Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

How Much Can Per-Image Classical Filter Selection Help Frozen Object Detectors in Low Light and Haze?

Melikşah Beşir

besir.meliksah@metu.edu.tr

Middle East Technical University

Erhan Ilhan Konukseven

Middle East Technical University

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11674291/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

How Much Can Per-Image Classical Filter Selection Help Frozen Object Detectors in Low Light and Haze?

Melikşah Beşir[1*] and E. İlhan Konukseven[2]

[1*]Robotics Graduate Program, Middle East Technical University, Ankara, 06800, Türkiye.
[2]Department of Mechanical Engineering, Middle East Technical University, Ankara, 06800, Türkiye.

*Corresponding author(s). E-mail(s): besir.meliksah@metu.edu.tr;
Contributing authors: konuk@metu.edu.tr;

Abstract

Simple image filters such as gamma correction and contrast-limited adaptive histogram equalization (CLAHE) are often placed in front of object detectors that work in low light or haze. Several studies suggest that choosing a different filter for each image could pay off, because an oracle that knows the best filter for every image clearly beats any single filter. We test this idea for detectors that cannot be retrained. We applied 16 classical filters one at a time to images from the low-light Exclusively Dark (ExDark) and hazy Real-world Task-driven Testing Set (RTTS) benchmarks and ran four detectors pretrained on the Common Objects in Context (COCO) dataset on every filtered image. The per-image oracle raised the mean average precision at an intersection-over-union of 0.5 (mAP@50) by up to 3.7 points, but three findings limit this gain. Most of it disappears under stricter localization. Most images have many equally good filters. And when we checked each oracle choice on a second, independently noise-perturbed copy of the same image, only about half of its gain survived. The surviving part still exceeded the gain of the best single filter, but a learned selector that picks a filter from eight global image statistics could not capture it and never beat the best single filter. For a frozen detector, one well-chosen filter was as good as the per-image selector we tested, and oracle results should be checked on repeated measurements before they are used as evidence for adaptive preprocessing.

Keywords: Object detection, Image enhancement, Low-light images, Dehazing, Filter selection, Reproducibility

1 Introduction

Object detectors that run outdoors, on vehicles, robots or surveillance cameras, regularly see images taken at night or in haze. A common remedy is to place an image-enhancement step in front of the detector. Classical operations such as gamma correction, contrast-limited adaptive histogram equalization (CLAHE) [1] and dark-channel-prior dehazing [2] are attractive for this role because they are fast, easy to interpret and need no training. Learned alternatives usually change the detector as well: image-adaptive YOLO (You Only Look Once) [3], gated differentiable image processing [4] and ERUP-YOLO [5] train the enhancement module together with the detector, and many recent low-light detectors are retrained end to end [6].

Many deployed systems cannot retrain their detector. A certified model may have to stay unchanged, or the detector may be supplied by a vendor as a black box. For such frozen detectors, a natural idea is to choose the most suitable filter separately for each image. Oracle analyses make this idea look promising. Awad et al. [7] kept, for each underwater image, the best of the original and nine enhanced versions and reported large gains over every uniform choice. AMIEOD (adaptive multi-experts image enhancement for object detection) [8] learns to select one of several enhancement experts for each image.

An oracle that picks the best filter after seeing the result is not, however, evidence that a usable policy will gain as much. There are three reasons. First, the best of many noisy outcomes is biased upward, a problem known as the optimizer's curse [9]; part of the oracle's advantage is luck. Second, when detection quality is measured by counting detected objects, several filters often produce the same count, so the "best filter" of an image is frequently not defined. Third, the oracle uses the image's ground truth, which a deployed system never sees.

We examine these three effects in a setting where the oracle can be computed exactly. Sixteen classical filters are applied one at a time to images from the low-light Exclusively Dark (ExDark) benchmark [10] and the hazy Real-world Task-driven Testing Set (RTTS) [11], and four frozen detectors pretrained on the Common Objects in Context (COCO) dataset are run on every filtered image. For every image and filter we count the correctly detected objects. We then compare four ways of choosing a filter: no filtering, the best single filter, a learned router that sees only image statistics, and the per-image oracle. To separate luck from real benefit, we also repeat the whole pipeline on two copies of every image with independent small noise, and check whether the oracle's choice on one copy still pays off on the other. Fig. 1 summarizes the design and the four questions we ask:

Q1. How much would per-image selection gain if the best filter were known?
Q2. Is there a single best filter for an image, or do many filters tie?
Q3. How much of the oracle's gain survives when its choice is checked on a second copy of the image?
Q4. Can a router that sees only image statistics beat the best single filter?

The answers temper the case for per-image selection. The oracle gains up to 3.7 points of mean average precision at an intersection-over-union (IoU) of 0.5 (mAP@50), but at most about one point at the stricter mAP@[.5:.95], which averages over IoU thresholds from 0.50 to 0.95, and most images have many equally good filters. Only about half of the oracle's gain survives the second copy. This surviving part is real and still exceeds what the best single filter achieves, but the learned router we tested could not capture it and never beat the best single filter. Our contributions are:

- a simple check for oracle and "mixed set" analyses: score each oracle choice on an independently perturbed copy of the input. The original oracle overestimates the reproducible gain and the checked oracle underestimates it, so together they bound it;
- evidence from four frozen detectors, including the detection transformer RF-DETR, and two degradation types: the oracle's advantage is concentrated at loose localization, dominated by ties, and only partly reproducible;
- practical guidance: for a frozen detector, mild gamma brightening for low light and CLAHE for haze matched the tested per-image router in all eight detector–dataset settings.

