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.

What Multi-Artifact CT Classifiers Learn: Shortcuts, Hidden Confusions, and Localisation Targets

R Kanishka Varshini
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee https://orcid.org/0000-0001-8829-9980

Research Article

Keywords: Artifact detection, attribution maps, computed tomography, image normalisation, multi- label classification, occlusion analysis, shortcut learn- ing, synthetic data generation.

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11733275/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.

What Multi-Artifact CT Classifiers Learn: Shortcuts, Hidden Confusions, and Localisation Targets

R Kanishka Varshini[1] and Mayank Goswami,[1,2]

[1]Department of Physics, IIT Roorkee
[2]MF School of AI and DS, IIT Roorkee
mayank.goswami@ph.iitr.ac.in

***Abstract*—Automated quality control in computed tomography begins by identifying which artifacts a reconstruction contains, because every correction is artifact-specific. Progress on this task is reported as a gain in F_1 or exact-match accuracy, always measured under one choice of image normalisation. We show that this choice decides what the reported number means. We build a synthetic corpus in which beam-hardening, ring, metal-streak and scatter operators are injected independently, so that the four-label presence vector is exact by construction and severity is recorded: 2,400 reconstructions split disjointly by phantom, with 360 more at severities outside the training range. A multi-label convolutional network reaches 0.9965 macro-F_1 and 0.9860 exact-match, against 0.9811 and 0.9302 for gradient boosting on fourteen hand-designed features. Three findings transfer beyond this corpus. First, the feature-based models read beam hardening and scatter from overall brightness rather than from cupping shape: two brightness features alone reach 0.797 macro-F_1, and the network loses nothing when that cue is removed. Second, the shortcut conceals the confusion the physics predicts, since removing brightness raises beam-hardening/scatter confusion from 6.6% to 42.1%. Third, occlusion shows that the annotated metal region is neither necessary nor sufficient for detection, so mask-overlap scores penalise a model for reading the correct evidence.**

***Index Terms*—Artifact detection, attribution maps, computed tomography, image normalisation, multi-label classification, occlusion analysis, shortcut learning, synthetic data generation.**

I. Introduction

Automated quality checking in computed tomography starts by finding which artifacts are present in a reconstructed image. Beam hardening, rings, metal streaks and scatter rarely appear alone. They occur together at different strengths, and the right correction depends on which ones are there. A multi-label classifier is trained to predict which artifacts are present, and its progress is reported as a gain in F_1 or exact-match accuracy. That number decides which detector gets used. But it is always measured under one choice of image normalisation. The same classifier, measured after a different choice, can be relying on completely different evidence.

Corresponding author: mayank.goswami@ph.iitr.ac.in

One accuracy number quietly mixes two different things: whether the model sees the shape of the artifact, or only sees the change in overall brightness that comes with it. Beam hardening and scatter both lower the overall brightness of the image. Both also produce cupping. A model can flag either one from brightness alone, without ever learning what cupping looks like. Both routes give a high score. Once the two are mixed, the comparisons the field relies on stop being reliable. The brightness cue also hides a confusion that the physics predicts. Beam hardening and scatter look alike, and a model that reads brightness never has to tell them apart.

Earlier work has built strong detectors and correction methods for single artifacts, metal most of all. It has also shown that simulation can give exact labels where hand annotation cannot. But this work mostly treats artifact detection as finished once accuracy is high, instead of asking what the accuracy rests on. A separate line of work scores heat maps against marked regions. It assumes the marked region is the evidence the model uses. Neither question has been asked for multi-artifact detection, where artifacts overlap and spread across the image by different amounts.

Three questions therefore remain open. First, when a multi-label artifact classifier scores well, is it reading the artifact or only a side effect of it? Second, if that side effect is taken away, which confusions appear that were hidden before? Third, when we check where the model looks, is the marked region really the evidence it uses?

We study these questions with a deliberately controlled design. We build a synthetic dataset in which four corruption steps are applied independently during projection and reconstruction. The list of artifacts in each image is therefore exact by construction, not annotated by hand. Each corruption also has a minimum strength that we calibrate, so no image is ever labelled with an artifact that cannot be seen. We then keep the dataset, the splits and the scoring code fixed, and change one thing: whether overall brightness survives normalisation. We train feature-based

baselines and a convolutional network under both settings. Finally, we test the explanations by erasing regions and re-running the model, rather than by overlap with a marked region. Our contributions are:

- **A shortcut decomposition.** Feature-based classifiers detect beam hardening and scatter mainly from overall brightness, not from the shape of the cupping. Two brightness features on their own recover most of the reported accuracy. A convolutional network loses nothing when that cue is removed, so it is reading the shape itself.
- **A hidden-confusion account.** The brightness cue hides the confusion between beam hardening and scatter. While the cue is available, the two are rarely confused. Once it is removed, feature-based models confuse them badly in both directions, while the convolutional network does not. The look-alike problem is real, and it only looked absent because of how it was measured.
- **A localisation-target finding.** By erasing regions and re-running the model, we show that the marked region is both needed and enough for ring artifacts, but is neither for metal. The model reads metal from the streaks that spread across the whole slice, while the marked region covers only the small insert. Overlap scores therefore punish the model for looking in the right place. We argue that erasing regions is the right test when an artifact spreads out.
- **A dataset-design trade-off, and what it means for measurement.** Making sure every labelled artifact is visible is exactly what pushes the task close to ceiling. We report two attempts to make the task harder that did not work, and why. We conclude that reported accuracy on multi-artifact detection depends on the normalisation choice, and that the evidence a classifier uses has to be measured, not assumed.

