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.

Class Activation Map Stability as an Uncertain Guide to Denoising Gain for Frozen Pedestrian Detectors Under Gaussian Noise

Vo Thanh Kiet
Technical University of Ostrava https://orcid.org/0009-0002-3278-8755
Rene Jaros
rene.jaros@vsb.cz

Technical University of Ostrava https://orcid.org/0000-0003-3346-6467
Ly Duc Minh
minh.ld@vlu.edu.vn

Van Lang University https://orcid.org/0000-0003-0200-7153
Petr Bilik
Technical University of Ostrava https://orcid.org/0000-0001-8655-778X
Radek Martinek
Technical University of Ostrava https://orcid.org/0000-0003-2054-143X

Research Article

Keywords: Pedestrian detection, Image denoising, Explainable artificial intelligence, Attention preservation, Preprocessing sensitivity, Image bootstrap

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11681327/v1

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

Additional Declarations: The authors declare no competing interests.

Class Activation Map Stability as an Uncertain Guide to Denoising Gain for Frozen Pedestrian Detectors Under Gaussian Noise

Vo Thanh Kiet [1], Rene Jaros [*1], Ly Duc Minh [†2], Petr Bilik [1], and Radek Martinek [1]

[1]Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, 17. listopadu 2172/15, Ostrava, 708 00, Czechia
[2]Faculty of Commerce, Van Lang University, Ho Chi Minh City, Vietnam

November 9, 2026

Abstract

Denoising can improve image quality without helping frozen pedestrian detectors. We ask whether the Attention-Preservation Score (APS), which correlates principal-component class activation maps of clean and processed images, tracks detection gain from denoising. We evaluate eight detectors, six pipelines and five Gaussian noise levels on 160 images, with geometry, polarity and attribution controls mainly on a reference detector, bootstrap intervals and an exploratory attention-guided pixel swap. Repairing incomplete patch coverage raises learned-denoiser APS by 0.09, detection change uncertain, and Block-Matching and 3D Filtering keeps the highest APS at the strongest noise in seven of eight detectors. APS is positively associated with detection gain for the reference detector, but pooled over detectors before repair the association is +0.02 (95% interval −0.15 to +0.17) and protocol-dependent; the primary swap interval includes zero. APS is a detector-specific stability measure, not a validated denoiser selector.

Keywords Pedestrian detection; Image denoising; Explainable artificial intelligence; Attention preservation; Preprocessing sensitivity; Image bootstrap

1 Introduction

Image denoising and pedestrian detection optimise different outcomes: a denoiser can improve peak signal-to-noise ratio (PSNR) or structural similarity (SSIM) [1] without improving the decisions of a frozen detector. Training a denoiser with a joint reconstruction and task loss couples the two objectives [2], but an existing detector may have to operate with a separately trained pipeline whose image transformations and detection performance must both be measured. The hypothesis of structural preservation proposes that denoising helps when it preserves the image structure needed by the downstream model [3]. An attention map is one operational way to examine this idea inside a detector, but map stability is only a proxy: a change in a class activation map (CAM) can reflect preprocessing geometry as well as the removal of useful image content.

We study the Attention-Preservation Score (APS), a measure of spatial activation stability relative to a reference map, on eight frozen pedestrian detectors, six processing pipelines and five levels of Gaussian noise. Unlike PSNR and SSIM, APS depends on the detector, since the same image pair can score differently at different checkpoints; that distinction does not establish a predictive advantage, because image metrics can still rank processing options for a fixed detector and APS needs a clean counterpart that may be unavailable in deployment.

*Corresponding author: rene.jaros@vsb.cz
†Corresponding author: minh.ld@vlu.edu.vn

With the detector weights fixed and a common evaluation split, we ask three questions. Which pipeline differences in APS survive changes to preprocessing geometry, map polarity, attribution and the repair of incomplete reconstruction coverage? How is APS associated with the denoising gain within and across the sampled detector groups, and how far does that association depend on the measurement protocol? What does a probe that swaps pixels selected by attention establish once full coverage, multiple control locations and image-level dependence are accounted for?

The main contributions of this study are summarised as follows:

- We specify APS as an operational measure of spatial activation stability with an explicit sign convention, a requirement for a clean reference, and a companion target-mass statistic ($\mathrm{APS}_{\mathrm{tgt}}$, reported only under the archived protocol), paired with controls for preprocessing geometry, map polarity, and attribution choice.
- Under a unified protocol hooking the activation map inside the detection forward pass, we evaluate eight fixed checkpoints of five You Only Look Once (YOLO) and three Residual Network (ResNet) variants: the APS–gain association has positive point estimates within each group, with intervals including zero when all noise levels are pooled; it is near zero for the eight checkpoints taken together and depends on the measurement protocol and on score offsets across checkpoints.
- We regenerate the three learned denoising pipelines at full patch coverage, turning the artifact from incomplete coverage into a controlled comparison in which Block-Matching and 3D Filtering (BM3D) retains the highest APS at $\sigma = 30$ in seven of the eight checkpoints.
- We rerun the exploratory pixel swap probe guided by attention on full-coverage images with 20 seeded random control locations per image; its primary interval still contains zero, preventing a claim of an isolated structural mechanism.

Section 2 reviews related work, Section 3 defines the attention map, APS, its association with denoising gain and the pixel swap probe, Section 4 describes the images, checkpoints, detection protocols and statistics, Section 5 reports the results, and Sections 6 and 7 discuss and conclude. The Supplementary Material, which follows the reference list of this preprint and whose tables are also supplied as a spreadsheet, contains the complete matrix of F_1 scores (the harmonic mean of precision and recall), the results retained under the archived protocol, implementation and provenance details, further sensitivity analyses and diagnostics, and the measurement and reproducibility details (Supplementary Appendices A–G).

2 Related Work

Evaluating whether image restoration preserves the spatial activation patterns of a frozen detector connects two lines of work, restoration and robustness for frozen detectors and the quantitative evaluation of saliency stability and faithfulness, with our previous studies of structural preservation in pedestrian detection, summarised at the end of this section. Supplementary Table S12 summarises representative prior studies alongside the scope of the present work.

2.1 Restoration for Frozen Detectors

Object detection spans convolutional and transformer designs with different feature and prediction pathways [4, 5, 6], but performance on clean images does not establish robustness to corrupted inputs. Classification experiments show that blur, noise and compression can disrupt recognition [7], ImageNet-C and ImageNet-P broaden this evaluation to common corruptions and perturbations [8], and corruption benchmarks for detection examine degraded inputs across object detection datasets [9]; none of these determines whether a particular denoiser helps a fixed pedestrian detector. Training conditions also matter: a network's reliance on shape and texture can change with its training distribution [10], so we do not interpret a detector family label, or correlation between its activation maps, as a direct measure of structural cue use. Unlike training a denoiser through a fixed recognition network [2], the present study evaluates separately trained pipelines placed before frozen detectors. Classical BM3D [11] and the Denoising Convolutional Neural Network (DnCNN) with residual learning [12] exemplify distinct restoration approaches, but an algorithm name alone does not

establish how a particular implementation, with its training data, noise settings and reconstruction coverage, affects detection.

2.2 Saliency Stability and Faithfulness

Attribution methods differ in the quantity they explain. Gradient-weighted Class Activation Mapping (Grad-CAM) [13] and its generalised extension Grad-CAM++ [14] use gradients of a scalar target; Score-weighted Class Activation Mapping (Score-CAM) [15] and Ablation-CAM [16] avoid gradients but still require a target, and Integrated Gradients [17] attributes a target along a path from a baseline input. Eigen Class Activation Mapping (EigenCAM) [18] instead projects activations and avoids choosing a prediction target, but it is insensitive to class: its dominant activation pattern need not explain a particular pedestrian prediction. For object detection, attribution using gradients requires specifying which class scores or detections form the target, a modelling choice rather than an inherent failure of gradient-based CAM, and Detector Randomized Input Sampling for Explanation (D-RISE) [19] perturbs inputs to explain both category and localisation responses. A map can thus remain similar while the detections it accompanies change, so agreement over the whole map alone does not test faithfulness to individual objects.

Quantitative saliency evaluation perturbs inputs with random masks [20], optimised deletion masks [21] or extremal perturbations under a mask-area constraint [22], and the Remove and Retrain (ROAR) benchmark retrains the model after removing the features an attribution ranks highest [23]; our intervention instead exchanges pixels between two processed versions of the same image at fixed weights. Stability also differs from other reliability criteria: randomising model parameters tests an explanation's dependence on learned weights [24], input invariance concerns transformations that preserve the model's relevant behaviour [25], in classification experiments label-preserving adversarial perturbations can substantially change explanations [26], and infidelity and sensitivity measure further properties [27]. Studies of the reliability of saliency metrics [28] and the Quantus toolkit [29] motivate specifying the evaluation criterion instead of treating one agreement statistic as a complete assessment, and representation comparison distinguishes different feature spaces [30].

The COnsistency-SEnsitivity (COSE) metric tests whether saliency maps of image classifiers change consistently under data augmentation and remain sensitive to changed predictions, and its authors report that saliency methods tend to explain transformer-based models better than convolutional ones [31]; Sarkar et al. [32], evaluating CAM robustness to noise in classification, caution that stable EigenCAM maps can be weakly responsive; stability and faithfulness must therefore be assessed separately. A close precedent for APS is Debiased-CAM [33], which scores the Grad-CAM maps of image classification and captioning models under blur, colour temperature shifts and changes between day and night lighting by their pixelwise Pearson correlation with the map that a conventionally trained model gives for the unperturbed image, the type of correlation that APS averages; training with auxiliary tasks that predict the explanation and the perturbation level raises both task performance and this score. The present use concerns a different outcome: the gain in pedestrian detection F_1 from placing a denoiser before a frozen detector. We do not claim that quantitative saliency stability or architecture dependence is new, or that the present experiment establishes a universal calibration for saliency metrics.

Among our previous studies, the Structural Preservation Score (SPS) [3] measures fidelity in image space without consulting detector activations; companion studies examine denoising and detector training strategies [34] and pipeline selection for frozen detectors with modern restorers and degradation model matching [35]. This analysis uses the same images from the Person and Person-Like Objects (PnPLO) dataset [36] and the same processing pipelines as that programme, but examines spatial activation stability in fixed checkpoints, and its numbers must not be interchanged with results from retraining detectors or other splits.

3 Method

3.1 An Operational Spatial Attention Map

Let f_d be a fixed detector checkpoint, x_i a clean image and $x_i^{(m,\sigma)}$ its version under pipeline m and noise level σ. Activations at a declared layer, $\boldsymbol{A} \in \mathbb{R}^{C\times H\times W}$, are reshaped into $\hat{\boldsymbol{A}} \in \mathbb{R}^{HW\times C}$ and centred within

each column to give $\tilde{\boldsymbol{A}}$. If $\tilde{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, the EigenCAM projection is:

$$\boldsymbol{M} = \text{reshape}(\tilde{\boldsymbol{A}}\boldsymbol{v}_1, H \times W). \tag{1}$$

"Attention map" is an operational name for this spatial activation summary, not learned attention weights or a faithful attribution of each detection. The principal component has arbitrary polarity, so the implementation orients it towards the activation energy map $\boldsymbol{E} = \sum_c \boldsymbol{A}_c^2$, flipping $\boldsymbol{M}$ when the centred inner product is negative, and applies min–max normalisation to $[0, 1]$. Independently oriented clean and processed maps need not have equivalent semantic polarity, particularly near a zero anchoring inner product, so signed and absolute correlations are reported separately.

3.2 Attention-Preservation Score

Preservation of input attention averages the spatial Pearson correlation of matched clean and processed maps:

$$\text{APS}_{\text{corr}}(d; m, \sigma) = \frac{1}{N} \sum_{i=1}^{N} \rho\Big(\boldsymbol{M}(f_d, x_i), \boldsymbol{M}(f_d, x_i^{(m,\sigma)})\Big). \tag{2}$$

The detector, hook layer and map grid are fixed within each pair, and a large score indicates preservation of this particular spatial activation pattern. The main results use the signed correlation; the mean $|\rho|$ sensitivity measures agreement up to polarity and is not an interchangeable definition of APS. The companion statistic for attention on the target is:

$$\text{APS}_{\text{tgt}} = \frac{\sum_{p \in \mathcal{B}} \boldsymbol{M}_p}{\sum_p \boldsymbol{M}_p}, \tag{3}$$

where $\mathcal{B}$ is the union of ground-truth (GT) boxes for the person class rasterised on the map grid, with zero for an empty target set and numerical safeguards for degenerate maps. The aggregate includes images without annotated persons; unlike APS_{corr} it requires annotations, and its absolute scale depends on box area and map normalisation.

3.3 Denoising Gain and Detector Dependence

We measure denoising utility with:

$$\Delta F_1(d; m, \sigma) = F_1(d; m, \sigma) - F_1(d; \text{noisy}, \sigma). \tag{4}$$

This is a gain relative to leaving the same noisy input untouched, not the accuracy lost to noise or the processed detector's absolute accuracy, and a small gain alone does not establish greater noise robustness. For each detector group we remove each (m, σ) condition mean from APS and ΔF_1 and correlate the residuals; the analysis restricted to the reference checkpoint removes means by σ instead. This estimates an association in the sampled checkpoints after the declared condition adjustment, rather than prediction performance on an unseen detector or a test of architecture as a causal factor.

3.4 Exploratory Pixel Swap Guided by Attention

At $\sigma = 30$ on the reference checkpoint, g_+ and g_- denote the BM3D and autoencoder (AE) versions of the same image; the autoencoder side is regenerated with the stored weights at complete patch coverage, so the graft avoids the coverage defect of Section 4.3, and its original incomplete grid is kept as a second input variant. A binary mask M selects the highest τ fraction of the clean EigenCAM map, bilinearly resized to the 640×640 inference grid, cropped to the recorded window without padding and resized to native image dimensions, with the exact required area. The primary fraction $\tau = 0.2$ was fixed before the controlled rerun in experiment kit v1.1 (Supplementary Appendix C). The graft targeting these high attention values (the high graft) is:

$$\hat{x}_{\text{high}} = (1 - M) \odot g_- + M \odot g_+. \tag{5}$$

Three further mask specifications were fixed as sensitivities before that rerun: the activation energy map at $\tau = 0.2$, and the EigenCAM map at $\tau = 0.1$ and $\tau = 0.3$. Crossing the four specifications with the two input variants gives eight cells, and a more favourable sensitivity does not replace the primary specification.

Each cell compares five arms: the BM3D and autoencoder images, the high graft, a reverse graft replacing BM3D pixels inside M with autoencoder pixels, and a random control translating M by a seeded random cyclic offset that preserves its area and overlaps the original mask by at most 25% of mask area; twenty such offsets are drawn per image from independent seeds, in place of the single saved location of the earlier probe (every draw satisfied the overlap rule; Supplementary Appendix G.6). Within each cell, the high, reverse and random grafts edit exactly the same number of pixels. These controls compare the location and direction of a pixel replacement and do not isolate high-frequency structure: equal edited area does not ensure equal consequences of replacement boundaries. The graft is applied to regenerated autoencoder images, not the archived files evaluated elsewhere, so we analyse the intervention as exploratory evidence about these processed images.

