Questioning Post-hoc Image Corruptions as a Proxy for Environmental Variation in Autonomous Driving

Swarnendu Ghosh
swarnendu.ghosh@fri.uni-lj.si
Faculty of Computer and Information Science,
University of Ljubljana
Ljubljana, Slovenia

Branko Šter
branko.ster@fri.uni-lj.si
Faculty of Computer and Information Science,
University of Ljubljana
Ljubljana, Slovenia

Abstract

Camera-driven cars are usually stress-tested by degrading recorded footage with standard image filters—darkening, fog-like veils—and checking how badly the vision network copes. We show this misleads, and that whether it does depends on the physics of the condition imitated. We take two conditions our simulator produces honestly—a night, made by dimming the lights before rendering, and a fog, made by exponential thickening with scene depth—and match each to the filter named after it on measured picture change. Nine independently trained agents then drive both. Fog is reproduced faithfully at every strength. Night is not: at a picture change of 38 grey levels the physical night costs 0.51 of driving performance and the darkening filter matched to it costs nothing measurable, and all nine agents agree (sign test, $p = 0.004$). We test the natural explanation—that real weather removes detail *by distance* while a filter cannot—on 20 CARLA presets with metric depth. The signature is real and separates fog from a depth-blind filter, but nightfall scores as depth-blind itself, and it is nightfall that puts our agents off the road. The asymmetry reappears in training: post-hoc augmentation leaves an agent slightly *worse* against physical conditions (8.68 laps of ten against 9.05 unaugmented), while re-rendered conditions complete every lap.

Keywords

robustness evaluation, autonomous driving, reinforcement learning, image corruptions, simulation-to-reality gap, CARLA

1 Introduction

A camera-driven car that behaves well in clear weather may fail in fog, at night, or where the lane paint has worn away. Finding out *which* conditions are dangerous is expensive, and driving in genuinely dangerous ones is not an option, so a shortcut has become standard: manufacture the weather afterwards, running ImageNet-C style filters [3] over footage recorded on a clear day—a practice carried into driving benchmarks [5] and into robustness *training* [4], though building it properly needs scene depth [7]. (A second shortcut measures *recognition* instead of driving; such offline metrics track closed-loop quality poorly [1].) The shortcut hides an assumption: that a filter painted over a photograph stands in for the physical process it is named after. We test it where we can measure what matters—whether the car stays on the road.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.79

Our contributions are: (i) a like-for-like comparison of two physical conditions against the filters named after them, on nine agents, showing that the filter reproduces fog faithfully and fails completely for night (Section 3); (ii) a distance statistic, against depth-blind controls on 20 CARLA [2] weather and illumination presets, that tells dense fog from a depth-blind filter but not a dangerous change from a harmless one (Section 4); and (iii) a closed-circuit experiment showing the same asymmetry in *training* (Section 5).

2 Experimental Setup

Task and agents. A road-following task rendered with panda3d at 84×84 pixels, three frames stacked; a bicycle model at 60 Hz physics and 20 Hz control. The car scores for progress and loses for drifting off the lane centre; leaving the road ends the episode at once, so a low score means a crash rather than sloppy driving. Forty road layouts train, four unseen ones measure. Ten agents are trained by proximal policy optimisation (PPO) [8] in Stable-Baselines3 [6] for 500k steps. **All training is on clean images only**, so Sections 3 and 4 report the zero-shot robustness of a clear-weather agent—the setting these benchmarks normally probe. Nine converge to 243.3 ± 4.7 reward; the tenth loses its policy at 235k steps, having reached only 151, and never recovers (Figure 1). Damage is a ratio normalised by clean performance and is undefined for an agent that cannot drive, so that run is excluded—by a criterion fixed before the curves were seen—and the other nine are evaluated. Clean skill is reproducible: their mean clean returns span 252.1 to 260.3, a spread of 3.2%. Robustness is not. Under the physical night at full strength the same nine lose between 0.046 and 1.535 of their clean performance: a thirty-four-fold spread in fragility behind a three-percent spread in skill, so we report the direction all agree on, not just the mean.