II. Problem Statement and Motivation

Why identification comes before correction. A quality-assurance pipeline has three decisions to make about a scan: accept it, flag it for a re-scan, or correct it. The third decision is the hard one, because corrections are artifact-specific. Metal artifact reduction operates on the sinogram around starved rays. Ring removal works on detector-column statistics. Beam-hardening correction fits a linearisation curve. Each assumes the artifact it was written for. Running the wrong one does not simply fail; it can add structure that was never in the object. A pipeline therefore has to know which artifacts are present before it can choose anything. Identification is not a side task. It is the step everything else depends on.

Why severity matters as well as presence. The same artifact at two strengths calls for two different decisions. A faint ring may be acceptable. A strong one may need a re-scan. A detector that only reports presence leaves the operator with the harder half of the judgement. Any useful system has to carry severity alongside the label, and that in turn requires training data where severity is known rather than estimated.

Why this is a multi-label problem. With four artifact types there are sixteen possible states, from a clean image to all four together. A single-label classifier cannot represent this. Treating the sixteen states as sixteen classes is possible but wasteful, because it hides the fact that the classes share structure, and it gives no way to handle a combination that was never seen in training. The correct framing is four independent outputs, each free to fire on its own. This also changes how performance should be read. Getting each label right on its own is not the same as getting all four right at once, and only the second tells us whether the pipeline can act on the output.

Why the data has to be built rather than found. Exact labels for every combination exist only when each corruption is injected on purpose. We searched the public datasets and found none that provides them. The closest is the AAPM metal-artifact-reduction corpus, which simulates metal with beam hardening and scatter present alongside it. Those two are never labelled separately, and rings do not appear at all. Other collections carry real scans with no artifact labels, or a single distortion class at low resolution. In every case the artifacts that co-occur are exactly the ones left unlabelled, which is the information this problem needs. Building the corpus is therefore not a convenience. It is the only route to the labels, and it also hands us the control the measurement question requires.

Why the measurement question is the one worth asking. Once a corpus like this exists, high accuracy is easy to reach, and that is the trap. A model can score well by reading the artifact, or by reading something that merely travels with it. The two cannot be told apart from an accuracy table. They can be told apart by changing the preprocessing and watching which models survive. Nobody has done this for multi-artifact detection, and without it a reported number says little about whether a detector will hold up on a scanner whose intensity scale differs from the one it was trained on.

Objectives. From this, the work sets out to do five things.

1) Build a multi-artifact corpus in which the presence vector is exact by construction, severity is recorded, and no label is attached to an artifact that cannot be seen.
2) Establish a fair classical baseline on hand-designed features, and a convolutional baseline, scored by identical code on identical splits.
3) Determine what evidence each of them actually uses, by removing the overall-brightness cue and measuring what is lost.
4) Measure the confusion between beam hardening and scatter under both conditions, since the physics predicts the two should be hard to separate.

5) Test whether the regions we mark as ground truth are the regions the model relies on, by erasing them and re-running the model.

III. Background and Problem Formulation

Image formation. For a monochromatic beam of energy E, the intensity reaching a detector bin after passing through the object along a ray ℓ follows the Beer–Lambert law,

$$I = I_0 \exp\!\left(-\int_\ell \mu(x,E)\,dx\right), \tag{1}$$

where $\mu(x,E)$ is the linear attenuation coefficient at position x. Taking the negative logarithm gives the line integral

$$p = -\ln\frac{I}{I_0} = \int_\ell \mu(x,E)\,dx. \tag{2}$$

Collecting p over detector bins d and view angles θ gives the sinogram $p(d,\theta)$, and filtered back-projection inverts it to recover μ. Every artifact we study is a departure from this idealised chain, introduced at a different point in it.

Beam hardening. A real X-ray tube emits a spectrum, not a single energy. Writing $w(E)$ for the normalised photon-number weight, the detected intensity is

$$I = I_0 \int w(E)\exp\!\left(-\int_\ell \mu(x,E)\,dx\right) dE, \tag{3}$$

so the measured line integral is $p_{\text{poly}} = -\ln\int w(E)\,e^{-p_E}\,dE$, with p_E the line integral at energy E. Because e^{-x} is convex, Jensen's inequality gives

$$p_{\text{poly}} \;\le\; \bar{p} \;=\; \int w(E)\,p_E\,dE, \tag{4}$$

with equality only when p_E does not vary with energy. The gap widens as the spread of p_E grows, and that spread grows with path length, because low-energy photons are absorbed first and the surviving beam hardens as it travels. Thick paths are therefore under-measured relative to thin ones. After reconstruction this appears as a depressed centre, known as cupping. A first-order water calibration, $p_{\text{cal}} = k\,p_{\text{poly}}$ with $k = \mu_w(E_{\text{eff}})/\int w(E)\mu_w(E)\,dE$, removes the linear part of this error and leaves the non-linear residual, which is the artifact proper.

Ring artifacts. A miscalibrated detector bin applies a fixed gain g_d to the intensity it reports,

$$I_{\text{meas}}(d,\theta) = g_d\,I(d,\theta) \implies p_{\text{meas}}(d,\theta) = p(d,\theta) - \ln g_d. \tag{5}$$

The error is a constant offset in one detector column, identical at every view angle. In parallel-beam geometry, bin d corresponds to rays at a fixed distance s_d from the rotation centre, so back-projecting a constant offset across all angles places it on a circle of radius $|s_d|$. The ring position is set by the faulty bin, and nothing else.