2 Related work

Enhancement for detection. Joint training adapts image processing to a detector: IA-YOLO predicts the parameters of a differentiable filter pipeline for each image [3], GDIP blends several differentiable operations through learned gates [4], and ERUP-YOLO combines pixel-wise curve adjustment with local filtering [5]. Restoration can also be optimized for detection directly [12, 13], or the detector can learn features that are less sensitive to illumination [14, 15]. Enhancers designed for visual quality, such as RetinexNet [16], Zero-DCE [17], EnlightenGAN [18] and self-calibrated illumination (SCI) [19] for low light, or AOD-Net [20], FFA-Net [21] and GridDehazeNet [22] for haze, produce outputs outside any fixed bank of

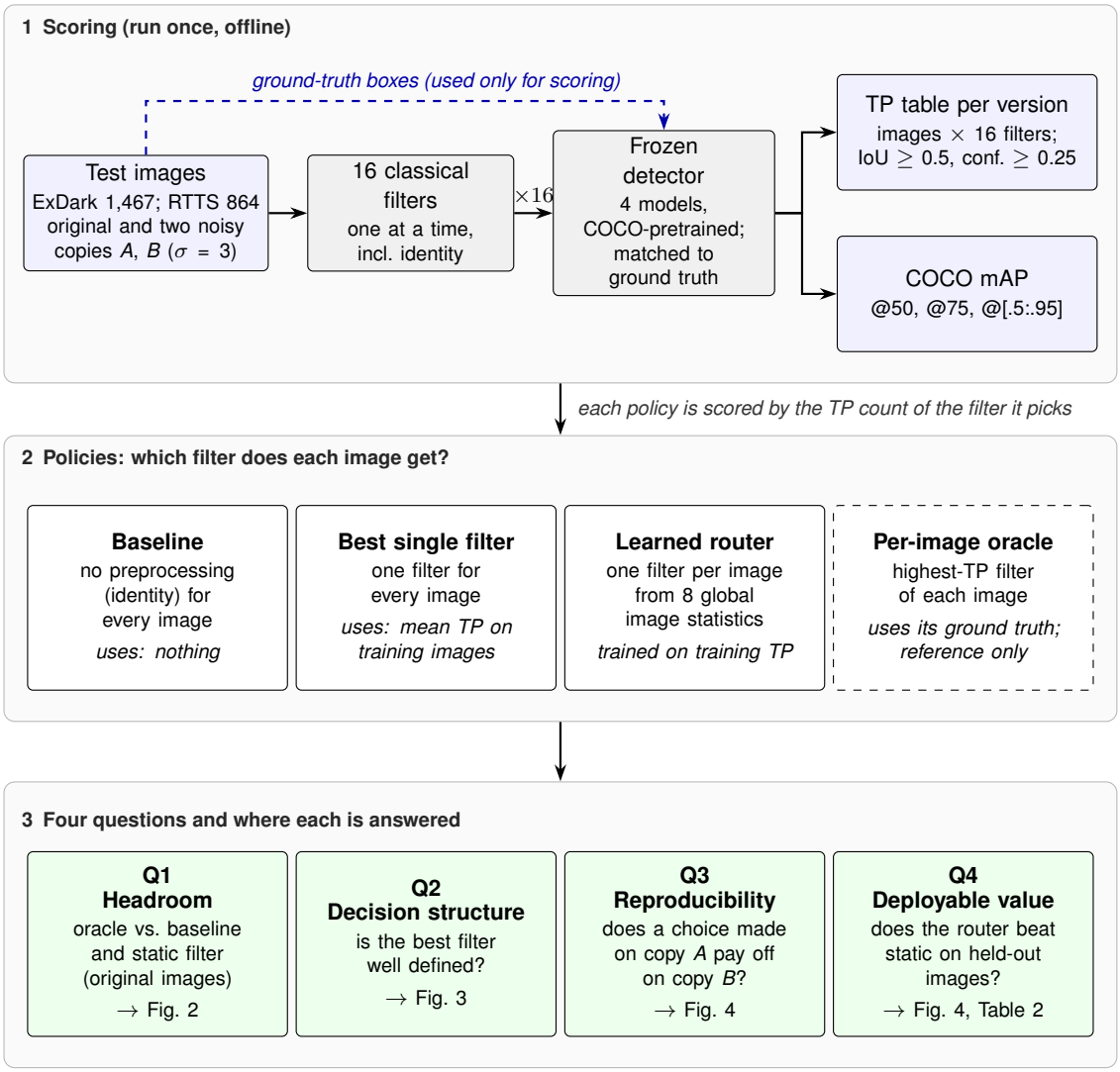

Fig. 1 Study design. (1) Every evaluation image is used as is and as two copies with small independent Gaussian noise (A and B), which imitate repeated captures of the same scene. Each version passes through every filter separately and then through a frozen detector, and the detections are matched to the ground truth. (2) Four policies choose one filter per image and are scored by the number of true positives (TP) of the chosen filter. Only the oracle sees the ground truth, so it is a reference rather than a usable method. (3) The four questions and where each is answered

classical operations. We do not study these methods; we isolate the choice among fixed classical filters for a frozen detector.