Two conditions, and the filters named after them. The simulator can make a scene worse in two ways. An **in-engine** condition changes the world and re-renders it; a **post-hoc** condition edits the finished frame with one of 14 ImageNet-C style filters. Each comes at six strengths *s1*–*s6*, the index setting a fraction s of that condition's own range ($s = 0.26$, 0.60 and 1.00 at the *s2*, *s4* and *s6* of Figure 2); the index was never calibrated between an engine condition and a filter, which is why we match on measured picture change instead. We study the two in-engine conditions whose named filter can actually reach that change. *Night* scales the ambient and directional lights by the factor $1 - 0.88\,s$ *before* rendering, leaving 12% of daylight at $s = 1$, with no headlights; its filter, *darken*, multiplies every pixel of the finished frame by $1 - 0.80\,s$. *Fog* thickens exponentially with scene depth, at density $0.004 + 0.055\,s\,\mathrm{m}^{-1}$, so $0.059\,\mathrm{m}^{-1}$ at $s = 1$; its filter, *fog*, blends each pixel x toward a flat veil, $x(1-a) + 235a$ with $a = 0.65\,s$. One is a change of illumination and the other depends on distance, and the two filters differ accordingly—a multiply against a blend.

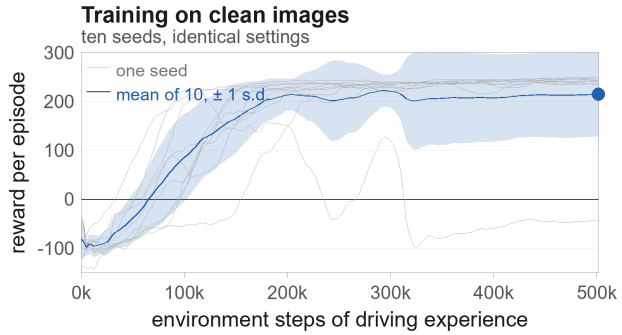


Figure 1: Reward per episode over the ten training seeds: the mean, a ±1 standard deviation envelope, and every run beneath. Nine converge to 243.3 ± 4.7; the tenth reaches 151, loses the policy at 235k steps and does not recover, which is what widens the envelope from 235k steps on.

Matching, and what we measure. We report the damage D a condition does to driving and the change L_1 it makes to the picture, both of Equation (1):

$$D = \frac{R_{\text{clean}} - R_{\text{cond}}}{|R_{\text{clean}}|}, \qquad L_1 = \frac{1}{N}\sum_i \left|x_i^{\text{cond}} - x_i^{\text{clean}}\right|, \quad (1)$$

where R_{cond} is the agent's mean episode return under that condition, R_{clean} its mean return on clean images, and x_i the ith of the N pixel values of a probe set of 2000 frames frozen *before* any driving result was inspected. $D = 0$ is no harm, $D = 0.1$ is driving 10% worse, $D > 1$ is crashing early; L_1 is on the 0–255 intensity scale. For each engine condition we keep the strength of its named filter whose L_1 is closest, so the two change the picture by the same measured amount and nothing else is held equal: at *s6* within 4% for night (39.0 against 37.4) and 1% for fog (42.4 against 42.9). Each agent drives 30 episodes per condition on identical seeds, so every comparison is paired agent by agent.

3 Same Change to the Picture, Different Outcome

Held to the same measured change in the picture, does a filter do the same damage as the physical condition it is named after? Write $\Delta = D_{\text{phys}} - D_{\text{filt}}$, positive when the physical condition is worse. For fog, yes. For night, no.

Fog first. At every strength the physical fog and its filter agree—the differences are −0.003, −0.006 and +0.093, and the nine agents split 3/9, 4/9 and 7/9 on which is worse, which is what disagreement about nothing looks like. A practitioner screening for fog with the fog filter would not be misled.