4 Experimental Setup and Statistical Analysis

4.1 Images, Checkpoints and Processing

The PnPLO dataset includes persons and distractors resembling people, such as statues and mannequins [36]. Our evaluation split has 160 images and 203 annotated persons; 67 images contain no person. It is not established as an independent holdout from all pipeline selection, because the archived training code for the convolutional autoencoder with particle swarm optimisation (CAE+PSO) fits denoiser hyperparameters on validation images, and optimising a selection criterion on finite evaluation data can introduce a selection bias [37] that resampling those images does not remove. A check of exact pixel identity across the training, validation and test directories also finds two evaluation images that duplicate training images and one that duplicates a test image, each annotated only with a person-like object and so among the 67 images without persons; they are retained, so the split is not fully disjoint from the training data, and near duplicates were not assessed. The analyses are conditional on this sample and the available trained checkpoints.

YOLOv9m is the reference detector. The broader analysis uses five YOLO variants (YOLOv8m, YOLOv9m, YOLOv10m [38], YOLOv11m and YOLOv12m [39]) and three YOLO detectors with ResNet-18, ResNet-50 [40], or ResNet-18 with a global attention mechanism (GAM) [41] backbone, called the groups of YOLO and ResNet variants; the first group is architecturally heterogeneous, so the grouping is an empirical stratification, not a claim of a shared backbone. In each detector the map is hooked at the last module returning one four-dimensional feature tensor before the detection head, on a 20×20 grid except for the two ResNet-18 variants, which give 40×40. Noise is additive Gaussian, $\sigma \in \{1, 5, 10, 20, 30\}$ on the 8-bit intensity scale, across six pipelines, the noisy image, Gaussian filtering, BM3D, autoencoder, DnCNN and CAE+PSO. The available processed images are one set of noise realisations per level, and we do not infer variability over fresh noise draws. Checkpoint identities, hook layers, processing parameters and provenance limits are in Supplementary Appendix C.

Two image sets are distinguished throughout: the archived images, the stored files that the original analysis evaluated, and the same images with the three learned pipelines (autoencoder, DnCNN and CAE+PSO) regenerated at full coverage with the stored denoiser weights, shortened to repaired coverage (learned pipelines only). The noisy, Gaussian filtering and BM3D images exist only as archived files, identical in both sets, and a regenerated output is not a reproduction of the archived file it replaces. A third set, the same regeneration on the original patch grid, is used to measure the coverage repair, to check continuity with the archived images and to test the sensitivity of the pixel swap.

4.2 Detection Protocol

All detection results use each detector's own frozen checkpoint trained on clean images, input size 640, confidence threshold 0.25 and non-maximum suppression (NMS) intersection-over-union (IoU) threshold 0.7, person class only. Two assignment rules are named wherever a number is reported. Under the "archived

protocol" a prediction goes to the person box with the highest IoU if IoU is at least 0.5 and that box is unmatched, with the attention map from a separate forward pass after resizing to a square; under the "unified protocol" predictions are assigned one to one by decreasing confidence, again at IoU at least 0.5, with the map hooked from the forward pass that produced the boxes. Assigned predictions are true positives (TP), unassigned predictions false positives (FP) and unmatched ground-truth boxes false negatives (FN), summed across images for micro-averaged precision, recall and F_1. Every quantity recomputed in the controlled rerun uses the unified protocol, including the coverage repair and the F_1 matrix (Supplementary Appendix A); quantities retained from the archived measurement path (Supplementary Appendix B) use the archived protocol and are labelled as such. Per-image scores under the two protocols are not interchangeable. A third, looser rule of the archived recovery diagnostic is defined in Supplementary Appendix F.

4.3 Incomplete Patch Coverage and Geometry Controls

The archived reconstruction by learned denoisers can leave bottom or right border strips without patch coverage. The strip detector of Supplementary Appendix C flags visibly dark strips in 126 of the 160 images (mean flagged area about 8.97% of the frame at $\sigma = 30$), and a direct count of the overlap-add patch grid finds pixels without any patch prediction on 127 stems; the two counts answer different questions, and both are unchanged across noise levels. These are preprocessing artifacts, not evidence that a denoiser's learned objective inherently removes task structure.

We compare three APS evaluations at the reference checkpoint: the "original" uses the full saved image and its CAM, the "masked" correlates only CAM cells fully inside the common valid region, and the "cropped" crops both image counterparts to that region before recomputing activations, a sensitivity analysis rather than a reconstruction of missing content. One image with incompatible stored exchangeable image file format (EXIF) geometry is excluded, leaving 159 matched images (the original 160-image scores are retained separately). Two reruns verified the archived measurement path against the original per-image values (Supplementary Appendix F); neither shows that the unified protocol reproduces archived APS.

4.4 Uncertainty and Sensitivity

For the attention and association analyses we apply the nonparametric bootstrap [42] at the image level (5,000 replicates, random seed 20260914), with one multinomial weight matrix shared by all detectors, pipelines, noise levels and image sets; each replicate recomputes mean APS, micro-averaged F_1, the difference from the noisy baseline and the adjusted correlation. The pointwise 95% percentile confidence intervals (CIs) quantify uncertainty from sampling images conditional on these checkpoints, processing outputs and noise realisations, not sampling of architecture families or training seeds. Nominal p-values of the partial-correlation test after adjustment for G condition groups use $N - G - 1$ degrees of freedom, still assume independent residuals and are treated only as exploratory diagnostics. Leave-one-checkpoint-out, signed versus absolute APS, and geometry analyses assess sensitivity without selecting the convention that gives the smallest p-value.

For the swap probe, intervals use 5,000 replicates from a separate generator (random seed 14092026), each drawing 160 image indices with replacement, so images are the resampling units; only 93 carry person annotations and the other 67 can contribute only false positives. The prespecified estimand in each cell is the micro-averaged F_1 of the high graft minus the mean over the 20 random translations, so the random locations enter as a fixed average rather than as a second level of resampling. All eight cells use one set of resample vectors and are dependent. Their intervals are pointwise exploratory summaries, without multiplicity adjustment or a claim of simultaneous coverage; the limits are the 2.5 and 97.5 percentiles. A separate exploratory sign-flip test of per-image differences is defined in Supplementary Appendix F; it tests a different estimand and does not replace the primary contrast.

5 Results

5.1 Attention Preservation in the Reference Detector

Figure 1 plots the reference checkpoint APS under the unified protocol on the archived images against noise level (values in Supplementary Table S13). At low noise the noisy image scores highest; from $\sigma = 10$ onwards BM3D has the highest mean among the saved pipelines, reaching 0.7821 at $\sigma = 30$ against 0.4550 for the noisy baseline and 0.5116 to 0.5505 for the three learned denoisers. These values describe the implemented pipelines, including their reconstruction artifacts. They do not establish a general ordering of classical and learned denoising algorithms. All 30 cells score higher than under the archived protocol (Supplementary Table S2), with the same best pipeline at every noise level, but per-image scores are not interchangeable between the protocols (Supplementary Appendix G.1).

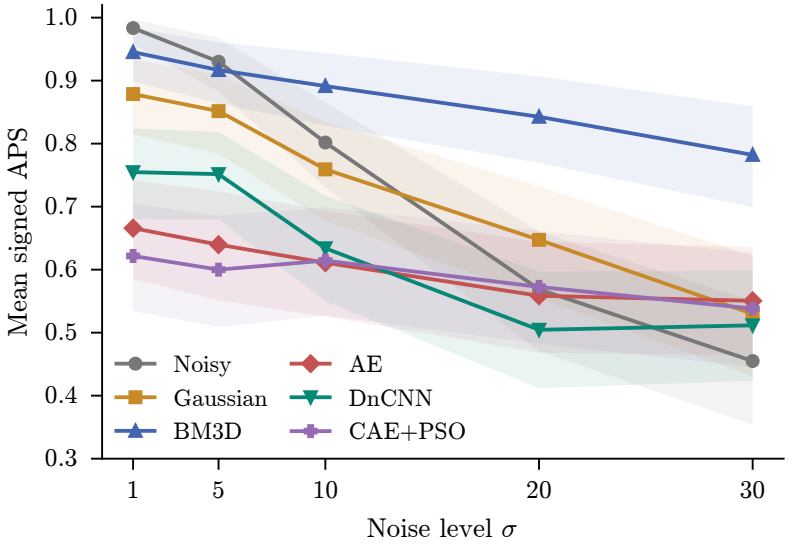


Figure 1: Reference YOLOv9m mean signed APS by noise level and pipeline, unified protocol on the archived images, with pointwise 95% bootstrap bands over images.

For the reference checkpoint, APS and denoising gain are positively associated across pipelines and noise levels under the unified protocol, with $r = 0.5034$ on the archived images after adjusting for σ (95% bootstrap CI over images $[0.118, 0.691]$) and $r = 0.6078$ on repaired coverage (learned pipelines only). The nominal test on 19 residual degrees of freedom (25 cells, five noise groups) gives $p = 0.0200$ on the archived images and $p = 0.0035$ on repaired coverage, and at $\sigma = 30$ alone the five denoising pipelines give $r = 0.8463$ from five observations. These descriptive associations do not validate a pipeline selector outside the evaluated conditions.

The matched geometry controls of Section 4.3 use the archived protocol on the reference checkpoint only. They preserve BM3D's leading mean at $\sigma = 30$ under original, masked and cropped evaluation (0.667, 0.669 and 0.661 on 159 images), but under cropping the three learned pipelines, whose original matched means lie below that of Gaussian filtering, rise above it as it falls from 0.416 to 0.314 (Figure 2 and Supplementary Table S14). The ranking below BM3D is therefore sensitive to input geometry, whereas the BM3D result on the reference model is more stable; even the original matched comparisons below BM3D are not clearly distinguished (Supplementary Appendix G.2). Cropping changes the detector input, so this contrast cannot separate artifact removal from changes in context, scale and map polarity.

Attribution and polarity act similarly: BM3D ranks first at $\sigma = 30$ under EigenCAM, activation energy, Grad-CAM and Grad-CAM++ (archived protocol; Figure 3) and under the absolute correlation $|\rho|$, but both choices reorder the pipelines below it, and the reversal between signed and absolute scores of Gaussian filtering

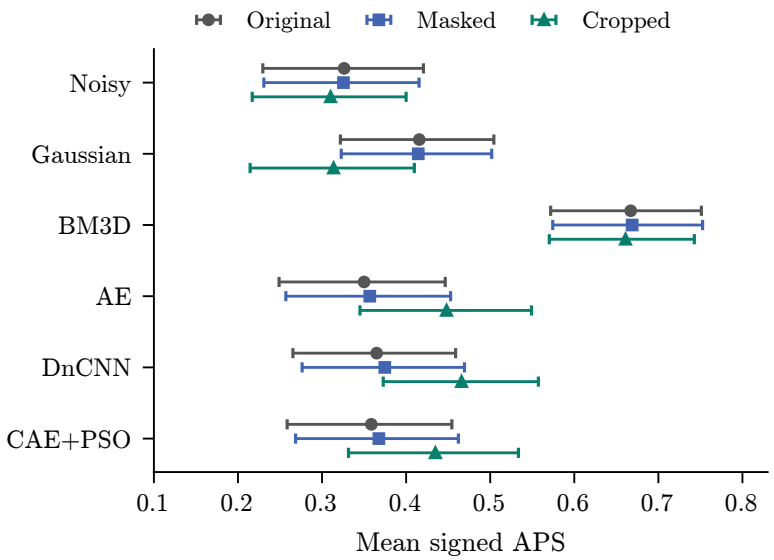


Figure 2: Geometry sensitivity of reference signed APS at $\sigma = 30$ on 159 matched images, archived protocol and reference checkpoint only, with pointwise 95% bootstrap intervals over images.

against the autoencoder and CAE+PSO disappears on repaired coverage (Supplementary Appendix G.5, which also reports the target mass statistic $\mathrm{APS}_{\mathrm{tgt}}$). The gradient methods take the summed score for the pedestrian class over anchors as their scalar target. The archived scores of the three alternative attributions are aggregates, so no confidence intervals are inferred.

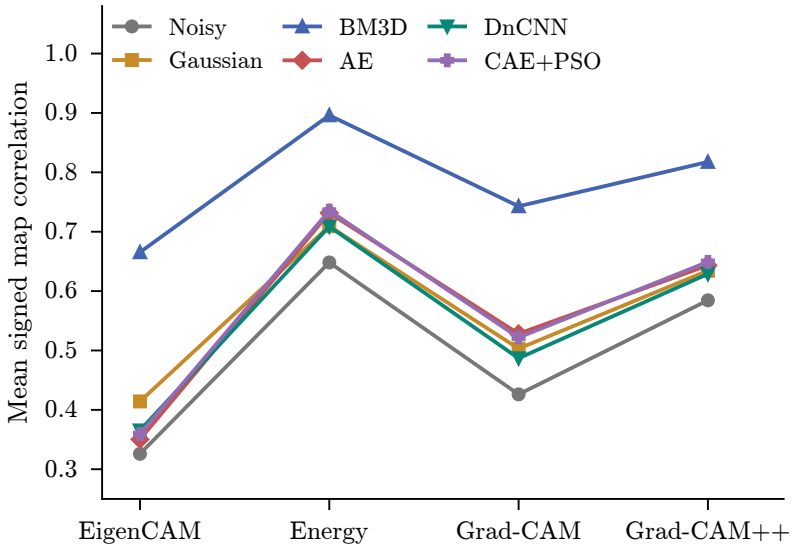


Figure 3: Attribution comparison at $\sigma = 30$ from archived aggregate scores under the archived protocol.

5.2 Coverage Repair and Changes in APS and Detection

Counted on the overlap-add patch grid, the coverage defect of Section 4.3 affects 1,905 of the 2,400 image cells for learned denoisers on 127 of the 160 stems, with an uncovered area averaging 23,646 pixels, or 8.91%

of the frame, on those stems. We therefore regenerated the three learned denoising pipelines with the stored denoiser weights on a repaired patch grid, giving the repaired coverage (learned pipelines only) image set, which covers every pixel at least once in all 2,400 cells (pixel changes in Supplementary Appendix G.3). The repair fixes the patch grid rather than filling black pixels, and the regenerated outputs are not identical to the archived ones: they are a new implementation with complete coverage, not a correction of the archived pipeline.