Does enhancement help recognition? Dehazing does not reliably improve image classification [23], and a large benchmark of enhancement methods in poor visibility found limited and inconsistent benefits for detection [24]. Low-light enhancement can help or harm both classification and detection [25]. Demir et al. [26] found that the visual quality of dehazed video and the accuracy of an object tracker on it do not move together. LIME-Eval [27] argues that detectors retrained on enhanced images overfit and evaluates enhancers with detectors pretrained on normal-light data, the setting we adopt. Closest to our work, Awad et al. [7] built a per-image "mixed set" of the best enhanced version of each underwater image, and AMIEOD [8] learns to select an enhancement expert per image. In AMIEOD's own ablation on ExDark, where detector and experts are trained

jointly, learned selection reaches 82.1 mAP@50 against 82.0 for the best single expert.

Per-instance selection and oracle bias. Choosing one of several fixed procedures for each input is the algorithm selection problem [28]. Its literature compares a learned selector with the best single procedure and with the per-instance oracle [29], and we use the same three reference points. Instance-aware test-time augmentation selects image transformations for a fixed network [30], and empirical upper bounds have been used to estimate the remaining headroom of detectors [31]. Because an oracle takes the best of noisy outcomes, its value is biased upward [9]. Scoring the choice on an independent sample, the "double estimator" of van Hasselt [32], removes this bias at the cost of a small downward bias. Our re-measurement test applies this idea to per-image filter selection. It is also motivated by the observation that small changes to an input can change a network's output [33]. Robustness can instead be built into the detector: common corruptions cause large accuracy drops for classifiers [34] and detectors [35], and augmentation schemes such as AugMix [36] expose models to many transformations during training.

3 Experimental setup

3.1 Filters

The filter bank contains 16 standard operations (Table 1). Each filter is applied on its own: filters are never chained or blended, and each image receives exactly one filter. The identity filter leaves the image unchanged and defines the baseline. Gamma correction maps normalized intensity $I \in [0, 1]$ to I^{γ}, so $\gamma < 1$ brightens and $\gamma > 1$ darkens. CLAHE equalizes local histograms on an 8×8 tile grid with a clip limit that bounds contrast amplification. Unsharp masking adds a scaled high-pass component, and dark-channel-prior dehazing removes haze with a strength ω, where $\omega = 1$ would remove all estimated haze. When an analysis needs one label for an image with several equally good filters, ties are broken by the order of Table 1.

Table 1 The filter bank. Each filter is applied on its own to the input image

Operation	Parameters	Filters
Identity (no filtering)	–	1
Gamma, brightening	$\gamma \in \{0.4, 0.6, 0.8\}$	3
Gamma, darkening	$\gamma \in \{1.5, 2.0, 2.5\}$	3
CLAHE, 8×8 tiles	clip $\in \{2, 4, 8\}$	3
Histogram equalization	–	1
Gray-world white balance	–	1
Unsharp masking	strength $\in \{0.5, 1\}$	2
Dark-channel dehazing	$\omega \in \{0.7, 0.9\}$	2
Total		16

3.2 Detectors and data

We use four detectors pretrained on COCO, all with frozen weights: YOLOv3u, YOLOv5lu and YOLOv8l from the Ultralytics family [37–39], and RF-DETR [40], a real-time detection transformer [41]. YOLOv3u and YOLOv5lu are modern anchor-free reimplementations, so we treat the three YOLO models as different convolutional detectors rather than as successive versions. Detector classes are mapped to each dataset's label set.

ExDark [10] contains real low-light photographs and RTTS [11] real hazy photographs. We use fixed evaluation subsets of 1,467 ExDark images (4,646 annotated objects, 3.17 per image) and 864 RTTS images (8,768 objects, 10.15 per image), about one fifth of each dataset, drawn once with random seed 42 and shared by all analyses. Because the detectors are frozen, these images are used only for evaluation; the router is trained and tested within them by cross-validation.

3.3 Scoring

For every image and filter we count the true positives (TP): detections with a confidence of at least 0.25 that match a ground-truth box of the same class one-to-one at an IoU of at least 0.5. The TP count of an image equals its number of objects times its recall at this operating point, so choosing the filter with the most TP maximizes per-image recall; false positives do not enter the count. A difference of Δ TP per image between two policies equals a change of $\Delta/\bar{n}$ in dataset recall, where $\bar{n}$ is the mean number of objects per image (3.17 for ExDark, 10.15 for RTTS). We use this to express

TP differences as recall points. Corpus detection quality is computed with pycocotools at an evaluation confidence of 0.001 as mAP@50, mAP@75 (IoU of 0.75) and mAP@[.5:.95]. For the Ultralytics detectors, non-maximum suppression uses an IoU threshold of 0.7.

3.4 Four policies

A policy assigns one filter to each image and is scored by the mean TP count of the chosen filters.

No filtering (baseline). The identity filter for every image.