Metal. Detected photon counts are Poisson,

$$N \sim \text{Poisson}\big(I_0 e^{-p}\big), \qquad \hat{p} = -\ln\frac{\max(N,1)}{I_0}. \tag{6}$$

The variance of $\hat{p}$ is approximately $1/N$, which grows exponentially with p. A dense insert pushes p high enough that N falls to a few counts, so those rays return wildly unreliable values. This is photon starvation. The ramp filter amplifies the resulting high-frequency errors, and back-projection spreads each one along its own ray, producing the streaks that radiate from the insert. Note where the signal lives: the corrupted measurements belong to every ray that crossed the metal, not to the metal region itself.

Scatter. Photons deflected out of their primary path land elsewhere on the detector and add a smooth, low-frequency field S to the primary signal I_p. The measured line integral becomes

$$p_{\text{meas}} = -\ln\frac{I_p + S}{I_0} = p - \ln\!\left(1 + \frac{S}{I_p}\right). \tag{7}$$

Because S varies slowly across the detector while I_p varies by orders of magnitude with path length, the ratio S/I_p is largest exactly where I_p is smallest. Thick paths are therefore under-measured most, and cupping follows.

Why two of the four look alike. Beam hardening and scatter both reduce to the same first-order form,

$$p_{\text{meas}} = p - \delta(p), \qquad \frac{d\delta}{dp} > 0, \tag{8}$$

an under-measurement that increases with path length. Their mechanisms are unrelated, but their leading image signature is identical: a depressed centre. They differ only in higher-order behaviour and in what happens outside the object, since scatter adds counts in air while beam hardening does not. This has a direct consequence for measurement. A classifier that reads only the depth of the depression cannot separate the two, while one that reads the full radial profile might. Any evaluation that does not distinguish these cases will report a confusion rate that depends on which of them the model learned.

Multi-label formulation. Let $y \in \{0,1\}^4$ record the presence of each artifact. The network produces logits $z \in \mathbb{R}^4$ and probabilities $\hat{y}_k = \sigma(z_k)$, trained with the summed binary cross-entropy

$$\mathcal{L} = -\sum_k \left[y_k \log \hat{y}_k + (1-y_k)\log(1-\hat{y}_k)\right]. \tag{9}$$

The alternative, a softmax over the sixteen combinations, forces the outputs to sum to one and treats each combination as an unrelated class. It cannot share structure between combinations that contain the same artifact, and it has no way to predict a combination absent from training. The sigmoid formulation avoids both problems, at the cost of assuming conditional independence between labels given the image.

Metrics. We report per-label F_1 and its macro average, Hamming loss, and exact-match accuracy, the fraction of images for which all four labels are simultaneously correct. The last is the demanding one. If per-label errors were independent with rates e_k, exact-match would be about

$\prod_k (1 - e_k)$, so it falls far faster than any per-label score. It is also the metric a pipeline actually needs, because a correction is chosen from the whole vector, not from one entry.

Attribution. Grad-CAM explains output k by weighting the channels of a convolutional layer by the spatial average of their gradients,

$$\alpha_c^k = \frac{1}{HW} \sum_{i,j} \frac{\partial z_k}{\partial A_{c,ij}}, \qquad L^k = \mathrm{ReLU}\left(\sum_c \alpha_c^k A_c \right). \tag{10}$$

In a multi-label setting there is no single predicted class, so one map is produced per output. One property of this definition matters later: when z_k is confidently negative, the weighted sum is negative almost everywhere and the ReLU returns an empty map. An empty map is a meaningful answer, not a failure.

Attribution as a causal question. Mask-overlap scores, such as the pointing game and intersection-over-union, compare L^k against a marked region R. They answer whether the model looks where we marked, and they assume that R is the evidence. That assumption can be tested directly. Writing $\hat{y}_k(I)$ for the predicted probability on image I, and $I_{\neg R}$, I_R for the image with R neutralised and with everything but R neutralised,

$$\underbrace{\hat{y}_k(I_{\neg R}) \ll \hat{y}_k(I)}_{\text{necessity}}, \qquad \underbrace{\hat{y}_k(I_R) \approx \hat{y}_k(I)}_{\text{sufficiency}}. \tag{11}$$

A region that is both necessary and sufficient is the evidence. A region that is neither is not, however well the heat map happens to overlap it. We also note that the pointing game has a chance rate of $|R|/|\Omega|$ for an image domain Ω, so scores quoted without it are not interpretable.

IV. Related Work

Correcting single artifacts. Most CT artifact work picks one artifact and removes it. Metal has had the most attention. Methods range from filling in the damaged sinogram readings [1] to networks that fix the projection and the image together [2], [3]. Rings are removed by suppressing the faulty detector columns, or by filtering the image in polar coordinates [4]. Beam hardening is corrected by fitting a curve measured on a known material [5], [6]. Scatter is estimated with convolution kernels, by simulating photon paths, or by a trained network [7]. All of these start from an artifact that is already known. Deciding which one is there is left to the operator. Challenge datasets label only the artifact that challenge is about [8]. We did not find a multi-label CT artifact classifier tested across combinations.

Synthetic CT data. Real scans do not come with a label for every pixel. Simulation is the usual way to get exact ones. Toolkits do the projection and reconstruction for standard scan geometries [9], [10]. Phantom libraries give test objects that anyone can regenerate [11]. Physics-based simulators model the X-ray spectrum, scatter and the detector in detail [12]. The usual pattern is to damage a clean image and train on the pair. We do the same, with one difference. Each corruption is a separate step that can be switched on or off, so the label is just a record of which steps ran.

Multi-label classification. The standard setup gives one output per label, each trained with its own cross-entropy [13]. Scores are per-label F_1, Hamming loss and exact-match accuracy. Later work adds losses that handle rare labels, and designs that let the labels inform each other [14]. Our labels are close to balanced by construction, so we use the plain setup.