Table 1 reports, over the 120 cells for learned denoisers and relative to the same images regenerated on the original patch grid (labelled "buggy" in Table 1 and Figure 4), a signed APS rise of +0.0896 [+0.0727, +0.1069], a rise of +0.1176 [+0.0985, +0.1367] in the score restricted to content, excluding padding, and a micro-averaged F_1 move of +0.0061 [−0.0044, +0.0174], with F_1 computed after pooling TP, FP and FN across all 120 cells and images; removing the eight DnCNN $\sigma = 30$ cells leaves these patterns in place. The repair gives a clear APS increase, while the detection change remains compatible with a small decrease or an increase of up to 0.0174, establishing neither equivalence nor absence of a detection effect. Figure 4 separates the five noise levels instead of pooling them as in Table 1, across the eight checkpoints and three learned denoisers, showing (a) signed APS averaged over the 24 cells per noise level and (b) micro-averaged F_1 from pooled detection counts across cells and images, on their own vertical scales; the bars are pointwise 95% paired percentile intervals from 5,000 bootstrap replicates over images. The APS rise holds at every noise level, with every interval excluding zero, and its pooled point estimate falls as noise increases; every F_1 interval includes zero.

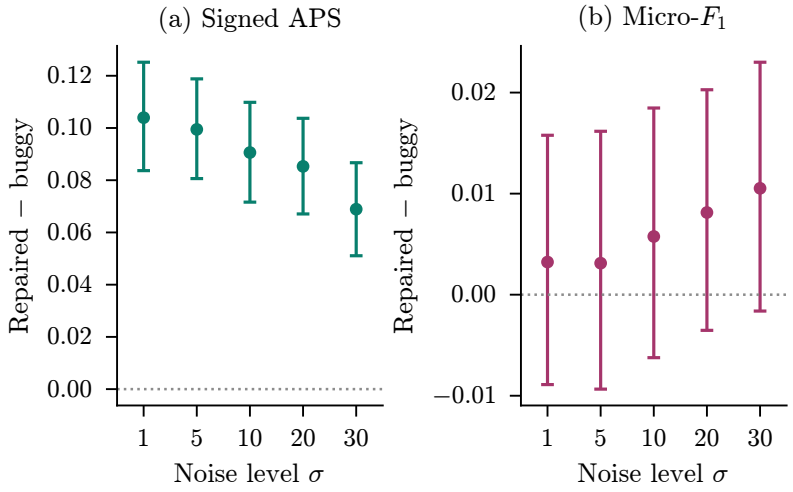


Figure 4: Changes after coverage repair by noise level, repaired minus images regenerated on the original grid: (a) signed APS and (b) micro-averaged F_1, with pointwise 95% bootstrap intervals over images.

Table 1: Effect of the coverage repair on the 120 learned-denoiser cells (repaired minus original-grid regeneration), with 95% bootstrap intervals over images; the final column excludes the eight DnCNN $\sigma = 30$ cells.

Quantity	Repaired − buggy [95% CI]	Cells > 0	Excluding DnCNN σ=30 [95% CI]
Signed APS	+0.0896 [+0.0727, +0.1069]	118/120	+0.0919 [+0.0746, +0.1095]
APS within content	+0.1176 [+0.0985, +0.1367]	120/120	+0.1205 [+0.1010, +0.1402]
Absolute APS	+0.0587 [+0.0480, +0.0699]	120/120	+0.0595 [+0.0486, +0.0707]
Micro-F_1	+0.0061 [−0.0044, +0.0174]	75/120	+0.0060 [−0.0048, +0.0177]

Pooled over the eight checkpoints at $\sigma = 30$ under the unified protocol on repaired coverage (learned pipelines only), mean signed APS is 0.8644 for BM3D, 0.8053 for CAE+PSO, 0.7697 for the autoencoder, 0.6921 for Gaussian filtering, 0.6887 for DnCNN and 0.6238 for the noisy baseline. BM3D leads for seven of the eight checkpoints, the exception being YOLOv8m, where CAE+PSO reaches 0.9749 against 0.9703. Restoring the

missing coverage therefore raises the learned pipelines without displacing the classical one overall.

The DnCNN $\sigma = 30$ cells are an exception we report rather than repair: their pooled micro-averaged F_1 is 0.6689 on the archived images but 0.4317 and 0.4369 on the regenerated images before and after the grid repair, both below the archived noisy baseline of 0.5145. The stored $\sigma = 30$ DnCNN weights under the declared regeneration path fail to reproduce the archived outputs (Supplementary Appendix C), so comparisons across those image sets do not assume continuity of the generating weights and pipeline. The regenerated autoencoder images also give lower F_1 than the archived ones at $\sigma = 30$, at all eight checkpoints (Section 5.4 gives the values at the reference checkpoint).

Figure 5 shows two PnPLO examples [36] at $\sigma = 30$ on the reference YOLOv9m checkpoint. We selected contrasting cases from the saved outcomes before inspecting photographs: among images containing persons, one case nearest the median positive per-image $F_1(\text{BM3D}) - F_1(\text{repaired AE})$ difference, restricted to BM3D recovering more true positives than the noisy input, and one case nearest the median negative difference. Ties use the numeric image identifier. These rules select images 27 and 6; the selection record in the deposited figures archive lists all candidate rankings. This illustration is selected by outcome, not representative sampling. Solid blue boxes mark TP, solid orange boxes FP, dashed magenta boxes FN and cyan boxes GT, with detection counts beneath each panel; the black borders in image 27 are already present in the clean source. On image 27 (Figure 5a), BM3D recovers one of three persons but also produces a false positive; the AE output regenerated at full coverage misses all three and also produces a false positive. On image 6 (Figure 5b), both denoisers detect the annotated person, but BM3D adds a false positive. All displayed detections use the unified protocol, under which an FP can overlap a real person when it duplicates an already matched detection or falls below the matching IoU threshold. Thus the examples expose both successful detections and remaining errors without estimating how often either occurs.

5.3 Associations Within and Across Detector Groups

After adjustment for the 25 denoiser–noise conditions, APS–ΔF_1 correlations on the archived images under the unified protocol are +0.46 (95% interval $[-0.03, 0.61]$) for the three ResNet variants, +0.24 ($[-0.04, 0.41]$) for the five YOLO variants and +0.02 ($[-0.15, 0.17]$) pooled. Table 2 (plotted in Figure 6) repeats the analysis on repaired coverage (learned pipelines only). The archived protocol's negative pooled estimates, −0.21 ($[-0.35, 0.04]$) over all noise levels and −0.35 ($[-0.54, -0.06]$) at $\sigma = 30$ (Supplementary Table S3), do not reproduce at their original size in either image set: at $\sigma = 30$ the archived images give a pooled −0.06 with an interval including zero; over all noise levels the pooled nominal independent-cell test on those images moves from $p = 0.00633$ to $p = 0.824$, and in both sets every family estimate pooled over noise levels is positive with an interval including zero. The single interval in Table 2 excluding zero, the ResNet group at $\sigma = 30$ on the archived images (+0.81, $[0.18, 0.93]$; Figure 6b), rests on three checkpoints spanning two CAM resolutions without multiplicity adjustment (Section 4.4) and is reported for completeness, not as a finding; the DnCNN $\sigma = 30$ cells have limited provenance in both image sets (Section 5.2), and none of these intervals establishes inference over a population of architectures.

Table 2: Adjusted APS–ΔF_1 correlations under the unified protocol on the archived images and on repaired coverage, with pointwise 95% bootstrap intervals over images (archived-protocol values in Supplementary Table S3).

	Archived images		Repaired coverage	
Checkpoint group	All noise levels	$\sigma = 30$	All noise levels	$\sigma = 30$
YOLO variants	+0.24 [−0.04, +0.41]	+0.23 [−0.18, +0.57]	+0.17 [−0.11, +0.38]	+0.25 [−0.16, +0.57]
ResNet variants	+0.46 [−0.03, +0.61]	+0.81 [+0.18, +0.93]	+0.52 [−0.08, +0.64]	+0.74 [−0.04, +0.87]
All eight checkpoints	+0.02 [−0.15, +0.17]	−0.06 [−0.29, +0.19]	+0.04 [−0.14, +0.21]	+0.15 [−0.10, +0.36]

At $\sigma = 30$ the group offsets in mean APS and ΔF_1 that produced the archived reversal shrink under the unified protocol (Supplementary Appendix G.4). Such opposing group offsets explain how a pooled association can differ in sign from estimates within each group [43], a caveat for grouped checkpoint analyses, not a claim

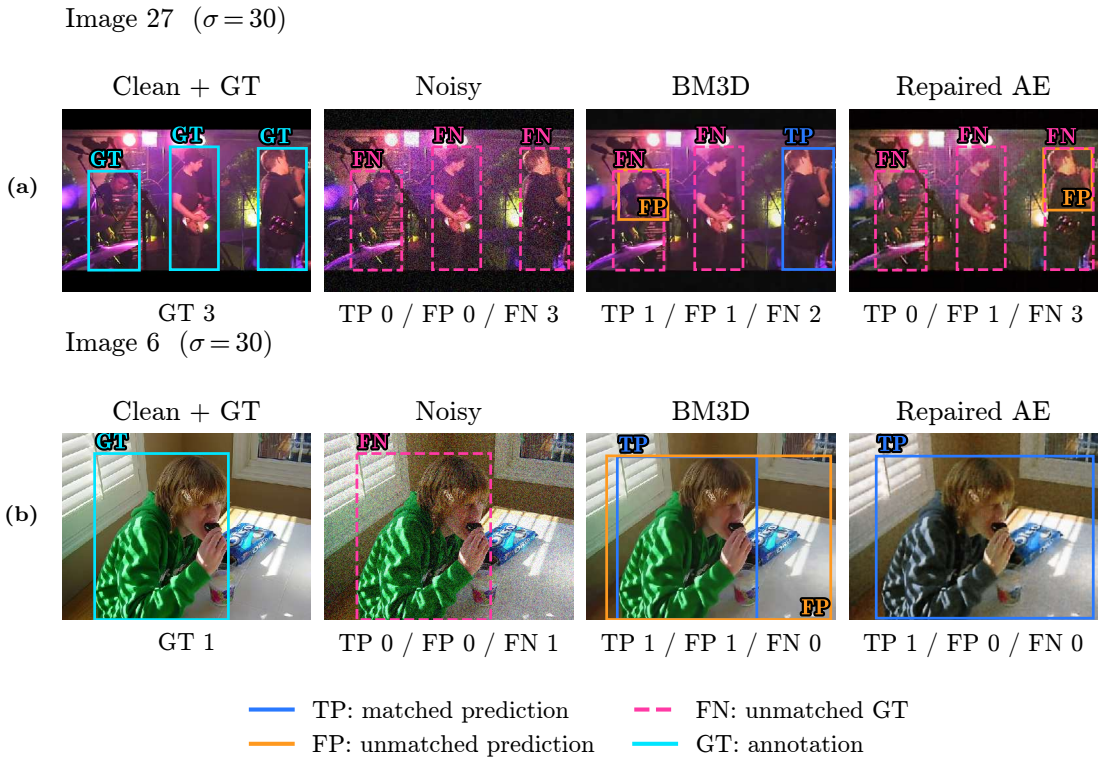


Figure 5: Inputs and stored detections at $\sigma = 30$ for the clean image with person annotations (GT only), the noisy input, the BM3D output and the autoencoder output regenerated at full coverage: (a) image 27 (top row) and (b) image 6 (bottom row).

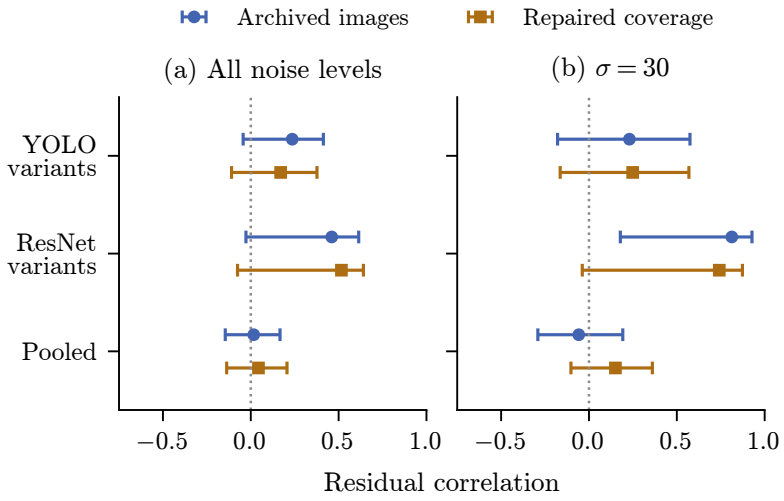


Figure 6: Adjusted associations between APS and denoising gain under the unified protocol on the archived images and on repaired coverage, with pointwise 95% bootstrap intervals over images: (a) pooled across all five noise levels and (b) restricted to $\sigma = 30$.

about the present estimates, whose only pooled value opposite in sign to both groups (-0.06 at $\sigma = 30$ on the archived images) has an interval including zero; training, feature scale, head design and the noisy baseline also differ between checkpoints, so the offsets are not attributed to backbone capacity. Rank associations within each condition and analyses omitting one ResNet checkpoint at a time are sensitivity diagnostics, as is omitting one YOLO variant, which moves that group's estimate between 0.00 (without YOLOv8m) and +0.39 (without YOLOv11m; Supplementary Table S10); no subset or score convention yielding a larger estimate is adopted as the reported result (Supplementary Appendices G.4 and D).

Within each condition, PSNR and SSIM do not vary across detectors, which does not make APS a better metric for selecting denoisers. At $\sigma = 30$, BM3D is best by F_1 for seven of the eight checkpoints on the archived images under the unified protocol, the exception being the ResNet-50 variant, where DnCNN reaches 0.6893 against 0.6872 (Figure 7), and for all eight on repaired coverage, where the F_1 of the regenerated DnCNN $\sigma = 30$ cells drops (Section 5.2); at $\sigma \leq 10$ the advantage of BM3D is less consistent (Supplementary Table S1). The experiment therefore demonstrates score variation across detectors, not a demonstrated advantage in choosing different denoisers for different detectors.

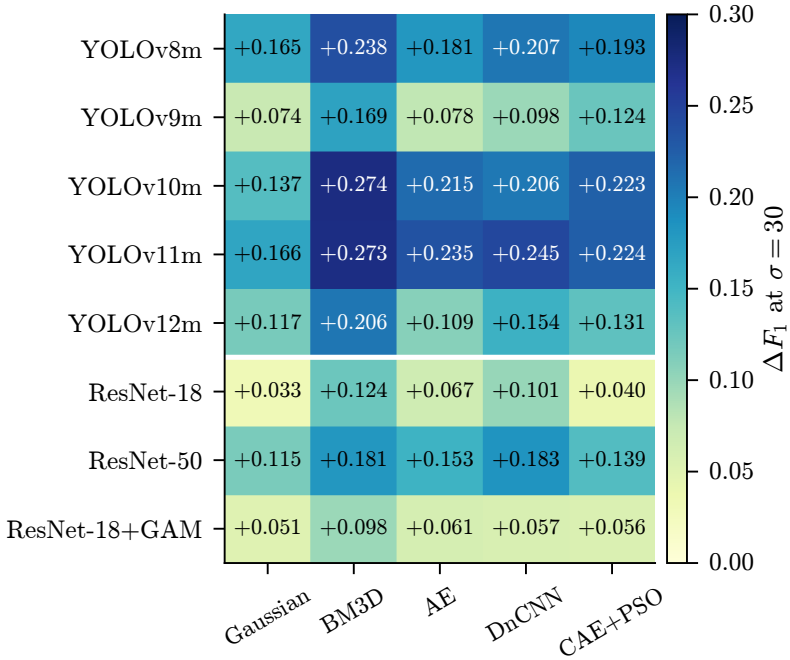


Figure 7: Denoising gain relative to the archived noisy baseline at $\sigma = 30$, on the archived images evaluated again under the unified protocol.