Best single filter. One filter for all images, the one with the highest mean TP count. For Q1 and Q2 it is chosen on all evaluation images; for Q3 and Q4 it is chosen on the training folds only and evaluated on held-out images, exactly like the router.

Learned router. A deployable policy that chooses a filter from eight global statistics of the unfiltered image: mean luminance, root-mean-square contrast, dark-channel mean, colorfulness, mean saturation, the 5th and 95th luminance percentiles, and the logarithm of the Laplacian variance, a sharpness measure. One ridge-regression model per filter predicts the TP count from these statistics, and the router picks the filter with the largest prediction. We use five outer folds repeated with five different splits. Within every training fold the statistics are standardized and the ridge penalty is chosen from 10^{-4} to 10^{4} by three-fold inner cross-validation. Models are trained on one noisy copy of the training images and scored on the other copy of the held-out images, in both directions.

Per-image oracle. The filter with the highest TP count on the image itself. It uses the ground truth and is therefore a reference, not a usable method. Its TP value is an exact upper bound for any policy that assigns one filter per image. The mAP of the detections it selects is not an upper bound on mAP, because corpus mAP does not add up over images.

3.5 Re-measurement test

To separate a real benefit from luck, we create two copies of every image, A and B, by adding independent Gaussian pixel noise with a standard deviation of 3 on the 0–255 scale, and run the full pipeline on both. The oracle's choice on copy A is then scored on copy B, and vice versa. When several filters tie on the first copy, we average the second copy's TP count over all of them. As checks, we also score the first tied filter in the order of Table 1 and the best and worst tied filters on the second copy. We report the share of the oracle's gain over no filtering that survives this check.

In expectation, the two oracle values bracket the gain that a per-image choice could truly reproduce. The oracle evaluated on the same copy overestimates it, because it profits from the noise of that copy. The oracle checked on the other copy underestimates it, because a choice made on one noisy copy is not always the best choice in expectation. This follows the single- and double-estimator results of van Hasselt [32]; a formal statement and proof for our setting, which also covers the averaging over tied filters, are given in Online Resource 1.

3.6 Statistics

Results on the original images (Q1 and Q2) are exact summaries of the evaluation subsets. Comparisons on the noisy copies (Q3 and Q4) use a paired bootstrap over images with 10,000 resamples and report 95% confidence intervals (CIs). Router predictions are out of fold, and the results of the five repeated splits are averaged per image before resampling. The intervals describe uncertainty from the sample of images given the fitted policies: they do not refit the models in each resample, they condition on one pair of noisy copies, and they are not adjusted for the eight detector–dataset comparisons.

3.7 Use of a language model

A large language model (Claude, Anthropic) was used to editing the manuscript and write the code that draws the figures from the result files. The study design, experiments and analyses are the authors' own. The authors reviewed and edited all text and take full responsibility for the content.

4 Results

4.1 Q1: How large is the oracle's gain?

On the original images, choosing the best filter for every image raises mAP@50 by 2.07 to 3.73

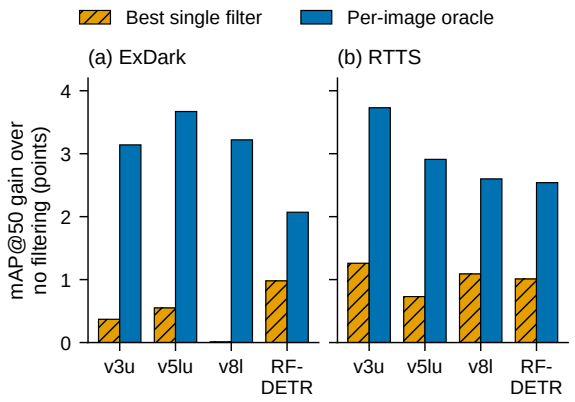

Fig. 2 Gain in mAP@50 over no filtering on the original images for (a) ExDark and (b) RTTS. Hatched bars: best single filter (one filter for all images), which was $\gamma = 0.8$ for ExDark and CLAHE with clip limit 2 or 4 for RTTS. Plain bars: per-image oracle. Exact values are listed in Online Resource 1, Table S1

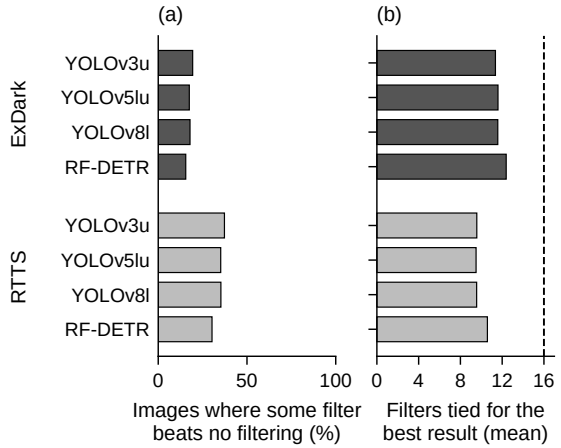

Fig. 3 How often the choice of filter matters. (a) Share of original images on which at least one filter finds more objects than no filtering. (b) Mean number of filters that tie for the highest TP count on an image (noisy copies, mean of copies A and B); the dashed line marks all 16 filters. Darker bars: ExDark; lighter bars: RTTS

points over no filtering, whereas the best single filter changes it by only 0.01 to 1.26 points (Fig. 2). In TP terms, the oracle finds 0.19 to 0.21 additional objects per image on ExDark and 0.44 to 0.60 on RTTS, or 6.1 to 6.8 and 4.3 to 6.0 recall points.