Explaining classifiers, and checking the explanations. Class activation maps and their gradient versions are the usual way to ask where a classifier looks [15], [16]. They are scored against a marked region, using the pointing game or intersection-over-union [17]. A separate line of work changes the image instead. It erases or hides parts of the image and watches what the output does [18], [19]. Sanity checks have shown that some maps barely depend on the model they explain [20]. What none of this asks is whether the marked region is right. Overlap scores treat it as correct by definition. That is safe when an artifact sits in one place. It is not safe when the artifact spreads.

Shortcut learning. Networks often read something that travels with the label instead of the label itself. Reported cases include a scanner token rather than the disease [21], texture rather than shape [22], and cues that only hold in one dataset [23], [24]. These are usually found after the model is deployed, once the side effect goes away. Our case is different in two ways. The brightness cue is a real physical consequence of the artifact, so using it is not obviously wrong. And because we build the dataset ourselves, we can take the cue away with one change to preprocessing. That makes it something we can measure, not something we find out later.

V. Method

A. Dataset construction

Design principle. The label is fixed before any physics runs. For each image we first draw which corruption steps will be applied and how strong each one is. We then simulate a scan that fails in exactly that way. The list of artifacts is a record of that draw, so it is exact by construction and nothing is annotated by hand.

Phantoms. Each object is a 256×256 map of material labels, not grey values. This is not a cosmetic choice. Beam hardening needs μ as a function of energy, and that can only be looked up if we know the material at each pixel. The outline is a closed curve built from random radial harmonics, which stops the network learning one fixed shape. The body is one of two polymers. Inside it we place one aluminium part, and one to five further

inclusions of air, polymer or aluminium. Every inclusion has to lie wholly inside the body. Attenuation values come from published mass attenuation coefficients multiplied by density. We interpolate them in log–log space to the energies we simulate. We generate 400 such phantoms.

Scan simulation. We use a parallel-beam geometry with 360 views over 180°, across a 12 cm field of view. Reconstruction is filtered back-projection with a ramp filter. The Radon transform is linear, so we project one mask per material once per phantom. Every energy and every configuration is then rebuilt as a weighted sum of those projections. This is why the whole corpus generates in minutes. It also means the single energy and multi energy paths share one physical model, rather than two separate pieces of code.

The four operators. Each one acts at the point in the chain where it physically belongs, and that ordering is part of the design. Fig. 1 shows their effects in isolation.

- *Metal* acts on the object. A steel insert is added to the material map before projection. We size it by *opacity*, $\tau = \mu \cdot \text{diameter}$, rather than by radius. Photon starvation only begins above $\tau \approx 8$, and sizing by radius gave phantoms whose metal could never streak at any strength. Severity scales the insert's attenuation.
- *Beam hardening* acts at beam formation. We take a Kramers bremsstrahlung spectrum at 120 kVp, harden it with aluminium filtration, and split it into seven energy bins. The line integrals are then combined as in (3). Severity is the filter thickness, from 2.0 mm down to 0.1 mm, so a higher severity means a softer beam. We apply first-order water calibration, which removes the global level shift and leaves the non-linear residual.
- *Scatter* acts at the detector. We take the scatter source as $1 - I$, since scatter is born where photons leave the primary beam, and convolve it with a broad kernel. This matters. Blurring I itself makes the scatter field follow the primary and gives a flat offset instead of cupping, which would have made the look-alike pair of (8) too easy to separate. The field is rescaled so the scatter-to-primary ratio means what it says inside the object.
- *Ring* is a detector fault, applied after scatter arrives. A set of bins gets a fixed gain error as in (5). We draw those bins so the ring radii land inside the sample, because a fault at the detector edge makes a ring in empty air.

Poisson detection comes last. The incident photon count is drawn log-uniformly over a twenty-fold range, so dose is a nuisance variable rather than a constant.

Severity-floor calibration. Drawing severity from zero upward would label artifacts that cannot be seen. That is label noise, in a dataset whose whole point is exact labels. So we calibrate a minimum strength for each artifact. For each of 24 calibration phantoms we simulate eight clean versions that differ only in the noise seed. This tells us how much a detection index moves under noise alone. A severity counts as detected when it shifts that index by at least four standard deviations. The floor is the smallest severity detected on at least 95% of phantoms. We calibrate at the noisiest dose in the range, so the guarantee holds everywhere.

The test is *paired within phantom*. A fixed threshold would ask the artifact to stand out against the object's own structure, which is the wrong question. Measured that way, detectability tracked base material and inclusion layout rather than severity. The floors we obtain are 0.02 for beam hardening, metal and scatter, and 0.65 for rings. Sampled severities are then mapped onto the range above each floor.

Composition and splits. Each phantom is rendered in six configurations, giving 2,400 images (Table I). The configurations are not drawn uniformly. We fix a quota per combination in advance and deal those slots across phantoms. This way all sixteen combinations appear in every split, no phantom gets the same combination twice, and each combination is spread over many different phantoms. Per-label positive rates land at 0.487. Splits are *by phantom*, 280 / 60 / 60, so no object appears in more than one split. Training draws severity from the middle of each range. A further 360 images reuse the test phantoms at severities outside that range, for the generalisation test.