5.4 Exploratory Pixel Swap Results

The controlled rerun called for by the archived probe applies the unified protocol to repaired coverage (learned pipelines only) at the primary cell fixed in advance in Section 3.4, grafting pixels with high attention against controls at 20 seeded random locations. The estimate of its prespecified contrast (Section 4.4) is $+0.0295$ micro-F_1, with a 95% bootstrap interval over images $[-0.0195, +0.0824]$ at seed 14092026, from 5,000 replicates. Table 3 lists the five arms; the 20 random grafts span 0.5000 to 0.5799 micro-F_1. The interval contains zero, so the rerun does not establish an effect specific to attention targeting. BM3D exceeds the autoencoder output regenerated at full coverage by $+0.1866$ $[+0.1247, +0.2536]$ micro-F_1, a substantial gap for the graft to move, and the high graft sits $+0.0647$ $[+0.0119, +0.1186]$ above that autoencoder arm; this precondition check is confounded by the edited pixel budget and is not an estimate of an attention effect.

The autoencoder arm is not continuous with the archived images: under the unified protocol its micro-F_1 is 0.612 on the archived images, 0.489 after regeneration with the stored weights and 0.516 after the coverage repair, although in a 10-image check the PSNR of regenerated and archived outputs against the clean images differs by only about 0.8 decibels (Supplementary Appendix C). The change of protocol, regeneration and repair together widen the archived probe's BM3D-minus-autoencoder gap of +0.072 to the +0.1866 above.

Table 3: Arm micro-F_1 at the primary cell (reference YOLOv9m, $\sigma = 30$, EigenCAM mask, $\tau = 0.20$) with pointwise 95% bootstrap intervals over images; the random row averages the 20 seeded control draws.

Condition	Micro-F_1 [95 % CI]	TP/FP/FN	Recall
BM3D	0.702 [0.650, 0.754]	138/52/65	0.680
Autoencoder (AE), repaired	0.516 [0.442, 0.584]	82/33/121	0.404
AE + high attention graft	0.580 [0.518, 0.641]	101/44/102	0.498
AE + random graft (mean of 20)	0.551 [0.495, 0.602]	92.0/38.8/111.0	0.453
BM3D + reverse graft	0.663 [0.605, 0.716]	122/43/81	0.601

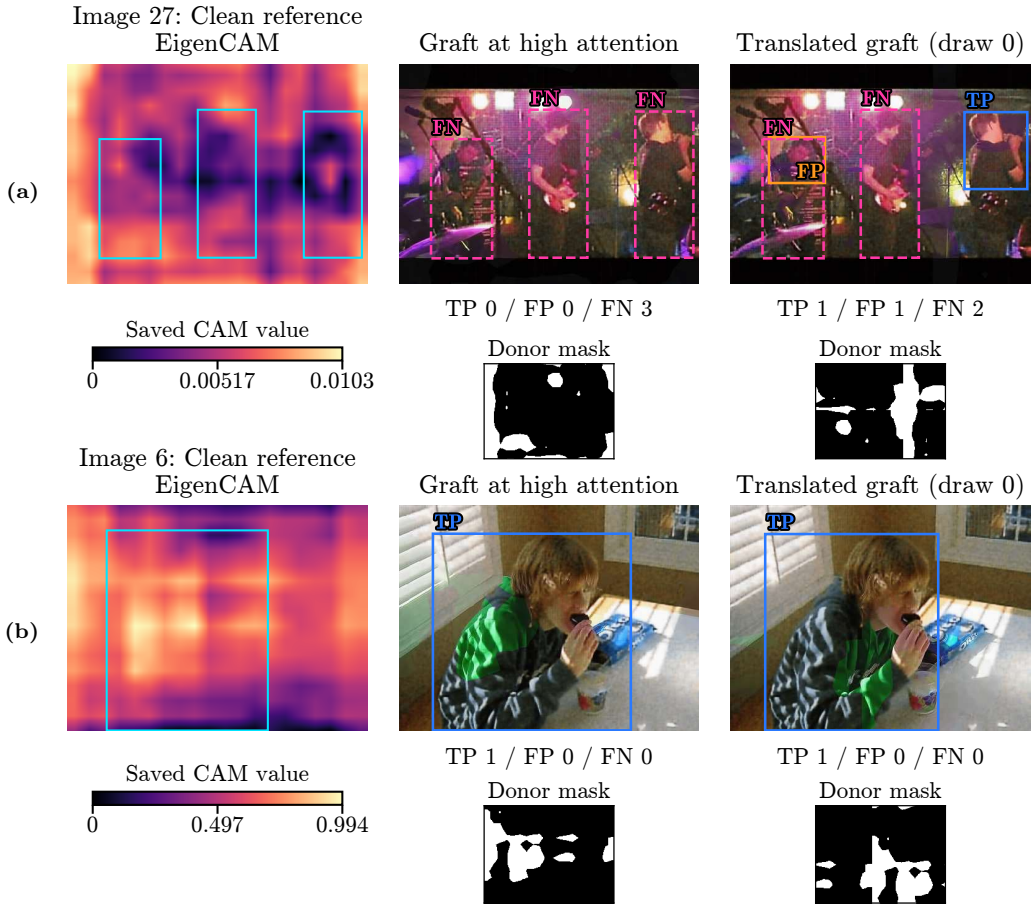
Table 4 crosses the four declared mask specifications with the two regenerated autoencoder image sets (repaired and "buggy" original grid); its Role column names the factors changed from the primary cell, and its Excl. 0 column marks intervals that exclude zero. Six of the eight intervals contain zero; the two that exclude it both use the activation energy map, a sensitivity fixed in advance, and neither is promoted. Under the EigenCAM map the estimate grows with the retained mask fraction, as consistent with an effect of the edited area as of targeting. The mask targeting high attention also edits more ground-truth pixels than the translated control: the mean per-image fraction of the union of person boxes edited is 0.153 against 0.102 (a factor of 1.50, rising to 2.67 under the activation energy map), assigning zero to images without persons. The design therefore cannot separate editing pixels with high attention from editing pedestrian pixels. Supplementary Appendix G.6 reports per-image, continuity, matched-coverage and recovery diagnostics, including a sign-flip test ($p = 0.039$) and a recovery count (23 against a random mean of 13.65) that favour the high graft but concern other estimands and do not resolve this.

Table 4: High graft minus the mean of 20 grafts at random locations for four mask specifications on the two regenerated image sets, with pointwise 95% bootstrap intervals over images.

Images	Map	τ	Δ micro-F_1 [95 % CI]	Excl. 0	Role
Buggy	EigenCAM	0.10	+0.0007 [−0.0428, +0.0448]	no	input + τ
Buggy	EigenCAM	0.20	+0.0413 [−0.0133, +0.0990]	no	input
Buggy	EigenCAM	0.30	+0.0489 [−0.0127, +0.1067]	no	input + τ
Buggy	Energy	0.20	+0.1065 [+0.0658, +0.1505]	yes	input + map
Repaired	EigenCAM	0.10	+0.0089 [−0.0288, +0.0476]	no	τ
Repaired	EigenCAM	0.20	+0.0295 [−0.0195, +0.0824]	no	primary
Repaired	EigenCAM	0.30	+0.0405 [−0.0149, +0.0943]	no	τ
Repaired	Energy	0.20	+0.0651 [+0.0188, +0.1141]	yes	map

Figure 8 illustrates the operation on the two cases of Figure 5. The first column shows the saved EigenCAM from the clean image projected to native coordinates, with cyan GT outlines and a separate colour bar for each image. Each map is min–max normalised on its own over its whole 20×20 grid, which spans the 640×640 inference input including the padding added to make the image square, so brightness is not comparable between images. On image 27 a peak in that added padding (not the black borders of the photograph) far exceeds every value inside the photograph, so its colour bar stops far below 1. The remaining columns show the high graft and translated control with stored detections and binary masks selecting 20% of the image (white uses BM3D donor pixels and black retains the AE recipient with complete coverage). The displayed translated control is draw 0, one realisation fixed for both cases independently of its outcome rather than the mean of 20 controls. On image 27 (Figure 8a) the high graft misses all three persons, whereas this control detects one and adds a false positive. On image 6 (Figure 8b) both grafts detect the sole annotated person

without a false positive. These image-level outcomes illustrate the edit and neither replace the 20-control aggregate nor resolve its uncertainty.



Separate CAM display scales; brightness is not comparable between images.

Figure 8: Primary pixel grafts at $\sigma = 30$ for the cases in Figure 5: the clean-reference EigenCAM map, the graft at high attention and translated control draw 0 with their 20% binary donor masks, for (a) image 27 (top row) and (b) image 6 (bottom row).

With complete denoiser coverage, one-to-one matching and 20 seeded control locations, the rerun's primary estimate is small and positive with an interval containing zero. One checkpoint, noise level, hook layer and denoiser pair, a graft that also changes local noise, colour and texture, the dependence on the saliency map and the asymmetry in ground-truth coverage all limit interpretation, so we report the intervention as exploratory evidence about these processed images, without a mechanistic conclusion.

6 Discussion and Limitations

The evidence supports a specific result for the reference detector: BM3D preserves the measured activation pattern well at high Gaussian noise, its lead survives the evaluated geometry and attribution sensitivities, and under the unified protocol it keeps the highest mean score at $\sigma = 30$ in seven of the eight checkpoints once the learned pipelines are regenerated at full coverage (Section 5.2). The association between APS and denoising gain depends on the protocol and on score offsets across checkpoints: the archived pooled reversal does not reproduce at its original size, every pooled interval in Table 2 includes zero, and the offsets behind the reversal shrink (Section 5.3). Interpreting attention preservation therefore requires reporting the grouping, score definition, measurement protocol and preprocessing, with explicit control of image geometry and sampling unit. The hypothesis of structural preservation remains a motivation, but these analyses do not isolate high-frequency structure as the cause of improved detection: coverage artifacts in learned denoisers, independently anchored CAM polarity and different feature representations can all influence the score, and faithfulness would require tests linking the map to particular predictions.

The scope is limited to fixed checkpoints, three ResNet variants, one set of noise realisations per level and one evaluation split, which archived denoiser selection code used and which shares three images with the training and test directories (Section 4.1); independent evaluation is needed before claims of predictive generalisation, and bootstrap intervals over images cannot replace replication over training seeds or architectures. The geometry controls cover YOLOv9m only; APS requires a clean reference and a declared hook layer, and $\mathrm{APS}_{\mathrm{tgt}}$ requires annotations. Neither a use as a denoiser selector, which should be tested on held-out conditions against SSIM and default-pipeline baselines with selection error or regret as the outcome, nor an attention regulariser, which would require a training experiment, is demonstrated by the present correlations. Gaussian noise does not capture the full range of signal-dependent sensor and acquisition corruptions; work on denoising real photographs explicitly models noise and camera processing [44]. External pedestrian evaluation should also examine scale and occlusion, as characterised in established pedestrian benchmarks and CityPersons [45, 46]; both are outside the present evaluation.

7 Conclusion

APS measures spatial activation stability in a fixed detector against a clean reference image. On the evaluated pedestrian pipelines, BM3D has the highest score at high noise under the tested controls for the reference model, and that lead survives full-coverage regeneration of the learned pipelines in seven of the eight checkpoints. Across the sampled checkpoints, the association between APS and denoising gain has positive point estimates within each group, a pooled estimate near zero and intervals from pooling all noise levels that include zero, and it shifts with the measurement protocol, so unqualified interpretation across detectors remains unreliable. Preprocessing and polarity sensitivities change the comparisons below BM3D, and the pixel swap probe remains exploratory: its primary interval includes zero and it cannot separate attention from pedestrian pixels. Independent data and faithfulness checks for individual predictions are required before broader utility or mechanism claims.

References

[1] Zhou Wang et al. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[2] Ding Liu et al. "When Image Denoising Meets High-Level Vision Tasks: A Deep Learning Approach". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. 2018, pp. 842–848. DOI: 10.24963/ijcai.2018/117.

[3] Thanh Kiet Vo et al. "A Structural Preservation Framework for Denoiser Selection in YOLO-Based Pedestrian Detection under Sensor Noise". Revised manuscript under review at Scientific Reports. 2026.

[4] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[5] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information". In: *European Conference on Computer Vision (ECCV)*. 2024, pp. 1–21. DOI: 10.1007/978-3-031-72751-1_1.

[6] Yian Zhao et al. "DETRs Beat YOLOs on Real-Time Object Detection". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 16965–16974. DOI: 10.1109/CVPR52733.2024.01605.

[7] Samuel Dodge and Lina Karam. "Understanding How Image Quality Affects Deep Neural Networks". In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. 2016, pp. 1–6. DOI: 10.1109/QoMEX.2016.7498955.

[8] Dan Hendrycks and Thomas Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *International Conference on Learning Representations (ICLR)*. 2019. URL: https://openreview.net/forum?id=HJz6tiCqYm.

[9] Claudio Michaelis et al. "Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming". 2019. DOI: 10.48550/arXiv.1907.07484. arXiv: 1907.07484.

[10] Robert Geirhos et al. "ImageNet-trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness". In: *International Conference on Learning Representations (ICLR)*. 2019. URL: https://openreview.net/forum?id=Bygh9j09KX.

[11] Kostadin Dabov et al. "Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering". In: *IEEE Transactions on Image Processing* 16.8 (2007), pp. 2080–2095. DOI: 10.1109/TIP.2007.901238.

[12] Kai Zhang et al. "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising". In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155. DOI: 10.1109/TIP.2017.2662206.

[13] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[14] Aditya Chattopadhay et al. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097.

[15] Haofan Wang et al. "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 111–119. DOI: 10.1109/CVPRW50498.2020.00020.

[16] Saurabh Desai and Harish Guruprasad Ramaswamy. "Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 972–980. DOI: 10.1109/WACV45572.2020.9093360.

[17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html.

[18] Mohammed Bany Muhammad and Mohammed Yeasin. "Eigen-CAM: Class Activation Map Using Principal Components". In: *International Joint Conference on Neural Networks (IJCNN)*. 2020, pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9206626.