Night is the opposite. At *s6* both conditions shift every pixel by about 38 grey levels—a match to within 4%—and the physical night costs 0.51 of clean driving performance while the darkening filter costs nothing measurable (+0.001). Every one of the nine agents loses more to the night than to its filter (sign test, $p = 0.004$), and the gap widens with strength: +0.013, +0.082, +0.511. Across the nine the night costs at least 4.6% and the filter at most 1.7%. Screening with the filter, a practitioner would file the most dangerous condition tested as the safest.

The difference is not size: at *s6* fog changes the picture *more* than night (42.4 against 39.0 grey levels) and costs under a quarter as much. What separates them is what the change is made of (Section 4).

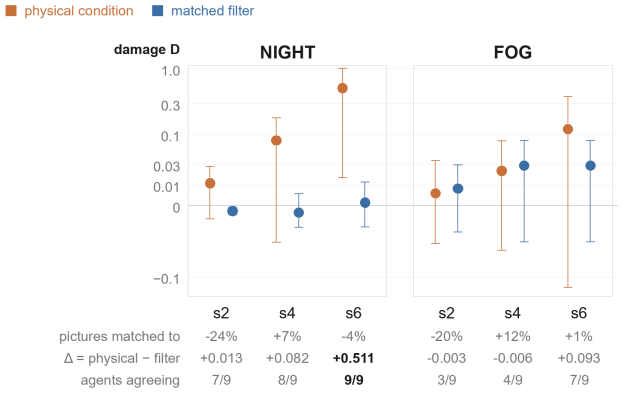


Figure 2: Mean damage D over the nine agents, on a symmetric-log axis: linear within ±0.01 of zero and logarithmic outside, so the near-zero and negative filter values stay visible. Whiskers are ±1 standard deviation *across agents*: agent-to-agent spread, not precision of the mean. Pictures matched to is the L_1 gap between the two conditions— the input held equal—negative when the filter changed the picture less. Agents agreeing counts how many of the nine found the physical condition worse, which the sign test uses: 9/9 gives $p = 0.004$ and 8/9 gives $p = 0.039$; no other column reaches $p < 0.05$.

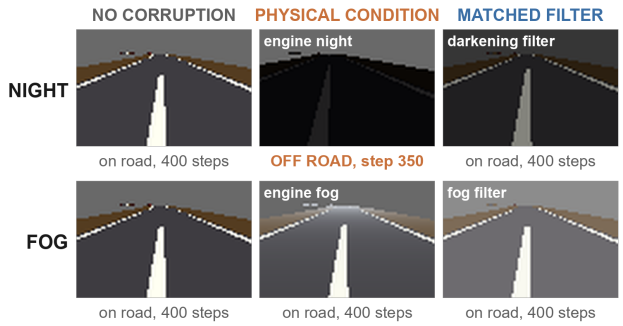


Figure 3: Both conditions on the same road, agent and random seed, driven by the most fragile of the nine agents. Each row shows the clean frame, the physical condition, and the filter matched to it on picture change. The engine night puts the car off the road while its filter does not; engine fog and its filter are alike, and both are harmless.

Figure 3 puts the two cases side by side, one episode each, driven by the most fragile of the nine—chosen because the failure is legible rather than typical. Under the engine night the car drifts 3.68 m off centre and leaves the road at step 350, while the darkening filter matched to it completes all 400 steps. Both fog conditions complete all 400 steps, and the engine fog and its filter are hard to tell apart.

4 A Distance Signature That Certifies Scattering, Not Danger

A filter sees a flat picture: what it does to a patch may depend on how that patch looks, but not on how far away it is. Physics is under no such constraint—fog thickens with the light path. That is the obvious explanation for Section 3; we test it on independent

data, where it survives as a statement about images and fails as an account of danger.

