WHEN PHYSICAL VIEWS SHOULD NOT BE POSITIVE PAIRS: CROSS-FREQUENCY COLLAPSE IN SELF-SUPERVISED LEARNING FOR EDDY CURRENT TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many scientific instruments acquire co-registered channels through distinct measurement operators, making cross-channel positive pairing tempting in self-supervised learning. Correspondence, however, is not exchangeability: pairing declares the operator difference nuisance for the task. When that difference carries target signal, pairing can suppress information the task needs. We test this failure mode in multi-frequency eddy current testing (ECT), where frequency controls penetration and the operator difference is known from physics. Holding backbone, objective, augmentation policy, split, and evaluation fixed, we change only the second-view channel. Same-frequency pretraining reaches 0.69–0.71 macro-AUPRC against a 0.047 random-ranking reference; all nine cross-frequency comparisons degrade transfer, with penalties approximately 5%–20% of their matched same-frequency macro-AUPRC, grouped physical-unit support after multiplicity correction, and adverse directions in all three pretraining seeds. The effect spans VICRegL, DenseCL, and SimCLR: mechanisms that avoid trivial collapse do not guarantee preservation of task-relevant operator variation when the view distribution encourages invariance to it. LP-FT shrinks the mean gap from 0.0688 to 0.0222 under full labels, bounding *cross-frequency collapse* to a supervision-dependent transfer penalty. PhysAug provides the constructive counterpart by deriving views from known sensing nuisances. Thus, when operator and target dependence are sufficiently known, candidate positive pairs can be audited at design time. Our controlled ECT study therefore motivates a broader principle for multi-channel scientific sensing: positive-pair validity depends on whether operator-dependent variation is nuisance or downstream signal, not merely on physical correspondence.

1 INTRODUCTION

Positive pairs define invariances: pulling two views together encourages SSL to ignore the variation between them (Chen et al., 2020; Tian et al., 2020). In scientific sensing, instruments often provide channels that are perfectly co-registered or synchronized and therefore tempting positive views. Yet correspondence does not imply task-level exchangeability. If the channel difference carries downstream information, enforcing equivalence can suppress structure the representation should preserve (Xiao et al., 2021; Xue et al., 2023). We call the resulting question one of *positive-pair validity*: given two real measurements of the same state, should they be declared equivalent?

A measurement model makes the question precise. For a channel c,

$$\mathbf{x}_c = H_c(\mathbf{z}) + \boldsymbol{\epsilon},$$

where $\mathbf{z}$ is the physical state, H_c is the channel-specific measurement operator, and $\boldsymbol{\epsilon}$ is measurement noise. Pairing channels c_1 and c_2 as positives declares the difference between H_{c_1} and H_{c_2} irrelevant. That declaration is valid when operator-dependent variation is nuisance for the target and invalid when it carries target information. When the measurement model and target dependence are sufficiently known, candidate views can therefore be audited before pretraining and without labels. We use this as an operational design criterion.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

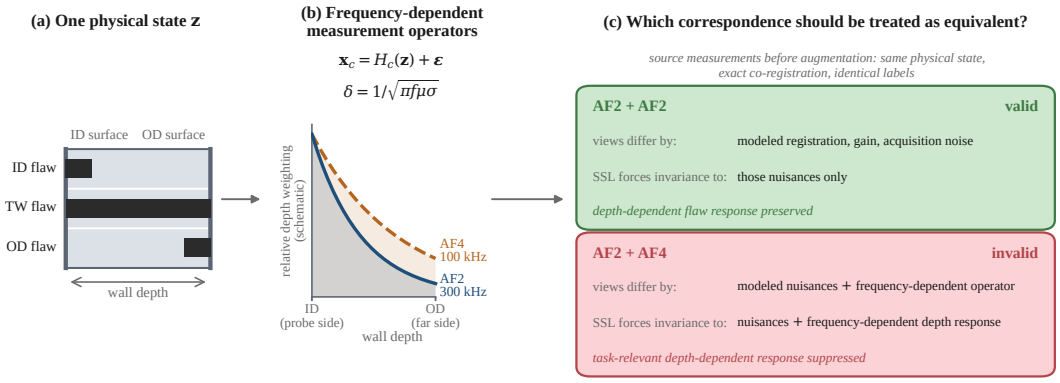


Figure 1: Positive-pair validity depends on measurement-operator exchangeability, not correspondence. **(a)** One tube-wall state contains ID, OD, and TW flaws at different depths. **(b)** Frequency changes through-wall weighting; lower-frequency AF4 penetrates deeper than AF2 (schematic, not calibrated response). **(c)** The source measurements are exactly co-registered and identically labeled, but AF2+AF4 additionally forces invariance to frequency-dependent depth response carrying task-relevant flaw-depth information.

Multi-frequency eddy current testing (ECT) of nuclear steam-generator tubing provides a controlled test (Figure 1). Here AF1–AF4 denote four co-registered axial frequency channels. Because penetration depth depends on frequency, these channels are spatially aligned measurements of the same region but are not exchangeable: higher frequencies emphasize near-surface response while lower frequencies probe deeper wall regions. Frequency therefore changes the measurement operator in a way that can remain relevant to inner-diameter (ID), outer-diameter (OD), and through-wall (TW) flaw detection.