[19] Vitali Petsiuk et al. "Black-Box Explanation of Object Detectors via Saliency Maps". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11443–11452. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Petsiuk_Black-Box_Explanation_of_Object_Detectors_via_Saliency_Maps_CVPR_2021_paper.html.

[20] Vitali Petsiuk, Abir Das, and Kate Saenko. "RISE: Randomized Input Sampling for Explanation of Black-box Models". In: *British Machine Vision Conference (BMVC)*. 2018, pp. 1–13. URL: https://bmva-archive.org.uk/bmvc/2018/contents/papers/1064.pdf.

[21] Ruth C. Fong and Andrea Vedaldi. "Interpretable Explanations of Black Boxes by Meaningful Perturbation". In: *IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 3449–3457. DOI: 10.1109/ICCV.2017.371.

[22] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. "Understanding Deep Networks via Extremal Perturbations and Smooth Masks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 2950–2958. DOI: 10.1109/ICCV.2019.00304.

[23] Sara Hooker et al. “A Benchmark for Interpretability Methods in Deep Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019, pp. 9737–9748. URL: https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

[24] Julius Adebayo et al. “Sanity Checks for Saliency Maps”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 9505–9515. URL: https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

[25] Pieter-Jan Kindermans et al. “The (Un)reliability of Saliency Methods”. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek et al. Vol. 11700. Lecture Notes in Computer Science. Cham: Springer, 2019, pp. 267–280. DOI: 10.1007/978-3-030-28954-6_14.

[26] Amirata Ghorbani, Abubakar Abid, and James Zou. “Interpretation of Neural Networks Is Fragile”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 3681–3688. DOI: 10.1609/aaai.v33i01.33013681.

[27] Chih-Kuan Yeh et al. “On the (In)fidelity and Sensitivity of Explanations”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019, pp. 10967–10978. URL: https://proceedings.neurips.cc/paper_files/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

[28] Richard Tomsett et al. “Sanity Checks for Saliency Metrics”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 6021–6029. DOI: 10.1609/aaai.v34i04.6064.

[29] Anna Hedström et al. “Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond”. In: *Journal of Machine Learning Research* 24.34 (2023), pp. 1–11. URL: https://jmlr.org/papers/v24/22-0142.html.

[30] Simon Kornblith et al. “Similarity of Neural Network Representations Revisited”. In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. 2019, pp. 3519–3529. URL: https://proceedings.mlr.press/v97/kornblith19a.html.

[31] Rangel Daroya, Aaron Sun, and Subhransu Maji. “COSE: A Consistency-Sensitivity Metric for Saliency on Image Classification”. In: *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2023, pp. 149–158. DOI: 10.1109/ICCVW60793.2023.00022.

[32] Syamantak Sarkar et al. “Assessing the Noise Robustness of Class Activation Maps: A Framework for Reliable Model Interpretability”. In: *Image and Vision Computing* 163 (2025), p. 105717. DOI: 10.1016/j.imavis.2025.105717.

[33] Wencan Zhang, Mariella Dimiccoli, and Brian Y. Lim. “Debiased-CAM to Mitigate Image Perturbations with Faithful Visual Explanations of Machine Learning”. In: *CHI Conference on Human Factors in Computing Systems (CHI ’22)*. 2022, pp. 1–32. DOI: 10.1145/3491102.3517522.

[34] Thanh Kiet Vo et al. “Evaluating the Impact of Gaussian Noise on YOLO-Based Object Detection: A Systematic Study on Noise Robustness and Denoising-Retraining Strategies”. Under review at Engineering Applications of Artificial Intelligence; preprint posted 3 July 2026. 2026. DOI: 10.20944/preprints202607.0222.v1.

[35] Thanh Kiet Vo et al. “Selecting Denoisers for Frozen Pedestrian Detectors Under Gaussian Noise and JPEG Acquisition: Degradation Matching over Restoration Fidelity”. Research Square preprint rs-10959104, version 1. 2026. DOI: 10.21203/rs.3.rs-10959104/v1.

[36] N. J. Karthika and Saravanan Chandran. “Addressing the False Positives in Pedestrian Detection”. In: *Electronic Systems and Intelligent Computing*. Ed. by Pradeep Kumar Mallick et al. Singapore: Springer Singapore, 2020, pp. 1083–1092. DOI: 10.1007/978-981-15-7031-5_103.

[37] Gavin C. Cawley and Nicola L. C. Talbot. “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation”. In: *Journal of Machine Learning Research* 11.70 (2010), pp. 2079–2107. URL: https://www.jmlr.org/papers/v11/cawley10a.html.

[38] Ao Wang et al. “YOLOv10: Real-Time End-to-End Object Detection”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2024, pp. 107984–108011. DOI: 10.52202/079017-3429.

[39] Yunjie Tian, Qixiang Ye, and David Doermann. “YOLOv12: Attention-Centric Real-Time Object Detectors”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2025, pp. 87151–87175. DOI: 10.52202/085713-2627.

[40] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[41] Yichao Liu, Zongru Shao, and Nico Hoffmann. "Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions". 2021. DOI: 10.48550/arXiv.2112.05561. arXiv: 2112.05561.

[42] Bradley Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26. DOI: 10.1214/aos/1176344552.

[43] Edward H. Simpson. "The Interpretation of Interaction in Contingency Tables". In: *Journal of the Royal Statistical Society: Series B* 13.2 (1951), pp. 238–241. DOI: 10.1111/j.2517-6161.1951.tb00088.x.

[44] Shi Guo et al. "Toward Convolutional Blind Denoising of Real Photographs". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1712–1722. DOI: 10.1109/CVPR.2019.00181.

[45] Piotr Dollár et al. "Pedestrian Detection: An Evaluation of the State of the Art". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.4 (2012), pp. 743–761. DOI: 10.1109/TPAMI.2011.155.

[46] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. "CityPersons: A Diverse Dataset for Pedestrian Detection". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4457–4465. DOI: 10.1109/CVPR.2017.474.

Supplementary Material

This Supplementary Material accompanies the main article "Class Activation Map Stability as an Uncertain Guide to Denoising Gain for Frozen Pedestrian Detectors Under Gaussian Noise" and provides Supplementary Appendices A–G and Supplementary Tables S1–S14: (A) the complete 240-cell detection F_1 matrix across the eight frozen pedestrian detectors, six processing pipelines, and five Gaussian noise levels under the unified protocol (Table S1), from which the denoising gain (ΔF_1) follows by main-text Eq. (4); (B) reference grids and single-control pixel swap results retained under the archived protocol for continuity (Tables S2–S4); (C) implementation details, detector hook configurations, denoiser architectures, convolutional autoencoder with particle swarm optimisation (CAE+PSO) hyperparameter sidecars, and the 16-candidate checkpoint provenance screen (Tables S5–S8); (D) score-convention, leave-one-checkpoint-out, and nominal degree-of-freedom sensitivity analyses under the unified protocol (Tables S9–S11); (E) a summary of representative prior studies (Table S12); (F) measurement details, the reproducibility materials and the determinism checks; and (G) further results and diagnostics (Tables S13 and S14): the reference Attention-Preservation Score (APS) grid, the geometry controls, coverage repair pixel changes, associations, attribution, target mass and polarity, and the pixel swap diagnostics. All supplementary tables are also provided as an accompanying spreadsheet.

Abbreviations used in this Supplementary Material: AE, autoencoder; APS, Attention-Preservation Score; AUC, area under the ROC curve; BM3D, Block-Matching and 3D Filtering; CAE, convolutional autoencoder; CAE+PSO, convolutional autoencoder with particle swarm optimisation; CAM, class activation map; CI, confidence interval; COSE, COnsistency-SEnsitivity; CPU, central processing unit; CSV, comma-separated value; CUDA, Compute Unified Device Architecture; cuDNN, CUDA Deep Neural Network library; dB, decibels; df, degrees of freedom; DnCNN, Denoising Convolutional Neural Network; D-RISE, Detector Randomized Input Sampling for Explanation; EigenCAM, Eigen Class Activation Mapping; EXIF, exchangeable image file format; GAM, global attention mechanism; GPU, graphics processing unit; Grad-CAM, Gradient-weighted Class Activation Mapping; IoU, intersection-over-union; MD5, Message-Digest Algorithm 5; MSE, mean squared error; PnPLO, Person and Person-Like Objects; PSNR, peak signal-to-noise ratio; PSO, particle swarm optimisation; ReLU, rectified linear unit; ResNet, Residual Network; RGB, red–green–blue; ROAR, Remove and Retrain; ROC, receiver operating characteristic; SHA-256, Secure Hash Algorithm 256-bit; SPS, Structural Preservation Score; SSIM, structural similarity; TP, FP and FN, true positives, false positives and false negatives; XAI, explainable artificial intelligence; YAML, YAML Ain't Markup Language; YOLO, You Only Look Once.

Appendix A F_1 of Frozen Detectors Across Pipelines and Noise Levels

Table S1 lists the detection F_1 of the eight frozen detectors trained on clean images for every (pipeline, σ) cell of the evaluation split of 160 images, with inference repeated on the archived images under the unified protocol (main-text Section 4.2), with ΔF_1 from main-text Eq. (4). These recomputed counts do not reproduce the archived study cell by cell: four of the 240 cells keep identical counts of true positives, false positives and false negatives, the mean absolute F_1 difference is 0.0126 (maximum 0.0477), 131 cells (54.6%) move by more than 0.01, and 13 of the 40 rows change the bold winner, 12 of them at $\sigma \in \{1, 5, 10\}$. The archived 240-cell matrix is in the results archive deposited on Zenodo (main-text Data Availability).

Table S1: F_1 recomputed with frozen detectors in the controlled rerun on the archived images, for eight detectors, six pipelines and five noise levels. Bold marks the best pipeline in each row.

σ	Detector	Noisy	Gaussian	BM3D	Autoenc.	DnCNN	CAE+PSO
1	YOLOv8m	0.752	0.754	0.747	0.743	**0.768**	0.740
	YOLOv9m	**0.741**	0.733	0.727	0.717	0.735	0.700
	YOLOv10m	**0.751**	0.742	0.747	0.684	0.742	0.679
	YOLOv11m	0.755	0.718	**0.755**	0.706	0.751	0.695
	YOLOv12m	**0.702**	0.695	0.690	0.672	0.700	0.651
	ResNet-18	0.686	0.680	**0.698**	0.649	0.694	0.631
	ResNet-50	0.722	0.708	0.722	0.681	**0.724**	0.661
	ResNet-18+GAM	**0.697**	0.681	0.689	0.693	0.686	0.663

σ	Detector	Noisy	Gaussian	BM3D	Autoenc.	DnCNN	CAE+PSO
5	YOLOv8m	0.748	0.752	**0.767**	0.731	0.750	0.736
	YOLOv9m	0.735	0.724	0.712	0.693	**0.737**	0.698
	YOLOv10m	0.733	0.714	**0.744**	0.703	0.731	0.674
	YOLOv11m	0.728	0.697	**0.735**	0.706	0.730	0.689
	YOLOv12m	**0.710**	0.698	0.704	0.679	0.708	0.638
	ResNet-18	0.689	0.646	**0.693**	0.656	0.685	0.649
	ResNet-50	0.717	0.695	**0.724**	0.665	0.715	0.663
	ResNet-18+GAM	0.691	0.683	**0.698**	0.688	0.690	0.667
10	YOLOv8m	0.737	0.709	0.748	0.739	0.747	**0.755**
	YOLOv9m	0.733	0.667	0.709	0.699	**0.757**	0.699
	YOLOv10m	**0.753**	0.702	0.745	0.685	0.727	0.716
	YOLOv11m	0.726	0.677	**0.745**	0.668	0.698	0.712
	YOLOv12m	0.688	0.690	0.693	0.665	0.680	**0.710**
	ResNet-18	**0.687**	0.628	0.665	0.644	0.682	0.640
	ResNet-50	0.718	0.679	**0.732**	0.674	0.711	0.675
	ResNet-18+GAM	**0.705**	0.650	0.701	0.660	0.697	0.679
20	YOLOv8m	0.644	0.697	**0.757**	0.697	0.597	0.723
	YOLOv9m	0.663	0.645	**0.713**	0.624	0.603	0.684
	YOLOv10m	0.618	0.637	**0.751**	0.646	0.596	0.688
	YOLOv11m	0.615	0.620	**0.735**	0.668	0.523	0.674
	YOLOv12m	0.556	0.648	**0.717**	0.640	0.546	0.684
	ResNet-18	0.632	0.581	**0.683**	0.615	0.619	0.626
	ResNet-50	0.650	0.649	**0.718**	0.636	0.643	0.658
	ResNet-18+GAM	0.662	0.645	**0.705**	0.644	0.660	0.648
30	YOLOv8m	0.516	0.681	**0.753**	0.697	0.723	0.709
	YOLOv9m	0.533	0.607	**0.702**	0.612	0.631	0.658
	YOLOv10m	0.457	0.594	**0.731**	0.672	0.663	0.680
	YOLOv11m	0.442	0.609	**0.715**	0.677	0.687	0.667
	YOLOv12m	0.500	0.617	**0.706**	0.609	0.654	0.631
	ResNet-18	0.549	0.582	**0.673**	0.617	0.650	0.589
	ResNet-50	0.506	0.621	0.687	0.660	**0.689**	0.646
	ResNet-18+GAM	0.597	0.648	**0.695**	0.658	0.655	0.653

Appendix B Results Under the Archived Protocol

These results were produced under the archived protocol (main-text Section 4.2) on the archived images and retained for continuity with the archived study; their per-image scores are not interchangeable with the measurements under the unified protocol in main-text Section 5.

Table S2: APS_{corr} by processing pipeline and noise level under the archived protocol (reference YOLOv9m, archived images, n=160). Signed score; bold marks the best pipeline at each level.

Pipeline	σ=1	σ=5	σ=10	σ=20	σ=30
Noisy	**0.938**	**0.835**	0.626	0.469	0.326
Gaussian filter	0.765	0.745	0.584	0.437	0.414
BM3D	0.865	0.830	**0.800**	**0.716**	**0.665**
Autoencoder	0.530	0.539	0.437	0.406	0.350
DnCNN	0.662	0.633	0.516	0.334	0.365
CAE+PSO	0.488	0.493	0.518	0.427	0.359

Under the archived protocol (Table S2) the reference detector's BM3D score at $\sigma = 30$ is 0.665, the highest in that column. Adjusted for σ, APS and denoising gain are positively associated at that checkpoint ($r = 0.584$, 95% bootstrap CI over images $[0.117, 0.763]$, nominal $p = 0.0055$ with residual degrees of freedom); at $\sigma = 30$ alone the five denoising pipelines give $r = 0.74$ on five observations.

Table S3: Adjusted APS–ΔF_1 correlations under the archived protocol, with pointwise 95% intervals from 5,000 joint bootstrap replicates over images. Analyses across all noise levels remove means within each method and noise level; $\sigma = 30$ analyses remove method means within each reported group.