Most of this gain is at loose localization. Because the oracle counts matches at an IoU of 0.5 and ignores false positives, its selections change mAP@75 by −0.40 to +1.56 points and mAP@[.5:.95] by only +0.09 to +1.05 points. Online Resource 1 lists the exact values (Table S1) and shows the three ExDark images with the largest oracle gain (Fig. S1); these were chosen for illustration and are not typical.

4.2 Q2: Is there a single best filter?

For most images there is not. On the original images, no filtering already reaches the oracle's TP count on 62.7 to 84.5% of images and comes within one true positive of it on 86.2 to 98.8%. Some filter strictly beats no filtering on only 15.5 to 37.3% of images (Fig. 3a). On the noisy copies, 84.5 to 96.0% of images have at least two best filters, and on average 9.5 to 12.4 of the 16 filters tie for the best result (Fig. 3b). A single "best filter" label is therefore usually an arbitrary pick among equally good options, which makes it a poor training target for a router.

Relative to the number of objects in each image, no filtering misses on average 4.1 to 7.6% of an image's objects compared with the oracle, and the best single filter 3.3 to 6.4%. The single filter helps mainly on images with at least two objects. On RTTS images with a single object it ties no filtering for RF-DETR and increases the shortfall for the three YOLO detectors (Online Resource 1, Tables S2 and S3).

4.3 Q3: Does the oracle's choice hold on a second copy?

On the noisy copies, the oracle evaluated on the same copy adds 0.21 to 0.70 true positives per image over no filtering (Fig. 4). This is slightly more than on the original images, which is consistent with the added noise giving the oracle more variation to exploit. When each choice is checked on the other copy, the gain falls to 0.11 to 0.39, so 46.6 to 65.1% of it survives. The checked oracle still beats the best single filter, chosen out of fold, by 0.064 to 0.185 true positives per image, with every 95% CI above zero (Table 2). This is 2.0 to 2.9 recall points on ExDark and 1.5 to 1.8 on RTTS. Only 33 to 55% of the oracle's original advantage over the single filter survives the check.

So a reproducible per-image headroom exists: up to sampling error, its gain over no filtering lies between the two oracle bars of Fig. 4, and because the checked oracle beats the single filter, that headroom exceeds what one filter provides.

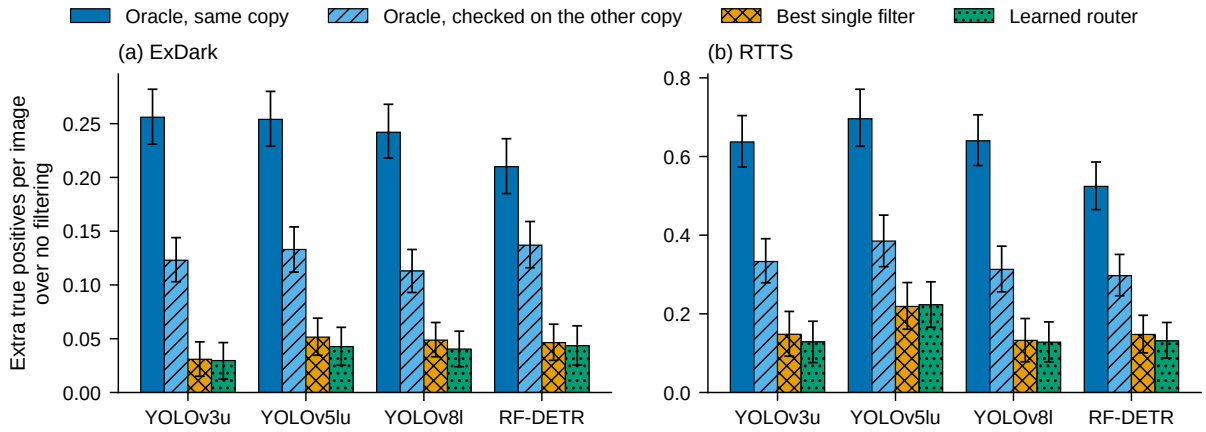

Fig. 4 Gains over no filtering on the noisy copies of (a) ExDark and (b) RTTS, in true positives per image. Within each group, the bars show, from left to right: the oracle scored on the copy where it chose (plain), the same oracle choice checked on the other copy (diagonal hatching), the best single filter (cross-hatching) and the learned router (dots). Error bars are 95% bootstrap CIs. The gain that a per-image choice could truly reproduce lies between the first two bars. The best single filter and the router are chosen on training folds and scored on held-out images

It is, however, defined by outcomes that require the ground truth of the image. How ties are handled matters for its size. Scoring the first tied filter instead of the average changes the result by at most 0.002 true positives per image on ExDark, but lowers it by 0.037 to 0.047 on RTTS. The best and worst tied filters fall on opposite sides of the single filter in every setting (Online Resource 1, Tables S4 and S5). Filters that tie on one copy can behave quite differently on the other.