Defining τ. Cut each frame into 8×8 patches; patch p has clean local contrast c_p, clean brightness b_p, image row r_p, depth d_p and retained contrast c'_p. Within each cell of a 10×10 (c, b) quantile grid, remove all a distance-blind operator could have produced:

$$\varepsilon_p = c'_p - \widehat{c'}(c_p, b_p, r_p), \qquad \delta_p = \widetilde{d}_p - \widehat{d}(\widetilde{r}_p), \tag{2}$$

$$\tau \;=\; \mathrm{corr}(\varepsilon,\, \delta), \tag{3}$$

where $\widehat{\cdot}$ is a within-cell least-squares fit, $\widetilde{\cdot}$ a rank transform and corr Pearson's; despite the symbol, τ is the statistic of Equation (3) and not Kendall's rank correlation. So $\tau = 0$ means contrast was lost evenly at all distances, $\tau < 0$ that the far field lost more, $\tau > 0$ the near field. The row term is essential: height in the frame proxies distance (Spearman $\rho = -0.869$), so a control without it is defeated by an operator reading pixel coordinates alone—a veil ramping with image row scores $\tau = -0.382$ without it and $+0.301$ with, the opposite sign.

Calibrating τ on changes that cannot see depth. A bootstrap propagates sampling noise but not the bias of the measure itself, so we calibrate directly: 32 operators with no access to depth—pointwise gammas, flat veils and eight of the 14 filters at three strengths—scored against every clear reference give an envelope $|\tau| \le 0.228$, set by defocus blur at night. A condition counts only if it leaves that band.

Result on independent data. We use a CARLA [2] capture of 20 weather and illumination presets over 10 fixed poses—identical geometry throughout—with a true metric depth channel, three cameras pooling to 7 419 patches per condition. Each is compared against the clear capture at the closest matching illumination, so only the named factor differs. Grouping the other 19 by what physically changed: *scattering* (nine presets of fog and rain) averages $\tau = -0.341$, and the five dense fogs, from -0.467 to -0.530, are the only conditions in the capture that leave the envelope—the four thinner ones stay inside, so the signature tracks optical thickness, not the label; *surface* change (six presets of wet, flooded and darkened road) averages -0.166, all inside; *light alone* (night, dusk, low sun) averages -0.078, all inside, while changing the picture nearly three times as much (L_1 109 against 40).

That last group is the finding. At comparable picture change τ tells dense fog from the benchmark's fog filter (-0.504 against -0.040; only the engine fog leaves the envelope), but cannot tell nightfall from the darkening filter matched to it (-0.117 against $+0.022$, both inside, at L_1 122 and 131; Figure 4). Yet a loss of light is exactly what our agents cannot survive: the engine night of Sections 3 and 5 costs them more than any other condition and is the only one that puts them off the road. The signature is real and tells dense fog from a filter; what it does not do is tell a dangerous change from a harmless one.

Nor do the standard fidelity metrics. Against the same references and masked pixels, the peak signal-to-noise ratio (PSNR) merely restates the pixel change: at 16.3, 16.4 and 7.4 dB for scattering, surface and light, and lower meaning more degraded, it calls the changes of light much the worst and cannot separate a scattering medium from a wet road at all. Structural similarity (SSIM) is no more discriminating—higher meaning more similar, its per-family ranges of $[0.27, 0.88]$, $[0.52, 0.86]$ and $[0.29, 0.76]$ overlap almost completely. On the fog pair both rate the engine