Checkpoint group	All noise levels	$\sigma = 30$
YOLO variants	$+0.16$ $[-0.17, +0.37]$	$+0.32$ $[-0.19, +0.64]$
ResNet variants	$+0.48$ $[-0.16, +0.66]$	$+0.77$ $[-0.17, +0.93]$
All eight checkpoints	-0.21 $[-0.35, +0.04]$	-0.35 $[-0.54, -0.06]$

The unified protocol does not reproduce the pooled reversal in Table S3: on the same archived images the pooled estimate moves from -0.205 to $+0.017$ over all noise levels and from -0.348 to -0.058 at $\sigma = 30$ (main-text Table 2). The only interval in Table S3 that excludes zero is the pooled $\sigma = 30$ row. The sensitivities to sign convention and checkpoint omission under the archived protocol in the deposited results archive gave positive point estimates within each family, except the YOLO estimate omitting YOLOv9m (-0.04), a negative pooled estimate under all three score conventions, and no interval pooling all noise levels that excludes zero. Over the 40 detector–denoiser cells at $\sigma = 30$ the unadjusted correlation under the archived protocol with denoising gain is approximately -0.08 for APS and $+0.33$ for SSIM; under the unified protocol APS gives $+0.096$ on the archived images and $+0.380$ on repaired coverage (learned pipelines only), while SSIM is stored only for the archived images, so the two measures cannot be compared on repaired coverage and are not compared under the unified protocol.

Table S4: Pixel swap rerun under the archived protocol with one control location at $\sigma = 30$ (reference YOLOv9m, archived images). Micro-F_1 intervals resample whole images, conditional on one random control location per image.

Condition	Micro-F_1 [95% CI]	Recall
BM3D	0.695 [0.644, 0.743]	0.680
Autoencoder (AE)	0.624 [0.563, 0.684]	0.571
AE + high attention graft	0.645 [0.592, 0.698]	0.635
AE + random graft	0.613 [0.549, 0.675]	0.576
BM3D + reverse graft	0.676 [0.618, 0.730]	0.621

Table S4 lists that rerun's arms, whose paired high-minus-random difference is $+0.0324$ $[-0.0258, +0.0882]$ and high-minus-autoencoder difference $+0.0213$ $[-0.0290, +0.0724]$, with 95% bootstrap intervals over images; neither comparison provides a precise positive effect estimate.

The archived swap run's exploratory recovery diagnostic uses the looser any-IoU coverage rule of Appendix F.1. Its 31-instance subset comprises boxes covered under BM3D but not the autoencoder. The high graft recovers 16 of them and the random graft six, with contingency counts of one recovered by both, 15 by high only, five by random only and ten by neither. An instance-level exact McNemar test gives $p = 0.0414$, but the 31 instances occupy only 24 images and joint image-level label exchange gives $p = 0.0990$, so we do not interpret that test as a robust demonstration of an intervention effect. The artifact from missing content further limits interpretation: all 31 selected instances occur in images with border strips; 13 boxes intersect the recorded strip region, including eight of the 16 recoveries by the high graft and two of the six random recoveries. This does not prove that every recovery repairs a missing region, because box intersection alone does not establish where the graft acted; it does show that this archived design cannot separate content restoration guided by attention from incomplete patch coverage.

Appendix G.6 reports a one-to-one analogue of this diagnostic under a different matching rule and on a different instance cohort, so the two are separate estimands and are not compared numerically.

Appendix C Implementation and Provenance of the Archived Experiment

This appendix documents the evaluation units and image preprocessing (Appendix C.1), the detector configurations and hooked feature stages (Appendix C.2, Table S5), the noise generation and denoiser implementations (Appendix C.3, Tables S6 and S7), additional metric conventions (Appendix C.4), and software and checkpoint provenance (Appendix C.5, Table S8).

Appendix C.1 Evaluation Units and Image Preprocessing

The archived analysis uses the 160 evaluation stems of the PnPLO validation directory (main-text Section 4.1). The clean directories hold 944 training, 160 validation and 235 test images, and of the two configured classes, person and person-like, it retains only class 0 (person). Checkpoint selection and the archived CAE+PSO hyperparameter search both used validation images. These results are therefore not an evaluation on a test set held out from all model selection decisions.

Clean and processed files are paired by filename stem. For APS, Pillow converts each file to red–green–blue (RGB) colour space and resizes it directly to 640×640, not preserving aspect ratio, then divides the pixel values represented as floating point by 255. Detection instead uses the Ultralytics prediction interface with `imgsz=640` and its own input preparation, so the two do not establish identical preprocessing geometry. Image quality scores are evaluated at the clean image's original dimensions, to which the processed image is resized; PSNR and SSIM use a data range of 255 and SSIM the colour channel axis. That preprocessing is the archived protocol; the unified protocol instead hooks every sweep row's CAM inside the `predict` forward pass that produced its detections, with `rect=False` and image size 640 (main-text Section 4.2). On the identical archived images, scores for each image under the archived and unified protocols correlate at $r = 0.1852$ and differ by a mean absolute error of 0.3389 over 38,400 paired rows.

Appendix C.2 Detector Configurations and Selected Feature Stages

Table S5 lists the eight configurations. The three custom ones combine residual stages with a multiscale neck and detection head based on YOLOv8, their YAML definitions using basic blocks with stage depths $(2, 2, 2, 2)$ for ResNet-18 and bottleneck blocks with depths $(3, 4, 6, 3)$ for ResNet-50; the GAM variant adds channel–spatial gating after neck feature blocks, including the selected final feature stage.

Table S5: Configurations, CAM hook points and hooked tensor shapes for the eight detectors. Indices are zero-based within the detector's top-level module sequence, and shapes are given as channels × height × width.

Configuration	Index	Selected module	Detection head	CAM tensor
YOLOv8m	21	`C2f`	Native YOLOv8	$576 \times 20 \times 20$
YOLOv9m	21	`RepNCSPELAN4`	Native YOLOv9	$480 \times 20 \times 20$
YOLOv10m	22	`C2fCIB`	Native YOLOv10	$576 \times 20 \times 20$
YOLOv11m	22	`C3k2`	Native YOLOv11	$512 \times 20 \times 20$
YOLOv12m	20	`C3k2`	Native YOLOv12	$512 \times 20 \times 20$
YOLOv8–ResNet-18	17	`C2f`	`Detect`	$512 \times 40 \times 40$
YOLOv8–ResNet-50	17	`C2f`	`Detect`	$2048 \times 20 \times 20$
YOLOv8–ResNet-18+GAM	21	`GAM_Attention`	`Detect`	$512 \times 40 \times 40$

The probe selects the last module before the detection head that returns one feature tensor with four dimensions: index `[-2]` for the five stock probes, and for the custom models the positive indices the earlier successful probe records in Table S5. "Native" in that table denotes the stock model's detection head, whereas the custom models use the multiscale `Detect` head declared in their YAML. APS correlation is calculated on each hook's native spatial grid before any display upsampling. The table also reports the hooked tensor shapes the rerun recorded: all APS inputs are 640×640, yet the hooked grids differ, the custom stage definitions permitting different downsampling schedules. Equal input size and a common position before the

head therefore do not establish a common feature resolution, a comparability caveat for the group of ResNet variants, which spans two feature resolutions internally. Grid size alone does not order the observed scores: the ResNet configuration with the highest mean APS is one of the 20×20 configurations, and configurations sharing a grid size occupy both extremes of the observed range. Its contribution to the group differences was not isolated.

Detector training specification and its limits. The archived training routine specifies at most 500 epochs, early stopping with patience 50, input size 640, seed 0, deterministic training, automatic optimiser selection, nominal initial learning rate setting `lr0`=0.01 and final learning rate fraction `lrf`=0.01, weight decay 5×10^{-4}, three warmup epochs, and mixed precision. Stock configurations use supplied weights; custom networks attempt to transfer compatible pretrained backbone tensors before detection training. Augmentation includes horizontal flips (probability 0.5), mosaic (1.0, disabled for the last ten epochs), translation 0.1 and scale 0.5, and available training argument files record batches of 64, 48, 64, 64, 48, 48, 24 and 48 in the model order of Table S5; these describe available runs, not the executed epoch count. All reported pipeline evaluations reuse checkpoints trained on clean images without fine tuning for individual pipelines.

Appendix C.3 Noise Generation and Denoising Implementations

The upstream code adds independent Gaussian samples to the image's three channels of 8-bit pixels, clips to $[0, 255]$ and casts back to 8-bit unsigned integers, so $\sigma \in \{1, 5, 10, 20, 30\}$ is on the 0–255 intensity scale. Its routine that generates the dataset seeds NumPy with the numeric noise level before iterating over images, but no manifest records the file iteration order, source version and random number state, so the seeding rule alone is not a claim of exact replay. Images are written through OpenCV with their existing extensions and default encoder settings, so encoding can add a step to the nominal Gaussian corruption. Table S6 describes the evaluated implementations, not all methods bearing the same names.

Table S6: Processing implementations reconstructed from the archived upstream code.

Pipeline	Implementation
Noisy	Stored Gaussian-corrupted image, without a denoising stage.
Gaussian filter	`cv2.GaussianBlur` with kernel widths 3, 5, 7, 9, 9 at the five noise levels; `sigmaX=0` lets OpenCV derive the filter width from the kernel.
BM3D	RGB input scaled to $[0, 1]$; `bm3d.bm3d_rgb` with `sigma_psd`= $\max(\sigma/255, 0.01)$, followed by clipping and conversion to an 8-bit image. Other options use the library defaults.
Autoencoder	Four encoder convolutions with 64, 64, 128, 128 channels, 3×3 kernels and two 2×2 max-pooling stages; a mirrored decoder with two transposed convolutions and a sigmoid RGB output.
DnCNN	A 17-layer residual noise estimator with 64 internal channels, 3×3 convolutions and intermediate batch normalisation and rectified linear unit (ReLU) activations; the network's estimated noise is subtracted from its RGB input.
CAE+PSO	An encoder–decoder with two pooling stages, batch normalisation, channel widths $f, f, 2f, 2f$ and a sigmoid output. PSO selects the learning rate, base width f and odd kernel size using validation reconstruction MSE.

Learned denoiser training. Each learned method has five saved networks, one per noise level, trained on random 128×128 RGB crops of clean images (smaller images resized), with clipped Gaussian noise at that level and inputs and targets scaled to $[0, 1]$. The autoencoder and DnCNN routines use four patches per training image per epoch, Adam at learning rate 10^{-3}, reconstruction MSE, a plateau scheduler (patience 5, factor 0.5) and 50 requested epochs. CAE+PSO uses two patches per image, six particles, ten iterations and eight training epochs per fitness evaluation, with search bounds $[10^{-4}, 5 \times 10^{-3}]$ for learning rate, $[32, 96]$ for base width and $[3, 7]$ for the kernel parameter, converted to an odd integer; inertia decreases from 0.9 to 0.4 and both PSO acceleration coefficients are 2. A new network is then trained for 50 requested epochs with the selected settings, which the retained CAE sidecars record in Table S7, at a batch size taken from graphics processing unit (GPU) memory rather than fixed globally. Existing checkpoint files are loaded without

training when present, so these rules do not establish each saved weight file's training history; training logs for each checkpoint and denoiser random number generator states are unavailable in the analysis bundle.

Table S7: CAE+PSO settings in the retained parameter sidecars. Learning rates are rounded here; the sidecars retain full precision.

σ	Base width f	Kernel width	Learning rate
1	80	3	9.13364×10^{-4}
5	45	3	1.62477×10^{-3}
10	69	3	3.95750×10^{-4}
20	38	3	8.22488×10^{-4}
30	96	3	5.53662×10^{-4}

Sliding window coverage. The archived inference function for learned denoisers uses 128×128 patches with stride 112, reflection padding to multiples of 128, overlap averaging, clipping and conversion back to 8-bit RGB. Its original loops over patch starting positions stop at the last position aligned with the stride without appending a patch at the far image boundary, so when the padded dimension and stride are incompatible some bottom or right pixels receive no predictions, and the normalisation clamps their count to at least one, leaving them at zero. Main-text Section 5.2 reports the regeneration of the three learned pipelines with the stored weights at complete coverage and the resulting scores.

Appendix C.4 Additional Metric Conventions

EigenCAM uses a forward hook and PyTorch's singular value decomposition (`torch.linalg.svd`), with a lower bound of 10^{-8} on the min–max normalisation denominator; APS correlation is set to zero when the product of centred map norms is no larger than 10^{-12}, and the robustness check based on gradients differentiates a summed output for the person class manually; the principal component computation does not require the `pytorch-grad-cam` package. For the historical summary of attention on target, the union of person boxes in main-text Eq. (3) is rasterised on the native CAM grid by truncating coordinates to integers and clipping them to the grid; empty person masks and maps with negligible total mass receive zero. The reported summary is the arithmetic mean over all 160 processed image stems, not the ratio of masses pooled across images, and implies no baseline normalised by box area or based on random locations.

The historical receiver operating characteristic (ROC) analysis contains 22,320 rows for combinations of image and condition, the 93 images containing people crossed with eight detectors, six pipelines and five noise levels. Its binary label is

$$y_i = \mathbf{1}\left\{ \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i + 10^{-9}} \geq 0.5 \right\}. \tag{S1}$$

For integer detection counts this labels strictly more than half of an image's persons as recovered, with the small denominator offset excluding exact 50% recall. False positives do not enter the label and negative images are excluded, so the ROC describes separation of this recall label on repeated observations across images and conditions, not a test of overall detection success on independent images (Appendix G.5).

Appendix C.5 Software and Checkpoint Provenance

The archived explainable artificial intelligence (XAI) notebook installs packages without version pins; one retained session output reports Python 3.12 and Ultralytics 8.4.60 alongside PyTorch 2.11.0+cu128, NumPy 2.0.2, OpenCV 4.13.0.92, Pillow 11.3.0 and SciPy 1.16.3; these are session observations, not a verified environment for every cached result. The notebook can reuse existing comma-separated value (CSV) rows; its later session could not import the custom ResNet module and reused cached results from the earlier notebook, whose probes of the custom models succeeded.

The original checkpoint paths follow `training_runs/<model>_noisy_noise_0/weights/best.pt`; the caches bind no row to a checkpoint checksum, manifest of input files and environment identifier, and

multiple snapshots exist under the same model names, which the controlled rerun resolves. In these paths, `noisy_noise_0` names the training run at noise level $\sigma = 0$, for which the upstream code adds no noise and passes the original clean dataset to training, and the retained training arguments of the six selected current checkpoints name that dataset; these are the detectors trained on clean images used throughout. The `ver_7` candidates whose file names contain `clean` (Table S8) are separate training runs. A screen of detection counts for each image in two conditions, noisy and BM3D at $\sigma = 30$ over the 160 evaluation images, scored every declared candidate, giving 320 comparisons each. Exactly one candidate per detector reproduces all 320, so the run required no manual acknowledgement of a new checkpoint set (`needs_ack` false), and Table S8 lists all 16 candidates, their matched counts, selected status and abbreviated paths. YOLOv12m, like YOLOv9m, resolves to the March `ver_5` file rather than to the file under its current training run directory. Full checksums and run identity are retained in the provenance records of the deposited results archive. Three limits apply. The screen compares detection counts, so it is not proof that two files hold identical tensors; it runs the prediction interface under the default letterbox geometry rather than `rect=False`, a difference of up to 1.151 pixels in box coordinates; and YOLOv9m's single declared candidate left its match uncontested. Numerical reproduction of figures from the retained CSV files remains a separate question.