TABLE I
DATASET COMPOSITION

Artifacts present	Combinations	Images each	Total
0 (clean)	1	300	300
1	4	150	600
2	6	120	720
3	4	120	480
4	1	300	300
Main corpus			2,400
Held-out severity set			360
Phantoms (train / val / test)			280 / 60 / 60

Normalisation, the controlled factor. We use two regimes. The *fixed* regime maps a constant attenuation window to $[0, 1]$, which is the analogue of clinical windowing and keeps the overall brightness level. The *per-image* regime maps each image's own 1st and 97th percentiles to $[0, 1]$, which takes that level away. We use the 97th percentile and not the 99.5th, and this was measured rather than guessed. A metal insert covers one to two percent of the frame. A 99.5th-percentile window therefore lands inside the metal and squeezes the body into the bottom few percent of the range. On metal and ring images, the contrast kept in the non metal object is 0.058 at the 99.5th percentile against 0.47 at the 97th.

B. Models

Hand-designed features. We compute fourteen measurements per image. They cover cupping, radial oscil-

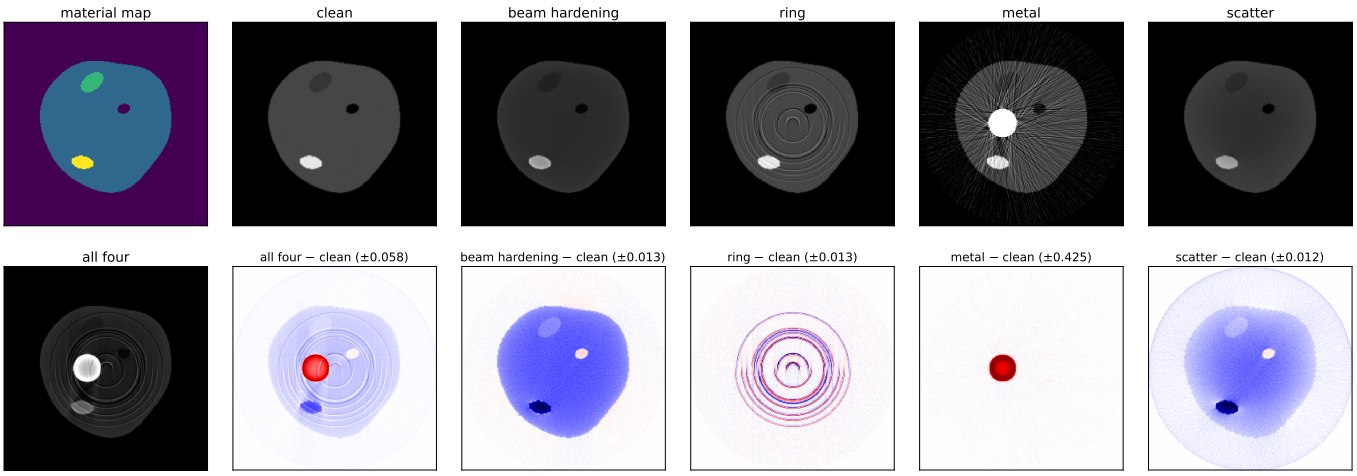


Fig. 1. Each corruption operator in isolation, on one phantom at severity 0.85. Top row: the material map, the clean reconstruction, and each artifact applied on its own. Bottom row: all four together, and the difference from the clean reconstruction in each case, on a symmetric scale whose range is given in the panel title. Beam hardening and scatter produce the same depressed centre, which is the shared signature (8) predicts.

lation coherence, bright-blob statistics, gradient energy, background level and texture. Two of them, the object mean and the background level, measure overall brightness. The other twelve describe shape or texture. This split is what the ablation uses.

Rule-based baseline. For each artifact we search all fourteen features, in both directions, over a grid of thresholds taken from the training distribution. We keep whichever single threshold gives the best validation F_1. We search instead of assigning features by hand on purpose. Which feature each rule picks is then a result, not an assumption.

Tree ensembles. We fit a Random Forest and an XGBoost model per label on all fourteen features. Hyperparameters are selected on validation by AUROC.

Convolutional classifier. We use a ResNet-18 with its first convolution rebuilt for one input channel and a final layer of four sigmoid outputs, trained with the summed binary cross-entropy of (9). For ImageNet initialisation we *average* the three input channels of the first convolution into one, rather than slicing one out. Averaging keeps the learned filters responding sensibly to a grey image. Training uses Adam at 10^{-4} with step decay, weight decay 10^{-4}, batch size 32, and up to 100 epochs. We stop early on validation macro-F_1 with patience 15.

Augmentation. We rotate about the image centre and flip horizontally and vertically. *Translation is deliberately left out.* Rings are concentric about the rotation axis, and cupping is centred on it. A shift destroys both signatures while leaving the label unchanged, which would teach the network that a displaced ring is still a ring. Intensity jitter is left out as well. Under the fixed regime it would erase the very cue the experiment measures.

C. Evaluation protocol

Selection discipline. Hyperparameters and per-label decision thresholds are chosen on validation only. The test split is scored once. We tune thresholds per label because exact-match accuracy is very sensitive to them, and the default of 0.5 is rarely the best choice. Every model is scored by one shared implementation of the metrics, so no difference between models can come from measuring them differently. Convolutional results are reported as mean $\pm$ standard deviation over three random seeds, varying both the training seed and the augmentation stream.

The shortcut ablation. Holding the corpus, the splits and the scoring fixed, we fit the best feature-based model on three feature sets: all fourteen, the twelve shape features alone, and the two brightness features alone. In parallel we train the convolutional model under both normalisation regimes. Only one factor changes in each case, so the cost of removing brightness is measured, not guessed.