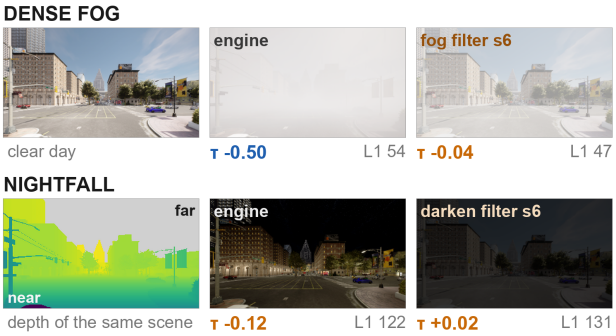


Figure 4: CARLA imagery with a registered metric depth channel, each engine condition beside the post-hoc operator that reproduces the same picture change. Dense fog erases the distance while the near kerb stays legible, and is the one condition here outside the envelope; nightfall changes the picture more than twice as much and sits inside it, beside the darkening filter. Blue marks τ outside the band, orange inside.

condition as the more degraded image, correctly (12.67 dB and SSIM 0.506 for the engine fog, 14.16 dB and 0.739 for its filter). On the nightfall pair both rate the *filter* as the more degraded image (5.43 dB and 0.249 for the filter, 6.19 dB and 0.385 for the engine night)—and it is the engine night, not the filter, that does all the driving damage.

The mechanism does not follow. It is tempting to conclude that depth-blindness is why the filters mislead. Our own renderer, where ground-plane depth is analytic, says otherwise. Define a condition's *depth tilt* as the slope of the residual ε of Equation (2) across six depth bins spanning 2–140 m, signed as τ is, so that a negative tilt again means the far field lost more. Across all 42 condition–strength pairs our renderer can produce, tilt and damage correlate at $\rho = +0.506$, and $+0.538$ as a partial correlation holding how much contrast was removed fixed—backwards for the explanation. The conditions that strip the far field hardest do the *least* damage, while the one that does by far the most tilts hard to the *near* field (night *s6*, tilt $+1.198$). We retract the explanation and keep the measurement. Nor is the signature universal among filters: blur and pixelation reach $\tau = -0.19$ and -0.29, distant things being smaller.

5 Ten Laps on a Closed Circuit

Everything so far is zero-shot; does augmentation fix it, and does what you augment *with* matter? We rebuilt the environment as a closed circuit—a smooth loop of 208 m, minimum corner radius 12 m, ten laps—so errors accumulate. Three regimes, two converged seeds each: **clean** sees no corruption, **filter-augmented** post-hoc filters only, **diverse** those filters *and* re-rendered conditions. Five engine conditions are held out of every regime and used only at test time.

Three things follow. All three regimes complete ten laps on the clear condition and on six of the nine test conditions; only the remaining three separate them. The clean-trained agent fails reliably: at engine night *s6* it completes 5.19 of ten laps and leaves the road in half of all runs, the two seeds agreeing closely (5.12, 5.26). Filter augmentation helps against that particular night (8.37 laps) but not against held-out physical conditions

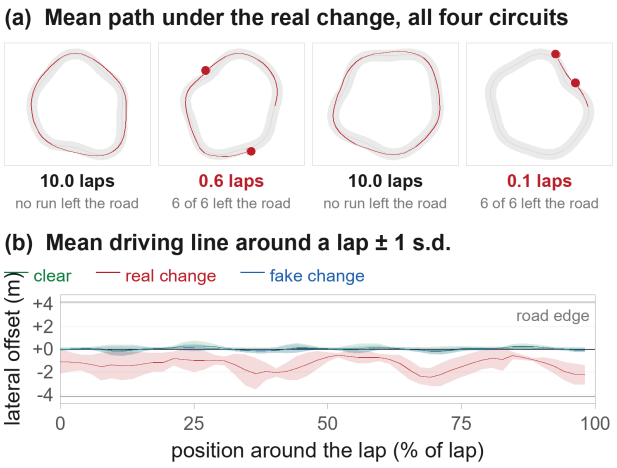


Figure 5: One clean-trained agent, ten laps, 24 runs per condition. (a) its mean path under the in-engine night on each evaluation circuit, drawn only over the part actually driven, with a marker wherever a run left the road. (b) the mean driving line around a lap, ±1 s.d.: the matched filter lies on the clear condition; the night does not.

in general: averaged over the five it completes 8.68 laps against the unaugmented 9.05, so of the two it is slightly the worse. On the post-hoc conditions it trained on it is flawless. The regime that also saw re-rendered conditions is best on every held-out physical condition, all ten laps in both seeds (10.01, lateral error 0.69 m); its one failure is a *post-hoc* condition (8.78 laps under the strongest darkening filter).