Table S8: Checkpoint candidate screen for the controlled rerun. Matched counts are agreements with the archived detection counts for each image over the 320 screen conditions; candidate paths are abbreviated relative to the run manifest.

Detector	Candidate	File	Matched	Selected
YOLOv8m	current	`yolov8m_noisy_noise_0/best.pt`	320/320	✓
YOLOv8m	ver7	`ver_7/yolov8m_clean_best.pt`	193/320	
YOLOv9m	march	`ver_5/yolov9m_noisy_noise_0_best.pt`	320/320	✓
YOLOv10m	current	`yolov10m_noisy_noise_0/best.pt`	320/320	✓
YOLOv10m	ver7	`ver_7/yolov10m_clean_best.pt`	204/320	
YOLOv11m	current	`yolo11m_noisy_noise_0/best.pt`	320/320	✓
YOLOv11m	ver7	`ver_7/yolo11m_clean_best.pt`	206/320	
YOLOv12m	current	`yolo12m_noisy_noise_0/best.pt`	179/320	
YOLOv12m	ver7	`ver_7/yolo12m_clean_best.pt`	191/320	
YOLOv12m	march	`ver_5/yolo12m_noisy_noise_0_best.pt`	320/320	✓
ResNet-18	current	`yolov8-resnet18_noisy_noise_0/best.pt`	320/320	✓
ResNet-18	ver7	`ver_7/yolov8-resnet18_clean_best.pt`	204/320	
ResNet-50	current	`yolov8-resnet50_noisy_noise_0/best.pt`	320/320	✓
ResNet-50	ver7	`ver_7/yolov8-resnet50_clean_best.pt`	185/320	
ResNet-18+GAM	current	`yolov8-resnet18-gam_noisy_noise_0/best.pt`	320/320	✓
ResNet-18+GAM	ver7	`ver_7/yolov8-resnet18-gam_clean_best.pt`	216/320	

The separate rerun with the reference checkpoint on 12 September 2026 uses the March YOLOv9m file `yolov9m_noisy_noise_0_best.pt` for its pixel swaps and geometry controls. At $\sigma = 30$, it reproduces the archived BM3D counts $(\mathrm{TP}, \mathrm{FP}, \mathrm{FN}) = (138, 56, 65)$ and autoencoder counts $(116, 53, 87)$ exactly. Its provenance records an NVIDIA A100 40-GB GPU, Python 3.13.15, Ultralytics 8.4.60, PyTorch 2.11.0+cu128 and NumPy 2.1.3; these observations are distinct from the original notebook session above.

The controlled rerun of 19 September 2026 ran on an NVIDIA A100-SXM4-40GB GPU with Python 3.13.15, PyTorch 2.11.0+cu128, torchvision 0.26.0+cu128, Ultralytics 8.4.60, CUDA 12.8, cuDNN 91900, NumPy 2.1.3, pandas 2.2.3, SciPy 1.16.3, Pillow 11.3.0 and OpenCV 4.12.0.88, on the evaluation split under the detection and bootstrap settings of main-text Sections 4.2 and 4.4, with `common_crop_aps` false. The manifest fingerprints 5,161 input files, and the run reports 10,453 seconds of compute against about 3 hours 58 minutes of elapsed time including the fingerprint preflight. The recorded revision metadata identifies the analysis kit as version 1.1; the internal `protocol_version` string still reads `2026-09-14-v1`. The provenance records of the deposited results archive retain the complete checksums of the notebook and declaration, together with the canonical run identity.

A central processing unit (CPU) check conducted by the authors outside that rerun compared regenerated and archived outputs for nine stored denoiser checkpoints: the three learned methods at $\sigma \in \{10, 20, 30\}$. It used 10 evaluation images and only the region covered by the original patch grid. The stored `dncnn_sigma30.pt`

improves PSNR over the noisy input by 0.6 dB, from 22.06 to 22.67 dB, whereas the archived DnCNN $\sigma = 30$ images reach 27.19 dB, a gain of 5.1 dB. The other tested checkpoints give PSNR against the clean reference within about 1 dB of the archived outputs: autoencoder ($\sigma = 30$): 25.02 against 24.23 dB, CAE+PSO ($\sigma = 30$): 24.63 against 25.35 dB, DnCNN ($\sigma = 20$): 25.52 against 25.48 dB, and DnCNN ($\sigma = 10$): 30.21 against 30.32 dB, with archived values listed first. Similar PSNR does not establish pixelwise reproduction. The stored DnCNN $\sigma = 30$ weights under the checked regeneration path are inconsistent with the archived outputs; continuity of the generating weights and pipeline remains unverified. This is a diagnostic on 10 images rather than a benchmark. The regenerated DnCNN $\sigma = 30$ cells are excluded from comparisons that assume identical archived and regenerated weights.

In the rerun of 12 September 2026, the geometry control marks a trailing row or column black when its maximum pixelwise mean RGB intensity is at most 8 on the 8-bit scale, and flags bottom or right strips wider than two pixels; this rule is the estimator behind the flagged images of main-text Section 4.3. For each image and noise level, the common valid region is the intersection of the six pipelines' valid regions, obtained by removing the maximum detected height of the bottom strip and width of the right strip; it is the full frame if none is flagged. Masked and cropped scores are computed inside it. Image `image (91)`, which has no annotated person, is excluded from the matched geometry comparison for incompatible stored EXIF geometry; the original swap experiment retains all 160 images. The same rerun's swap mask thresholds the bilinearly upsampled clean CAM at the value selecting its top 20% of pixels, ties retained. Its random control seeds a NumPy generator for each image with the first eight hexadecimal digits of the filename stem's Message-Digest Algorithm 5 (MD5) digest and draws cyclic offsets between one-quarter and three-quarters of each image dimension, trying up to 20 locations for an overlap of at most 25% of mask area. Although the code has a fallback after the final trial, all 160 retained random masks satisfy the overlap criterion. These details identify the saved probe at a single location alone; main-text Section 5.4 reports the controlled rerun that replaces it, applying the graft to images from learned denoisers regenerated at complete coverage (main-text Section 5.2).

Appendix D Additional Statistical Sensitivities

These tables repeat the association analysis of main-text Section 5.3 under the unified protocol on the archived images. Two cells exclude zero, the $|r|$ convention in the YOLO group (Table S9) and the leave-one-checkpoint-out analysis omitting YOLOv11m (Table S10); neither is promoted to the reported estimate (main-text Section 4.4). Changing the score convention neither regenerates nor validates the saliency masks, the leave-one-checkpoint-out intervals describe image resampling only, and the nominal p-values of Table S11 are exploratory diagnostics, not primary uncertainty estimates.

Table S9: Adjusted APS–ΔF_1 correlations pooling all noise levels under three scalar score conventions, computed under the unified protocol on the archived images with joint bootstrap intervals over images.

Convention	YOLO variants	ResNet variants	Pooled
Signed r	$+0.236$ $[-0.043, +0.413]$	$+0.461$ $[-0.027, +0.614]$	$+0.017$ $[-0.145, +0.166]$
$\lvert r \rvert$	$+0.351$ $[+0.053, +0.485]$	$+0.468$ $[-0.002, +0.646]$	$+0.106$ $[-0.071, +0.240]$
$\max(r, 0)$	$+0.281$ $[-0.002, +0.438]$	$+0.468$ $[-0.012, +0.634]$	$+0.047$ $[-0.118, +0.188]$

Table S10: Leave-one-checkpoint-out association sensitivity under the unified protocol on the archived images.

Group	Omitted checkpoint	Residual r [95% CI]
YOLO variants	YOLOv8m	0.00 $[-0.26, +0.24]$
YOLO variants	YOLOv9m	$+0.16$ $[-0.18, +0.41]$
YOLO variants	YOLOv10m	$+0.28$ $[-0.04, +0.48]$
YOLO variants	YOLOv11m	$+0.39$ $[+0.02, +0.57]$
YOLO variants	YOLOv12m	$+0.31$ $[-0.04, +0.52]$
ResNet variants	ResNet-18	$+0.41$ $[-0.20, +0.68]$
ResNet variants	ResNet-50	$+0.34$ $[-0.47, +0.66]$
ResNet variants	ResNet-18+GAM	$+0.58$ $[-0.07, +0.78]$

Table S11: Nominal independent-cell tests with corrected residual degrees of freedom, under the unified protocol on the archived images.

Group	N cells	Residual df	Nominal p
YOLO variants	125	99	0.01751
ResNet variants	75	49	0.00067
All eight checkpoints	200	174	0.82406

Appendix E Summary of Representative Prior Studies

Table S12 summarises representative prior studies relative to the present study (main-text Section 2). Bracketed numbers refer to the reference list of the main article.

Table S12: Summary of representative prior studies on image restoration for vision tasks, spatial attribution reliability, and structural preservation relative to the present study.

Study / Representative works	Domain and task	Restoration or attribution focus	Scope relative to the present study
Dabov et al. [11]; Zhang et al. [12]	Natural-image denoising	Nonlocal collaborative filtering (BM3D) and residual convolutional denoising (DnCNN)	Optimise and evaluate image-space fidelity (PSNR and SSIM) without consulting frozen detector activations.
Dodge and Karam [7]; Hendrycks and Dietterich [8]; Michaelis et al. [9]	Image classification and object detection under corruption	Recognition degradation under noise, blur, compression, and benchmark corruptions	Quantify task performance loss under corruption but do not pair denoiser selection with internal spatial attention stability.
Liu et al. [2]	High-level vision under degradation	Denoiser trained with a joint reconstruction and task loss through a fixed recognition network	Adapts the denoiser to one recognition network during training rather than evaluating separately trained pipelines before frozen detectors.
Selvaraju et al. [13]; Chattopadhay et al. [14]; Muhammad and Yeasin [18]; Petsiuk et al. [19]	Visual explanation for classification and object detection	Gradient-weighted CAMs (Grad-CAM, Grad-CAM++), principal-component CAM (EigenCAM), and detector mask sampling (D-RISE)	Define spatial attribution maps for individual inputs but do not test whether clean-reference map correlation tracks the detection gain from denoising.
Adebayo et al. [24]; Hooker et al. [23]; Yeh et al. [27]; Hedström et al. [29]	Saliency evaluation and sanity checks	Parameter randomisation, feature removal and retraining (ROAR), infidelity/sensitivity, and Quantus metrics	Evaluate attribution reliability for classifiers, partly through retraining (ROAR), rather than fixed-weight object detectors across restoration pipelines.
Daroya et al. [31]; Sarkar et al. [32]; Zhang et al. [33]	Saliency consistency across augmentations, architectures and image perturbations	Consistency and sensitivity under data augmentation (COSE); CAM robustness under input noise; correlation with the clean-image map, raised by auxiliary tasks (Debiased-CAM)	Classification and captioning studies; Sarkar et al. report that stable EigenCAM maps can respond weakly to changed predictions, motivating separate stability and faithfulness checks.
Vo et al. [3, 34, 35]	Pedestrian detection under image degradation	Image-space Structural Preservation Score (SPS), training regimes, and denoiser selection	Evaluate image-domain metrics or detector regimes without internal activation preservation or controlled pixel swap interventions.
Present study	Pedestrian detection under Gaussian noise with frozen detectors	Unified-protocol APS, full-coverage denoiser regeneration, and 20-draw pixel swaps guided by attention	Evaluates eight frozen YOLO and ResNet checkpoints across six pipelines and five noise levels with explicit geometry, polarity, and bootstrap controls.

Appendix F Measurement and Reproducibility Details

This appendix gives measurement and reproducibility details summarised in main-text Sections 4.2–4.4 and in its Data Availability statement.

Appendix F.1 Recovery Rule of the Archived Diagnostic

The archived recovery diagnostic uses a third, looser rule: a ground-truth instance is "covered" when any prediction overlaps it at IoU at least 0.5, so one prediction can cover more than one instance; its rerun analogue counts an instance as caught only under the one-to-one assignment. The two are different estimands on different cohorts, kept distinct throughout; neither can be reconstructed as an identical set of true positives from aggregate F_1 alone.

Appendix F.2 Verification of the Archived Measurement Path

The reference checkpoint rerun of 12 September 2026 verified that checkpoint against the original baselines: over the 4,800 matched original APS values the median absolute difference was about 3.7×10^{-5}, 98.5%

were below 10^{-3} and the maximum was about 7.1×10^{-3}, with no sign changes. This documents numerical agreement without claiming every row met the 10^{-3} threshold. The later controlled rerun repeated that archived path on all 160 evaluation stems as a canary, matching the archived per-image values to within 2.07×10^{-4}. Both checks concern the archived path, and neither shows that the unified protocol reproduces archived APS.

Appendix F.3 Exploratory Sign-Flip Test

A separate exploratory statistic averages the difference for each image between the high graft F_1 and the mean F_1 of random grafts. We apply 20,000 random sign assignments to its nonzero image differences with seed 1, under the working null of independent, sign-symmetric paired image differences. This tests a different estimand from the primary micro-F_1 contrast: images without persons have zero per-image F_1, whereas their false positives contribute to the primary pooled counts. It is not a prospectively randomised trial or a replacement primary test.

Appendix F.4 Reproducibility Materials

The reproducibility materials are deposited, as three archives to be extracted side by side, in the Zenodo record that distributes this preprint (its DOI is shown on the record's landing page). The results archive contains the per-image numerical outputs, the configuration records, the derived results and the generated tables behind the reported statistical estimates, including the archived-protocol 240-cell detection matrix and sensitivity tables not reprinted here, and the complete checkpoint provenance records; it excludes the crop accounting export `crop_gt_accounting.csv`, whose ground-truth exclusion policy is applied nowhere in the reported analysis. The code archive contains the analysis source, with a single command that reproduces and verifies the statistical estimates and tables in a separate directory, the rendering code for all eight figures (main-text Figures 1–8), the denoiser weight check summarised in Supplementary Appendix C, and the LaTeX source of the preprint deposited in the same record. The figures archive contains these eight figures and, for the two selected example images, their saved detections, maps from clean images, verified masks and selection record, but not, as separate files, the photographs or the images made from them, without which main-text Figures 5 and 8 cannot be regenerated. Reproducing the analyses and figures requires no neural inference. The reported statistical estimates, the generated tables and the data plots were computed from stored experimental outputs with this code, apart from the diagnostics and summaries named below. The authors validated them with checkpoint and input checksums, matching and reconstruction tests, and an independently written reanalysis of detection counts and bootstrap intervals. The diagnostics and summaries computed outside this code, and not recomputed by it, include the paired Wilcoxon tests of Appendix G.2, the McNemar test, strip counts and SSIM correlation of Appendix B, the target-mass means and ROC summary of Appendix G.5, the agreement and determinism checks of Appendix F.2 and Appendix F.5, the letterbox difference of Appendix C.5, and the strip-detector statistics and duplicate check of main-text Sections 4.3 and 4.1. The supplementary tables are also supplied as a spreadsheet with the Supplementary Material of this article. The full dataset and trained checkpoints are not redistributed. The inference outputs of the controlled rerun were produced by experiment kit v1.1, which is not included in the archives apart from its core protocol module. The provenance records in the results archive identify its notebook and checkpoint declaration by complete checksums; the internal protocol version string still records the earlier value and is not the basis for this identification. Numerical reproduction uses the deposited exports. Re-executing inference additionally requires the identified detector and denoiser weights and processed image files, whose complete checksums are retained in the provenance records described in Appendix C. These model files are not redistributed and no public checkpoint archive is available; requests concerning access to the identified files can be directed to the corresponding authors, subject to applicable permissions.