Attribution. We produce Grad-CAM maps per output at two depths, giving 8×8 and 16×16 resolution before upsampling. Where ground-truth masks exist we report the pointing game, intersection-over-union, and the share of heat inside the mask divided by the mask's share of the image. Every overlap score is reported next to its chance rate, which for the pointing game is the mask area fraction. We dilate ring masks by four pixels before scoring, since a map at this resolution cannot trace a 1.5-pixel annulus. The tolerance is reported rather than hidden. Finally we erase each marked region, and separately keep it in isolation. The removed area is filled with the local median of a surrounding ring, so the edit adds no new edge that could itself read as an artifact.

VI. Results

Every model is scored by the same code, on the same phantom-disjoint splits. Convolutional results are mean ± standard deviation over three seeds.

Benchmark. Table II gives the four model families on the held-out test split. Two things stand out. The convolutional model gains only about 0.016 macro-F_1 over the best tree ensemble. But it gains about 0.060 exact-match, which cuts the images with at least one wrong label from 25 to 5 out of 358. Getting each artifact right on its own was close to solved already. Getting all four right at once was not. The rule-based baseline shows the same split from the other side. It loses 26 points of exact-match while losing only 7 of macro-F_1. Single thresholds find each artifact well enough, and almost never produce a correct vector. That is the gap a pipeline would actually feel.

TABLE II
Multi-label artifact classification performance (F_1 and exact-match)

Model	BH	Ring	Metal	Scat.	Macro-F_1	Exact
Rule-based	0.759	0.942	0.969	0.986	0.9139	0.6676
Random Forest	0.936	0.971	1.000	0.994	0.9759	0.9050
XGBoost	0.968	0.959	1.000	0.997	0.9811	0.9302
CNN (per-image)	0.991	1.000	0.995	1.000	0.9965	0.9860

XGBoost shows no seed variance at all. The setup chosen on validation draws no random subsamples, so boosting on fixed data gives the same answer every time. All the uncertainty in this comparison comes from the convolutional side.

The brightness shortcut in the feature-based models We first measured which features individual labels depend on in XGBoost. Beam hardening relies on object mean intensity (0.72 of its total importance), while scatter is governed almost entirely by background level (0.96).Both of those measure overall brightness.Both features capture overall brightness rather than spatial geometry. In contrast, ring artifacts depend on radial oscillation and metal streaks on peak intensity, as expected from their physical structures.

Table III evaluates XGBoost across these feature subsets to confirm the shortcut.Two brightness features alone are sufficient to achieve a macro-F_1 of 0.797. Conversely, removing brightness costs 18 points in exact-match accuracy. This degradation falls almost entirely on beam hardening and scatter, while ring and metal detection remain intact. Hand designed features thus depend heavily on brightness cues for the cupping artifacts.

Robustness of the convolutional network The convolutional network behaves very differently. When brightness is removed via normalization, its score barely moves (macro-F_1 shifts by only $+0.0002 \pm 0.0032$, and exact-match by $+0.0009 \pm 0.0126$, which is essentially zero). While the brightness shortcut accounts for nearly a fifth of the hand-designed model's exact-match score, the neural network learns real spatial shapes and ignores the shortcut completely.

TABLE III
Feature-subset ablation for XGBoost across shape and brightness features.

Feature set	BH	Ring	Metal	Scat.	Macro-F_1	Exact
All 14	0.968	0.959	1.000	0.997	0.9811	0.9302
Shape only (12)	0.884	0.974	0.994	0.867	0.9298	0.7542
Brightness (2)	0.777	0.657	0.772	0.983	0.7972	0.3184

Unmasking the hidden confusion. Table IV measures how often each model mistakes one cupping artifact for the other when only one is present. When brightness cues are available, XGBoost appears to easily separate them, misclassifying scatter as beam hardening only 6.6% of the time and never in reverse.However, removing brightness exposes the look-alike problem predicted by the physics. Without brightness cues, XGBoost's confusion surges to 42.1% and 27.6% across both directions, confirming it was relying on brightness statistics rather than visual shapes. In contrast, the convolutional model maintains a low error rate (4.4% and 0.0%) under the same conditions, proving it distinguishes the two by learning subtle differences in the cupping geometry itself.

TABLE IV
Cross-label confusion between beam hardening and scatter across models and feature sets.

Model	Scatter → false BH	BH → false scatter
XGBoost, all features	6.6%	0.0%
XGBoost, shape only	42.1%	27.6%
CNN (per-image)	4.4%	0.0%

Data composition against quantity. We tested whether training only on single artifacts allows a model to detect artifact combinations, or if it must see mixtures during training. Because combining artifacts naturally produces more images, we added a third, size-matched set to separate data variety from total dataset size (Table V).At equal sizes, learning from mixtures costs just -0.016 macro-F_1 on single artifacts, and extra data in Arm C recovers this entirely.While mixed training slightly affects single-artifact recall, the critical finding is on combinations (Table V, Multi), a model trained only on isolated artifacts collapses, dropping from 0.951 to 0.173 exact-match because it cannot separate overlapping visual features it has never seen.

Difficulty does not scale with artifact count. Counterintuitively, classification difficulty does not rise monotonically with the number of defects. Across 0 to 4 artifacts, exact-match accuracy follows a U-shaped dip.For XGBoost it runs 1.000, 0.920, 0.898, 0.931, and 0.956, while for the convolutional network it runs 1.000, 1.000, 0.963, 0.986, and 1.000. Both models struggle most at two

TABLE V
PERFORMANCE COMPARISON UNDER VARIED TRAINING SET COMPOSITION AND SIZE (MACRO-F_1 / EXACT-MATCH).