Training on filters buys robustness mainly to filters: what transferred was having seen the renderer change the world, not the picture changed afterwards.

The same change, faked. Figure 5 follows one clean-trained agent around all four circuits under three conditions matched on measured picture change—clear, the in-engine night ($L_1 = 38.3$) and the darkening filter closest to it ($L_1 = 36.9$)—24 runs each. Clear and filtered are indistinguishable: both complete 10.01 laps with no excursion, at a mean lateral error of 0.23 m clear and 0.18 m filtered—the filter does not degrade driving at all. The engine night completes only 5.19 laps and puts half of all runs off the road, and where it survives its driving line loosens almost sixfold (0.23 to 1.34 m). Survival is a property of the circuit: two of the four are driven for all ten laps, and on the other two every one of six runs leaves the road inside the first lap—a single test circuit would have reported total failure or none.

6 Conclusion

Manufacturing weather with filters fails selectively, which is worse than failing outright. Held to the same change in the picture the fog filter reproduces physical fog at every strength, while the darkening filter reports as harmless the one condition that reliably puts the car off the road—so a screen that passes on fog gives no assurance about night. Why remains open: the natural explanation—that a filter cannot place its damage by distance—is true of image statistics but not of the driving outcomes, because the change that does the damage is itself depth-blind. The same asymmetry governs training: post-hoc augmentation left our agents slightly worse against physical conditions than no augmentation, and re-rendered ones left them better than either.

Scope and future work. These are simulator measurements: the CARLA part concerns image statistics, not driving, and panda3d fog is not scattering physics. We test only night and fog, the two conditions whose named filter can match them on picture change; widening the set first needs filters that can imitate the others. Our night has no headlights, which plausibly overstates what darkness does to a lit vehicle; nor can one darkening factor imitate a headlit road, bright near and dark far, a case we have not tested. Our agents are small convolutional PPO policies. All nine lost more to the night than to its filter at full strength despite a thirty-four-fold spread in fragility, but stronger or pretrained agents may lose less to both, shrinking the gap without making the filter more faithful; end-to-end architectures in a full CARLA closed loop would test this. Two converged seeds per regime leave the closed-circuit comparison an indication rather than a proof, and a third filter-augmented run that failed to converge (−167 mean episode reward; all others +509 to +713) was excluded and replaced. Finally, τ is blind to illumination, leaving absolute luminance the likeliest, and testable, remaining explanation.

Acknowledgements

This publication is co-funded by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie COFUND Postdoctoral Programme grant agreement No. 101081355–SMASH, and by the Republic of Slovenia and the European Union from the European Regional Development Fund. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work was also supported by the Slovenian Research and Innovation Agency (ARIS), research programme P2-0241 "Synergetics of complex systems and processes".

References

[1] Felipe Codevilla, Antonio M. López, Vladlen Koltun, and Alexey Dosovitskiy. 2018. On offline evaluation of vision-based driving models. In *European Conference on Computer Vision (ECCV)*, 246–262.

[2] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio López, and Vladlen Koltun. 2017. CARLA: an open urban driving simulator. In *Conference on Robot Learning (CoRL)*.

[3] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*.

[4] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. AugMix: a simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*.

[5] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. 2019. Benchmarking robustness in object detection: autonomous driving when winter is coming. *arXiv:1907.07484*.

[6] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-baselines3: reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22, 268, 1–8.

[7] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126, 9, 973–992.

[8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv:1707.06347*.