4.4 Q4: Can a learned router use this headroom?

The router improves on no filtering in every setting, by 0.030 to 0.223 true positives per image, but not on the best single filter (Fig. 4, Table 2). The router-minus-single-filter differences range from −0.0193 to +0.0043 true positives per image, or −0.28 to +0.04 recall points. Seven of the eight estimates are negative and no interval lies above zero; the only interval that excludes zero, for YOLOv3u on RTTS, lies below it. The upper ends of all intervals are at most 0.025 true positives per image, well below the reproducible headroom of 0.064 to 0.185 in the same settings. Replacing the logarithm of the Laplacian variance with the raw variance changes the differences by at most 0.0065 and does not change the conclusion (Online Resource 1, Tables S6 and S7).

5 Discussion

Why the oracle overstates the gain. Three mechanisms separate the oracle from what a usable policy achieves, and each corresponds to one of our questions. Ties come first: most images reach the same TP count under many filters, so the oracle's advantage comes from a minority of images, and even there its choice is often one of several equally good filters. Luck comes second: the oracle takes the best of 16 noisy outcomes, and only one third to one half of its advantage over the single filter survives a second measurement of the same image. The remaining headroom, finally, is defined by which borderline detections become true positives under which filter. A router would have to predict this from the image alone, and a linear model on eight global statistics could not.

These results do not contradict the large mixed-set gains reported for underwater images [7]; they show how such gains should be read. A mixed set built from the best result per image is an optimistic ceiling, and only its reproducible part is a realistic target. AMIEOD's ablation points the same way: learned selection beat the best single expert by only 0.1 mAP@50 (82.1 versus 82.0) [8]. The gains of adaptive front ends seem to come mainly from adapting the detector or the enhancement itself, not from choosing among fixed operations.

Table 2 Differences to the best single filter on the noisy copies, in true positives per image, with 95% bootstrap CIs. "Oracle, checked" is the oracle's choice scored on the other copy; the router and the single filter are chosen out of fold

Data	Detector	Oracle checked on the other copy − single filter	Router − single filter
ExDark	YOLOv3u	0.092 [0.075, 0.110]	−0.0013 [−0.0108, 0.0082]
	YOLOv5lu	0.082 [0.065, 0.099]	−0.0088 [−0.0194, 0.0018]
	YOLOv8l	0.064 [0.048, 0.080]	−0.0083 [−0.0170, 0.0003]
	RF-DETR	0.090 [0.074, 0.108]	−0.0029 [−0.0124, 0.0065]
RTTS	YOLOv3u	0.185 [0.145, 0.228]	−0.0193 [−0.0375, −0.0015]
	YOLOv5lu	0.166 [0.124, 0.211]	+0.0043 [−0.0164, 0.0248]
	YOLOv8l	0.181 [0.137, 0.228]	−0.0045 [−0.0225, 0.0132]
	RF-DETR	0.150 [0.114, 0.189]	−0.0163 [−0.0319, 0.0000]

Practical guidance. For a frozen detector, a single filter was as good as per-image routing among the policies we tested. Its benefit is modest but holds on held-out images: 0.031 to 0.052 true positives per image on ExDark (1.0 to 1.6 recall points) and 0.133 to 0.219 on RTTS (1.3 to 2.2 recall points). The best single filters were mild gamma brightening ($\gamma = 0.8$) for low light and CLAHE with a clip limit of 2 or 4 for haze. The single filter helps mainly on images with several objects and can hurt on sparse hazy images. A gate that only decides between no filtering and the single filter is therefore a more plausible next step than free choice among sixteen filters. Such a gate should be fixed before evaluation and tested on new data. Latency, false positives and localization quality, which our TP score does not capture, should also enter a deployment decision. More generally, a per-image oracle or mixed-set result should be reported together with a re-measurement check of the kind used here and with a usable comparator such as the best single filter.

Limitations. The oracle covers only the 16 operations of the bank; learned enhancers, continuous parameters, combinations of filters and spatially varying operations can produce images outside it. The TP score ignores false positives and localization beyond an IoU of 0.5, and a score that accounts for both, such as localization-recall-precision [42], could change how often filters tie. We tested one linear router on eight global statistics; nonlinear models and features taken from the detector's own outputs remain untested. Gaussian noise is a controlled re-measurement, not a full camera-noise model, and our intervals condition on one pair of noisy copies; more pairs and noise levels would show how stable the transfer is. The evaluation subsets cover about one fifth of each dataset and two degradation types, and joint training of enhancement and detection is excluded by design. Finally, modern detectors may gain little from the bank because training on large, strongly augmented datasets already makes them largely insensitive to its photometric changes; we did not test this explanation.

6 Conclusion

We asked whether choosing a classical filter separately for each image helps frozen object detectors in low light and haze. A per-image oracle gains 2.1 to 3.7 mAP@50 points over no filtering but at most about one point at mAP@[.5:.95]. The choice is dominated by ties, and most images gain nothing from any filter. Only about half of the oracle's gain survives when its choices are checked on a second copy of each image, and a learned router on global image statistics matches but does not beat the best single filter. Oracle gains therefore overstate what the deployable policies we tested achieved. Checking oracle choices on a repeated measurement costs little and, in expectation, bounds the reproducible gain from both sides; we recommend it for any oracle or mixed-set analysis of adaptive preprocessing. Whether richer features or a score that accounts for false positives can recover the reproducible part of the headroom remains open.