Arm	n train	Single	Multi	All
A — singles only	630	1.000 / 1.000	0.656 / 0.173	0.724 / 0.480
B — combined, matched	630	0.984 / 0.966	0.991 / 0.951	0.990 / 0.961
C — combined, full	1,680	1.000 / 1.000	0.994 / 0.969	0.995 / 0.980

artifacts rather than four. When all four defects occur together, every detector fires, leaving little ambiguity. With two artifacts, the model must solve a harder problem that is correctly selecting which two of the four are active while ruling out the rest.

Held-out severities. When tested on artifact strengths outside the training range, both models degrade, but the convolutional network degrades far less. XGBoost's exact-match drops by roughly 10 points (from 0.930 to 0.833), whereas the convolutional model drops by only about 5 points (from 0.986 to 0.940). However, out-of-distribution testing reveals an interesting contrast regarding normalization. When evaluating unseen severities, a fixed-window setup retains more accuracy than per-image normalization, gaining $+0.0112 \pm 0.0033$ in macro-F_1. While overall brightness provides no genuine benefit on standard data, it acts as a helpful backup cue when the model encounters artifact severities it has never seen before.

Localisation. Table VI scores the attribution maps against ground-truth masks, with chance rates. Without those chance rates the numbers mislead. Ring masks cover 30% of the image, so 0.960 is only a $\times 3.2$ lift. Metal's apparently poor 0.280, against a mask covering 2.2%, is $\times 12.8$ lift demonstrating far sharper spatial selectivity.. The maps also come back *empty* for 61–84% of images where the artifact is absent, and never when it is present. That is what (10) predicts for a confidently negative logit. The network does not invent evidence for artifacts that are not there.

TABLE VI
ATTRIBUTION MAP LOCALIZATION SCORES AND RELATIVE LIFT AGAINST GROUND-TRUTH MASKS.

Layer	Artifact	Mask	Point.	Chance	Lift	IoU
8×8	Ring	29.7%	0.960	0.297	$\times 3.2$	0.625
8×8	Metal	2.2%	0.280	0.022	$\times 12.8$	0.043
16×16	Ring	29.7%	0.989	0.297	$\times 3.3$	0.244
16×16	Metal	2.2%	0.029	0.022	$\times 1.3$	0.030

Are the marked regions the evidence? Table VII tests whether the network actually looks at the marked artifact regions using occlusion test. We evaluate each artifact by erasing its marked region, and separately by keeping only that region.The ring and metal artifacts behave in opposite ways. For rings the marked region is both necessary and sufficient, in the sense of (11). The annotation and the evidence agree, and localisation is genuinely solved. For metal it is neither. Erasing the insert leaves detection untouched, and the insert on its own recovers barely more than half. The network reads metal from the streak field across the slice, just as Section III predicts. The annotation marks the small insert at the hub of those streaks.

TABLE VII
OCCLUSION TEST EVALUATION SHOWING MEAN ARTIFACT PROBABILITY AND DETECTION RATES.

Artifact	Original	Region erased	Only region kept
Ring	0.999 (100%)	0.167 (54%)	0.994 (100%)
Metal	1.000 (100%)	0.999 (100%)	0.359 (57%)

This resolves the low overlap score from Table VI(IoU 0.043) The low score does not mean the explanation failed, rather, the ground-truth mask covers only the center insert (2.2% of the image), whereas the network focuses on the surrounding streaks. It also explains why higher resolution heatmaps (16×16) scored worse on metal. As the map sharpens, it pinpoints individual streak lines sitting a median of 55 pixels away from the center, moving further away from the tiny center mask.

VII. LIMITATIONS

The task is close to saturation, and our own design caused it. The best model reaches 0.9965 macro-F_1, leaving little headroom. This follows directly from the severity floor of Section V. Guaranteeing that every labelled artifact is detectable is what makes the problem solvable to the ceiling. We tried twice to raise difficulty without weakening that guarantee, and both attempts failed in instructive ways. Narrowing and lowering the sampled severity band changed nothing, because the test band moves with the training band and the model simply relearns a lower threshold. Adding a twenty-fold dose variation also changed nothing, because the hand-designed features average over large regions or over all 360 views and are therefore insensitive to noise level. The remaining lever is the four-sigma detection criterion itself. Relaxing it would admit marginal artifacts and produce a harder task, at the cost of the label guarantee. We judged honest labels the more valuable property, but it is a trade and not a free choice.

The four corruptions are independent by construction; in reality they are not. A metal insert hardens the beam and scatters photons. Our generator injects metal, beam hardening and scatter as separate operators drawn independently, which is exactly what makes the presence vector exact. It also means the corpus contains combinations that physics would rarely produce, and omits the causal coupling that a real scan would show. This bounds the confusion results in particular: we measure how separable two artifacts are when they occur independently, not when one produces the other.

No sim-to-real check is possible. We searched the public datasets and found none carrying presence labels for several artifact types, so not even a qualitative numeric comparison on real data is available. Every result here is internal to the simulation, and the sim-to-real gap is unmeasured rather than small.

Geometry and physics are simplified. We use 2D parallel-beam acquisition rather than cone-beam with FDK reconstruction, a constraint of the available hardware. Artifact signatures are unaffected, but absolute geometry is not reproduced. Attenuation values are reference-range tabulations rather than exact cross-section data, scatter is modelled by convolution rather than Monte Carlo transport, and only one metal is used. Titanium was excluded because at these energies it requires an implausibly large insert to reach photon starvation.

Rings have a narrow usable severity range. Their calibrated floor is 0.65, so the training band spans roughly 0.72 to 0.93 of the parameter range. Ring severity is therefore poorly resolved, and the ring arm of the held-out-severity test is correspondingly weak.