Appendix F.5 Determinism

Determinism was measured rather than assumed. Two independent sessions with the same run signature, each on an NVIDIA A100 graphics processing unit (GPU), recomputed 878 of 878 sampled unit files with identical

bytes and complete filename containment across all seven unit sections, and produced identity, geometry, selection and screen exports with identical bytes; each unit filename is the Secure Hash Algorithm 256-bit (SHA-256) digest of its canonical key. The runner disables TensorFloat-32 arithmetic for matrix multiplication and Compute Unified Device Architecture Deep Neural Network library (cuDNN) operations and sets the cuDNN deterministic flag. Three limits apply. The run signature excludes the stage list, the resume directory and the field acknowledging the checkpoints, so it does not certify which stages were executed, and the identity record does not itself store the determinism switches. The smallest matched IoU in the run is 0.500000192, so one accepted true positive sits just above the matching threshold. Reproduction with identical bits is therefore demonstrated across two sessions with identical software, not across environments.

Appendix G Further Results and Diagnostics

This appendix reports results and diagnostics summarised in main-text Section 5. Main-text section, table, figure and equation numbers are marked as such; all other numbers refer to this Supplementary Material.

Appendix G.1 Reference Grid and Protocol Comparison

Table S13 lists the reference checkpoint APS grid plotted in main-text Figure 1. On the same archived images, all 30 cells are higher under the unified than the archived protocol, by +0.128 on average, while the best pipeline at each noise level is unchanged in 5 of 5 cases; Table S2 retains the grid under the archived protocol. The two protocols' scores are nevertheless not interchangeable per image: over the 38,400 paired image scores their Pearson correlation is 0.185 and their mean absolute difference 0.339, so aggregate orderings carry across where individual scores do not.

Table S13: APS_{corr} by processing pipeline and noise level for the reference YOLOv9m checkpoint, unified protocol on the archived images (n=160). Signed score; bold marks the best pipeline at each level.

Pipeline	σ=1	σ=5	σ=10	σ=20	σ=30
Noisy	**0.983**	**0.930**	0.802	0.569	0.455
Gaussian filter	0.879	0.852	0.759	0.647	0.530
BM3D	0.945	0.917	**0.891**	**0.843**	**0.782**
Autoencoder	0.666	0.640	0.611	0.559	0.551
DnCNN	0.755	0.752	0.634	0.505	0.512
CAE+PSO	0.622	0.600	0.614	0.573	0.538

Appendix G.2 Geometry Controls

The matched geometry controls of main-text Section 4.3 use the archived protocol on the reference checkpoint only, and belong with Table S2 rather than Table S13. They preserve BM3D's leading mean at $\sigma = 30$ under original, masked and cropped evaluation, but change the comparison between Gaussian filtering and the learned pipelines, whose cropped means exceed the mean for Gaussian filtering, unlike their original matched means (Table S14, main-text Figure 2). The ranking below BM3D is therefore sensitive to input geometry, whereas the BM3D result on the reference model is more stable. Even under the original matched evaluation, paired Wilcoxon comparisons do not clearly distinguish Gaussian filtering from the autoencoder, DnCNN or CAE+PSO ($p = 0.762$, 0.804 and 0.736). Cropping changes the detector input, so this contrast cannot separate artifact removal from changes in context, scale and map polarity, nor measure denoisers at complete reconstruction coverage; main-text Section 5.2 reports that measurement. Paired method comparisons in this geometry analysis are exploratory.

Table S14: Reference signed APS at $\sigma = 30$ on 159 matched images under original, masked and cropped evaluation, archived protocol and reference checkpoint only. Brackets contain pointwise 95% bootstrap intervals over images.

Pipeline	Original	Masked	Cropped
Noisy	0.326 [0.230, 0.421]	0.325 [0.231, 0.415]	0.310 [0.217, 0.400]
Gaussian	0.416 [0.322, 0.504]	0.414 [0.323, 0.502]	0.314 [0.214, 0.410]
BM3D	0.667 [0.572, 0.751]	0.669 [0.574, 0.753]	0.661 [0.570, 0.743]
Autoencoder	0.350 [0.249, 0.446]	0.357 [0.257, 0.453]	0.448 [0.345, 0.549]
DnCNN	0.365 [0.265, 0.459]	0.375 [0.276, 0.469]	0.466 [0.373, 0.557]
CAE+PSO	0.359 [0.258, 0.454]	0.368 [0.269, 0.462]	0.435 [0.331, 0.534]

Appendix G.3 Coverage Repair

Relative to the images regenerated on the original patch grid, the repaired grid (main-text Section 5.2) changes pixels already covered by a mean absolute error of 0.0632 over all 2,400 cells on the 8-bit scale, at most 0.2783 in one cell, against 110.80 in the previously uncovered region of the 1,905 affected cells, and the 495 already fully covered cells still change by 0.0280 on average through the altered overlap-add reweighting. Main-text Figure 4 separates the five noise levels of the repair effect on APS and F_1.

Appendix G.4 Associations Within and Across Detector Groups

Main-text Figure 6 plots the estimates of main-text Table 2. At $\sigma = 30$, where the adjustment removes pipeline means within each reported group, the single interval excluding zero is the ResNet group on the archived images, +0.81 ([0.18, 0.93]; main-text Figure 6b). That cell rests on three checkpoints spanning two CAM resolutions, with the twelve reported estimates sharing one set of resample indices and no multiplicity adjustment (main-text Section 4.4), and is reported for completeness, not as a finding. The DnCNN $\sigma = 30$ cells have limited provenance in both image sets (main-text Section 5.2). None of these intervals establishes inference over a population of architectures.

At $\sigma = 30$ the group offsets producing the reversal under the archived protocol shrink under the unified protocol: the gap in mean APS (YOLO minus ResNet) moves from the archived protocol's −0.2126 to −0.1357 on the archived images and −0.1121 on repaired coverage, and the mean ΔF_1 gap from +0.0833 to +0.0804 and then to +0.0282.

Under the unified protocol on the archived images the rank association within each condition is positive in 16 of the 25 YOLO conditions, 15 of the ResNet conditions and 14 pooled, two YOLO and one pooled condition being exactly zero. These conditions reuse the same models and images. Analyses omitting one ResNet checkpoint at a time on the same cells give adjusted correlations of 0.34 without ResNet-50, 0.41 without ResNet-18 and 0.58 without ResNet-18+GAM; with only two remaining checkpoints each these are sensitivity diagnostics, not evidence that a capacity class explains the effect, and no subset or score convention yielding a larger estimate is adopted as the reported result; see Appendix D.

At $\sigma = 30$, BM3D is best by F_1 for seven of the eight checkpoints on the archived images under the unified protocol (main-text Figure 7 and Section 5.3). That advantage occurs at high noise: over $\sigma = 1$, 5, 10, 20 and 30, BM3D is best among the five denoising pipelines for 3, 6, 4, 8 and 7 of the eight checkpoints on the archived images. That count excludes the untouched noisy image, which itself outscores BM3D for five, two and four checkpoints at $\sigma = 1$, 5 and 10, and for none at $\sigma = 20$ or 30 (Table S1). The experiment therefore demonstrates score variation across detectors, not a demonstrated advantage in choosing different denoisers for different detectors.

Appendix G.5 Attribution, Target Mass and Polarity

The rerun recomputed only EigenCAM and the activation energy map, so the attribution comparison below uses the archived protocol and the archived aggregate scores of Table S2. We compare EigenCAM with activation energy, Grad-CAM and Grad-CAM++, the gradient methods taking the summed score for the pedestrian class over anchors as their scalar target. Across the six pipelines and $\sigma \in \{10, 20, 30\}$, their Spearman correlations with EigenCAM scores are 0.942, 0.948 and 0.959, falling to 0.543, 0.486 and 0.543 at $\sigma = 30$ alone. BM3D ranks first under all four attributions at the reference checkpoint, while the complete ordering below BM3D is not preserved (main-text Figure 3). The three alternative attributions lack per-image scores, so no confidence intervals are inferred.

The archived $\mathrm{APS}_{\mathrm{tgt}}$ means range only from 0.258 to 0.269 at $\sigma = 30$, include the 67 images without persons and have no chance baseline normalised for area, so we report their weak discrimination between pipelines without treating absolute target mass values as evidence of faithful localisation. A weak ancillary per-image diagnostic of the same archived protocol pools images containing persons across checkpoints and conditions: the area under the receiver operating characteristic (ROC) curve (AUC) separating images with more than half their persons recalled from the remainder is approximately 0.54 for APS, 0.56 for SSIM and 0.52 for PSNR. This pooled descriptive result does not establish an operational per-image gate.

Using $|\rho|$ instead of signed correlation changes several comparisons below BM3D. On the 160 archived images at $\sigma = 30$ under the unified protocol, the mean absolute correlations at the reference checkpoint are 0.9202 for BM3D, 0.7527 for Gaussian filtering, 0.7238 for CAE+PSO, 0.7036 for the autoencoder, 0.6957 for the noisy baseline and 0.6828 for DnCNN, so DnCNN falls from fifth under the signed convention to last. The BM3D lead survives, but the contrast of Gaussian filtering with the autoencoder and CAE+PSO changes sign: Gaussian filtering at 0.5298 sits below the autoencoder at 0.5505 and CAE+PSO at 0.5382 under the signed convention, and above both in absolute terms. That sign change is a property of the archived cells: on repaired coverage (learned pipelines only), excluding the regenerated DnCNN $\sigma = 30$ cells (main-text Section 5.2), Gaussian filtering stays below the autoencoder and CAE+PSO under either convention, so the sign change disappears. This sensitivity concerns the meaning of the score, not random sign nondeterminism between repeated executions.

The activation energy map, the one alternative attribution the rerun recomputed, alone compares like for like: on the same grid of 18 cells at the reference checkpoint, with six pipelines and $\sigma \in \{10, 20, 30\}$, the Spearman correlation between energy and EigenCAM scores moves from 0.9422 under the archived protocol to 0.9505 under the unified protocol, and the $\sigma = 30$ value from 0.5429 to 0.7714.

Appendix G.6 Pixel Swap Diagnostics

Each random control translates the high mask by a seeded cyclic offset that is accepted only when its overlap with the original mask is at most 25% of mask area (main-text Section 3.4). Across the 25,600 rows for random grafts every draw satisfied this limit, and the realised overlap averaged 0.080 of mask area.

Main-text Table 4 reports the eight cells, from 5,000 bootstrap replicates over images at seed 14092026; its intervals are dependent, pointwise and exploratory, without simultaneous coverage or a multiplicity adjustment (main-text Section 4.4). An exploratory per-image summary at the primary cell favours the high graft without separating attention from ground-truth coverage: the high graft exceeds the random mean on 43 of the 160 images (26.9%), falls below on 31 (19.4%) and ties on the other 86, of which 67 have no ground truth. A sign-flip test of the mean per-image F_1 difference gives $p = 0.039$; this is the separate exploratory estimand of main-text Section 4.4, not a test of the primary micro-F_1 contrast.

The rerun's BM3D arm connects it to the archived probe: 138/52/65 true positives, false positives and false negatives under one-to-one matching in main-text Table 3, against the archived 138/56/65 recorded in Appendix C, whose micro-F_1 is 0.6952. The autoencoder arm shows no such continuity: from the archived probe's 0.624, re-inference of those archived images under the unified protocol moves it by -0.012 to 0.612, regeneration with the stored weights costs 0.123 micro-F_1 (0.489) and the coverage repair returns 0.027 (0.516). Detection is therefore more sensitive to regeneration than the roughly 0.8 decibels (dB) PSNR agreement between regenerated and archived outputs indicates. Grafting the autoencoder images regenerated

at full coverage against the archived BM3D reference widens the archived probe's +0.072 precondition gap to +0.1866.

The mask targeting high attention edits more ground-truth pixels than the translated control. The mean per-image fraction of the union of person boxes edited is 0.153 against 0.102, assigning zero to images without persons, a factor of 1.50 at the primary cell and 2.67 under the energy map. Ground-truth coverage is associated with F_1 across the 20 control draws (arm-level $r = +0.397$; within-image $r = +0.249$ after removing image means).

A post hoc diagnostic selects, for each image, the $k = 3$, 5 or 10 random controls closest to the high mask's ground-truth fraction, with ties resolved by draw index. It averages their TP, FP and FN before computing the matched arm's micro-F_1, a different estimand from the primary mean of micro-F_1 values computed for each draw. High minus matched ranges from +0.028 to +0.029, with all bootstrap intervals over images including zero. This comparison remains poorly balanced: mean matched coverage ranges from 0.099 to 0.111 against 0.153 for the high mask, and no random draw reaches its coverage on 50 of 160 images. Matching selects observed controls; it does not extrapolate a regression. Moreover, coverage could be a mediator if an informative map lands on pedestrians, so conditioning on it could remove part of the association of interest.

The design cannot separate editing pixels with high attention from editing pedestrian pixels. Equal mask area also permits different perturbation amplitudes: the ratio of high to random mean absolute pixel change is 1.03 at the primary cell. The primary interval resamples images only; 93 of the 160 images carry person annotations, and the 20 random locations enter as a fixed average whose standard deviation across draws (0.019) is of the same order as the estimate (0.030).

A secondary diagnostic repeats the recovery diagnostic of Appendix B under the rerun's one-to-one matching rather than the looser archived rule (Appendix F.1); the two are different estimands on different cohorts and are not compared numerically. Its cohort comprises 61 ground-truth instances over 45 images detected by BM3D but missed by the autoencoder output regenerated at full coverage. The high graft catches 23 (37.7%) against a mean across random draws of 13.65 (22.4%, range 7 to 18, none reaching 23), the reverse graft 43. This is a secondary count on a fixed cohort, not a second estimate of the primary estimand.