
21. Hu J, Zhao Q, Li G. DLDA: Unified dual-level domain adaptation for low-light object detection. Proc AAAI Conf Artif Intell 2026;*40*(6):4851–9. doi: 10.1609/aaai.v40i6.42488.

© 2026 The Author(s). Published by Qassim University under CC BY-NC-SA 4.0.