Single architecture, single phantom family. We report one convolutional backbone. All phantoms come from one procedural family, a polymer body with inclusions, and that choice was itself forced: with aluminium bodies in the pool, the same detector fault produced a ring roughly three times less visible, so ring severity would have meant different things on different objects. Generalisation to other object classes and material regimes is untested.

Auxiliary severity prediction is implemented but not evaluated. The network carries optional severity regression heads, and severity labels are recorded for every image, but we report no results for them. Claims in this paper concern presence only.

VIII. Conclusion

We built a synthetic multi-artifact CT corpus in which four corruption operators are injected independently, so that the artifact presence vector is exact by construction, and we used it to ask what multi-label artifact classifiers actually learn.

The narrow answer is that the task is solvable. A convolutional classifier reaches 0.9965 macro-F_1 and 0.9860 exact-match, against 0.9811 and 0.9302 for the best tree ensemble on hand-designed features. Its advantage lies almost entirely in joint prediction rather than in any single label, which is the property a correction pipeline depends on. This is also the least interesting finding in the paper, and it follows from our own decision to guarantee that every labelled artifact is detectable.

Three results transfer beyond this corpus.

First, a high accuracy number can rest on a correlate of the artifact rather than the artifact. Beam hardening and scatter were being read from overall brightness, and two brightness features alone recovered most of the reported accuracy. The convolutional model lost nothing when that cue was removed, so the two model families were scoring similarly for different reasons. Ablating a suspected shortcut feature is cheap, and we think it should be routine before an artifact detector is compared or deployed.

Second, the shortcut concealed the confusion the physics predicts. Beam hardening and scatter produce the same first-order signature, and with the brightness cue available they were separated by two different intensity statistics rather than by shape. Removing it raised confusion from 6.6% to 42% for feature-based models while leaving the convolutional model unaffected. A confusion rate measured under one preprocessing convention says little about the task and a great deal about the representation.

Third, mask-overlap scores assume the marked region is the evidence. By erasing regions and re-running the model, we showed this holds for rings and fails for metal, where the classifier reads a streak field spanning the slice while the annotation marks the compact insert. The usual metrics then penalise a model for attending to the correct evidence. Where an artifact's signature is non-local, occlusion is the appropriate instrument.

Taken together, these indicate that reported accuracy on multi-artifact identification is confounded by preprocessing convention, and that the evidence a classifier relies on must be measured rather than assumed. The corpus, generator and evaluation code are seeded and regenerate end to end in minutes, so every result here can be reproduced without redistributing the data. The clearest next step is a second corpus in which the operators are physically coupled, so that metal produces its own beam hardening and scatter. That restores realism at the cost of exact labels, and the trade we describe in Section VII reappears in a sharper form.

IX. Video

https://youtu.be/AaKpczGDRo0

References

[1] E. Meyer, R. Raupach, M. Lell, B. Schmidt, and M. Kachelrieß, "Normalized metal artifact reduction (NMAR) in computed tomography," *Medical Physics*, vol. 37, no. 10, 2010.

[2] Y. Zhang and H. Yu, "Convolutional neural network based metal artifact reduction in x-ray computed tomography," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, 2018.

[3] W.-A. Lin, H. Liao, C. Peng, X. Sun, J. Zhang, J. Luo, R. Chellappa, and S. K. Zhou, "DuDoNet: Dual domain network for CT metal artifact reduction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] B. Münch, P. Trtik, F. Marone, and M. Stampanoni, "Stripe and ring artifact removal with combined wavelet–Fourier filtering," *Optics Express*, vol. 17, no. 10, 2009.

[5] G. T. Herman, "Correction for beam hardening in computed tomography," *Physics in Medicine and Biology*, vol. 24, no. 1, 1979.

[6] Y. Kyriakou, E. Meyer, D. Prell, and M. Kachelrieß, "Empirical beam hardening correction (EBHC) for CT," *Medical Physics*, vol. 37, no. 10, 2010.

[7] J. Maier, S. Sawall, M. Knaup, and M. Kachelrieß, "Deep scatter estimation (DSE): Accurate real-time scatter estimation for x-ray CT using a deep convolutional neural network," *Journal of Nondestructive Evaluation*, vol. 37, 2018.

[8] E. Haneda *et al.*, "AAPM CT metal artifact reduction grand challenge," *Medical Physics*, 2025, vERIFY: full author list, volume, DOI.
[9] W. van Aarle, W. J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K. J. Batenburg, and J. Sijbers, "Fast and flexible x-ray tomography using the ASTRA toolbox," *Optics Express*, vol. 24, no. 22, 2016.
[10] A. Biguri, M. Dosanjh, S. Hancock, and M. Soleimani, "TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction," *Biomedical Physics & Engineering Express*, vol. 2, no. 5, 2016.
[11] D. Kazantsev, V. Pickalov, S. Nagella, E. Pasca, and P. J. Withers, "TomoPhantom, a software package to generate 2D–4D analytical phantoms for CT image reconstruction algorithm benchmarks," *SoftwareX*, vol. 7, 2018.
[12] M. Wu *et al.*, "XCIST – an open access x-ray/CT simulation toolkit," *Physics in Medicine & Biology*, vol. 67, 2022, vERIFY: full author list, issue, article number.
[13] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, 2007.
[14] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.
[17] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, 2018.
[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*, 2014.
[19] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *British Machine Vision Conference (BMVC)*, 2018.
[20] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
[21] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Medicine*, vol. 15, no. 11, 2018.
[22] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *International Conference on Learning Representations (ICLR)*, 2019.
[23] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, 2019.
[24] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, 2020.