Supplementary information. Online Resource 1 (ESM_1.pdf) contains the formal statement and proof of the bracketing result,

the full result tables with confidence intervals (Tables S1–S7) and qualitative examples (Fig. S1). Online Resource 2 (ESM_2.zip) contains the per-image true-positive counts of all 16 filters for every detector, dataset and image version (original images and copies A and B).

Acknowledgements. The authors thank Hidroan Ankara Hidrolik Mak. San. Tic. Ltd. Şti. (Ankara, Türkiye) for providing access to the NVIDIA RTX A4000 workstation used for the experiments.

Statements and Declarations

Funding. No funding was received for conducting this study. Hidroan Ankara Hidrolik Mak. San. Tic. Ltd. Şti. provided in-kind access to computing resources (see Acknowledgements).

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Compliance with ethical standards. This article does not contain any studies with human participants or animals performed by any of the authors. Ethics approval: not applicable. Consent to participate and consent for publication: not applicable.

Data availability. This study used two public datasets: the Exclusively Dark (ExDark) dataset [10] (https://doi.org/10.1016/j.cviu.2018.10.010), available from its official repository at https://github.com/cs-chan/Exclusively-Dark-Image-Dataset, and the Real-world Task-driven Testing Set (RTTS) of the RESIDE benchmark [11] (https://doi.org/10.1109/TIP.2018.2867951), available at https://sites.google.com/view/reside-dehaze-datasets. The data generated in this study, the per-image true-positive counts of all 16 filters for every detector, dataset and image version (original images and two noise-perturbed copies), together with the indices of the evaluation images, are included in the Supplementary Information (Online Resource 2). They are sufficient to recompute the oracle, tie, regret and transfer results and the single-filter results on the original images. The image features and cross-validation fold assignments used for the router analysis are available from the corresponding author on reasonable request.

Code availability. The analysis code is available from the corresponding author on reasonable request.

Author contributions. Melikşah Beşir: conceptualization, methodology, software, investigation, formal analysis, writing – original draft. E. İlhan Konukseven: supervision, writing – review and editing. Both authors read and approved the final manuscript.

References

[1] Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Heckbert, P.S. (ed.) Graphics Gems IV, pp. 474–485. Academic Press, San Diego (1994). https://doi.org/10.1016/B978-0-12-336156-1.50061-6

[2] He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(12), 2341–2353 (2011) https://doi.org/10.1109/TPAMI.2010.168

[3] Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., Zhang, L.: Image-adaptive YOLO for object detection in adverse weather conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 1792–1800 (2022). https://doi.org/10.1609/aaai.v36i2.20072

[4] Kalwar, S., Patel, D., Aanegola, A., Konda, K.R., Garg, S., Krishna, K.M.: GDIP: Gated differentiable image processing for object detection in adverse conditions. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 7083–7089 (2023). https://doi.org/10.1109/ICRA48891.2023.10160563

[5] Ogino, Y., Shoji, Y., Toizumi, T., Ito, A.: ERUP-YOLO: Enhancing object detection robustness for adverse weather condition by unified image-adaptive processing. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025). https://doi.org/10.1109/WACV61041.2025.00833

[6] Liu, Z., Tian, W., Zhang, A., Li, Q., Hu, M., Qi, W., Tuo, J.: Balancing noise suppression and detail recovery for efficient low-light object detection. Signal, Image and Video Processing **20**(4), 245 (2026) https://doi.org/10.1007/s11760-026-05322-5

[7] Awad, A., Saleem, A., Paheding, S., Lucas, E., Al-Ratrout, S., Havens, T.C.: Revisiting underwater image enhancement for object detection: A unified quality–detection evaluation framework. Journal of Imaging **12**(1), 18 (2026) https://doi.org/10.3390/jimaging12010018

[8] Huang, X., Chen, H., Zhang, W., Dai, X., Li, Y., Qing, L., He, X.: AMIEOD: Adaptive multi-experts image enhancement for object detection in low-illumination scenes. arXiv preprint arXiv:2605.06084 (2026) https://doi.org/10.48550/arXiv.2605.06084

[9] Smith, J.E., Winkler, R.L.: The optimizer's curse: Skepticism and postdecision surprise in decision analysis. Management Science **52**(3), 311–322 (2006) https://doi.org/10.1287/mnsc.1050.0451

[10] Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. Computer Vision and Image Understanding **178**, 30–42 (2019) https://doi.org/10.1016/j.cviu.2018.10.010

[11] Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2019) https://doi.org/10.1109/TIP.2018.2867951

[12] Sun, S., Ren, W., Wang, T., Cao, X.: Rethinking image restoration for object detection. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 4461–4474 (2022)

[13] Li, C., Zhou, H., Liu, Y., Yang, C., Xie, Y., Li, Z., Zhu, L.: Detection-friendly dehazing: Object detection in real-world hazy scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8284–8295 (2023) https://doi.org/10.1109/TPAMI.2023.3234976

[14] Cui, Z., Qi, G.-J., Gu, L., You, S., Zhang, Z., Harada, T.: Multitask AET with orthogonal tangent regularity for dark object detection. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2553–2562 (2021). https://doi.org/10.1109/ICCV48922.2021.00255

[15] Hong, M., Cheng, S., Huang, H., Fan, H., Liu, S.: You only look around: Learning illumination-invariant feature for low-light object detection. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 37, pp. 87136–87158 (2024)

[16] Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (BMVC) (2018)

[17] Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1780–1789 (2020). https://doi.org/10.1109/CVPR42600.2020.00185

[18] Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: EnlightenGAN: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing **30**, 2340–2349 (2021) https://doi.org/10.1109/TIP.2021.3051462

[19] Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5637–5646 (2022). https://doi.org/10.1109/CVPR52688.2022.00555

[20] Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: AOD-Net: All-in-one dehazing network. In: IEEE International Conference on Computer Vision (ICCV), pp. 4770–4778 (2017). https://doi.org/10.1109/ICCV.2017.511

[21] Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: FFA-Net: Feature fusion attention network for single image dehazing. In: AAAI Conference on Artificial Intelligence, vol. 34, pp. 11908–11915 (2020). https://doi.org/10.1609/aaai.v34i07.6865

[22] Liu, X., Ma, Y., Shi, Z., Chen, J.: GridDehazeNet: Attention-based multi-scale network for image dehazing. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7314–7323 (2019). https://doi.org/10.1109/ICCV.2019.00741

[23] Pei, Y., Huang, Y., Zou, Q., Lu, Y., Wang, S.: Does haze removal help CNN-based image classification? In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science. Springer, ??? (2018). https://doi.org/10.1007/978-3-030-01249-6_42

[24] Yang, W., Yuan, Y., Ren, W., Liu, J., Scheirer, W.J., Wang, Z., Zhang, T., Zhong, Q., Xie, D., Pu, S., *et al.*: Advancing image understanding in poor visibility environments: A collective benchmark study. IEEE Transactions on Image Processing **29**, 5737–5752 (2020) https://doi.org/10.1109/TIP.2020.2981922

[25] Wu, X., Lai, Z., Zhou, J., Gao, C., Hou, X., Zhang, Y.-n., Shen, L.: Low-light enhancement effect on classification and detection: An empirical study. arXiv preprint arXiv:2409.14461 (2024) https://doi.org/10.48550/arXiv.2409.14461

[26] Demir, H.S., Rajbharti, N., Sciarappo, S., Blain, J., Ozev, S.: Evaluating the impact of dehazing algorithms on object tracking performance. Signal, Image and Video Processing **19**(17), 1401 (2025) https://doi.org/10.1007/s11760-025-04996-7

[27] Li, M., Zhao, H., Guo, X.: LIME-Eval: Rethinking low-light image enhancement evaluation via object detection. arXiv preprint arXiv:2410.08810 (2024) https://doi.org/10.48550/arXiv.2410.08810

[28] Rice, J.R.: The algorithm selection problem. Advances in Computers **15**, 65–118 (1976) https://doi.org/10.1016/S0065-2458(08)60520-3

[29] Kerschke, P., Hoos, H.H., Neumann, F., Trautmann, H.: Automated algorithm selection: Survey and perspectives. Evolutionary Computation **27**(1), 3–45 (2019) https://doi.org/10.1162/evco_a_00242

[30] Kim, I., Kim, Y., Kim, S.: Learning loss for test-time augmentation. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33 (2020)

[31] Borji, A., Iranmanesh, S.M.: Empirical upper bound in object detection and more. arXiv preprint arXiv:1911.12451 (2020) https://doi.org/10.48550/arXiv.1911.12451

[32] van Hasselt, H.: Double Q-learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 23, pp. 2613–2621 (2010)

[33] Azulay, A., Weiss, Y.: Why do deep convolutional networks generalize so poorly to small image transformations? Journal of Machine Learning Research **20**(184), 1–25 (2019)

[34] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (ICLR) (2019)

[35] 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 preprint arXiv:1907.07484 (2019) https://doi.org/10.48550/arXiv.1907.07484

[36] 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: International Conference on Learning Representations (ICLR) (2020)

[37] Redmon, J., Farhadi, A.: YOLOv3: An

incremental improvement. arXiv preprint arXiv:1804.02767 (2018) https://doi.org/10.48550/arXiv.1804.02767

[38] Jocher, G.: YOLOv5 by Ultralytics. https://github.com/ultralytics/yolov5 (2020). https://doi.org/10.5281/zenodo.3908559

[39] Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO. https://github.com/ultralytics/ultralytics (2023)

[40] Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., Peri, N.: RF-DETR: Neural architecture search for real-time detection transformers. In: International Conference on Learning Representations (ICLR) (2026). arXiv:2511.09554. Code: https://github.com/roboflow/rf-detr

[41] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV), pp. 213–229 (2020). https://doi.org/10.1007/978-3-030-58452-8_13

[42] Oksuz, K., Cam, B.C., Akbas, E., Kalkan, S.: Localization recall precision (LRP): A new performance metric for object detection. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science. Springer, ??? (2018). https://doi.org/10.1007/978-3-030-01234-2_31

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2.zip