We exploit this structure as a positive-pair intervention. Holding the backbone, SSL objective, PhysAug policy, split, and downstream evaluation fixed, we compare AF2+AF2 same-frequency positives with AF2+AF1/AF3/AF4 cross-frequency positives, so only the second-view channel changes. Same-frequency pretraining is effective: the three objectives reach 0.7147 (VICRegL), 0.7022 (DenseCL), and 0.6896 (SimCLR) macro-AUPRC against a 0.047 random-ranking reference. All nine low-label same-minus-cross contrasts remain supported after grouped physical-unit inference and within-family Benjamini–Hochberg (BH) correction. The penalty spans SimCLR instance discrimination (Chen et al., 2020), DenseCL dense correspondence (Wang et al., 2021), and VICRegL variance–covariance regularization (Bardes et al., 2022). We use *cross-frequency collapse* for this reproducible, supervision-dependent loss of transferability from forcing non-exchangeable channels into one equivalence class. It does not mean constant embeddings, universal failure of multi-frequency learning, or irreversible information loss; LP-FT recovers much of the gap under stronger supervision.

The same measurement model that invalidates a view can construct valid ones. PhysAug derives invariances from measurement nuisances rather than natural-image convention, encoding probe-registration, gain, and acquisition-noise variation, and is markedly more stable than a generic policy under exactly the registration nuisance it models. The general claim is that positive-pair validity in multi-channel sensing depends on measurement-operator exchangeability, not on shared semantics, shared provenance, or co-registration alone. ECT is the controlled setting where that dependence is checkable against known physics; other multi-channel systems face the same design question, but our experiments do not establish cross-channel harm outside this setting.

Contributions. **(1) Positive-pair validity as measurement-operator exchangeability.** We give a design criterion that can be audited from the measurement model and target dependence, before pretraining and without labels. A controlled counterexample makes this distinction concrete: genuine, correctly co-registered measurements of the same state can be invalid positive equivalents when their operator difference carries downstream information. **(2) A reproducible transfer penalty and its recovery boundary.** The same-versus-cross intervention replicates across three SSL objectives and three independent pretraining seeds under grouped physical-unit inference, and LP-FT bounds the

failure to a supervision-dependent, partially recoverable transfer penalty rather than information destruction. **(3) Measurement-grounded constructive view design.** PhysAug builds positives from known sensing nuisances (probe registration, gain variation, and acquisition noise) and delivers replicated stability under the modeled registration nuisance, with a specificity control showing that the stability is tied to that nuisance rather than general robustness.

2 Background and Setup

ECT is an electromagnetic non-destructive testing method: probe coils induce eddy currents in a conductive tube wall, and flaws perturb the measured complex response. For a Transmit–Receive (T-R) array probe, the resulting data are calibrated 2D sensor maps indexed by circumferential coil row and axial probe position, not natural images. Frequency is physically meaningful: skin depth follows $\delta = 1/\sqrt{\pi f \mu \sigma}$, so lower frequencies penetrate deeper wall regions. Thin-wall tubing inspection routinely uses multiple frequencies, with the exact channels set by probe design, tube material, and wall thickness (Xiang et al., 2000). Here the axial channels AF1–AF4 are recorded at 400, 300, 200, and 100 kHz; this frequency-depth structure is the basis for our cross-frequency collapse hypothesis. Figure 2 summarizes the inspection geometry, calibrated probe-response coordinates, and 2D patch representation used throughout this work.

Each example is a patch $\mathbf{x} \in \mathbb{R}^{1\times 16\times 256}$ from the imaginary component of the AF2 300 kHz axial channel, with multi-label target $\mathbf{y} = (y_{\mathrm{ID}}, y_{\mathrm{OD}}, y_{\mathrm{TW}}) \in \{0,1\}^3$ derived from flaw boxes using a 30% minimum-overlap rule. AF2 is the anchor input channel because it is the balanced analysis channel for this tubing material and wall thickness, sensitive to both low-severity and through-wall flaws. The task is patch-level detection: a positive patch is one whose inspected axial–circumferential window overlaps a labeled flaw region, and no pixel-level mask or flaw size is predicted. Macro-AUPRC over ID, OD, and TW is the primary metric because positives are rare, with macro-AUROC secondary. View validity is therefore a measurement question rather than a visual-preprocessing one: after probe-response calibration, a flip, a large phase rotation, or a frequency exchange can change the physical variable a patch represents, not merely how it looks.

Experiments use calibration-standard tube scans from CAL14008 and CAL19003 under tube-level holdout: I003 is the fixed test tube (441 patches, 47 positives) and I004 is validation (357 patches, 39 positives), neither entering SSL pretraining or supervised training. The SSL pool contains 2,343 patches with 168 positives (7.2%).

Downstream adaptation uses the same positive-budget protocol for SSL probes, LP-FT, and the supervised CNN: $f = 10\%$ uses 55 of 168 training positives plus all 2,175 backgrounds, while $f = 100\%$ uses all 168 with the same background pool. Here f is the target effective positive sampling rate under the weighted sampler, not a fraction of unique labels (Appendix A); Figure 5 summarizes the split.

All methods use the same small ECTBackbone (four convolutional blocks, 64-dimensional representation, ≈24.6K backbone parameters). We hold the backbone fixed and compact by design. Holding architecture constant isolates view construction from model capacity; whether larger encoders interact with the effect is untested.

SSL pretraining is repeated independently with seeds $\{42, 7, 21\}$. Downstream seeds $\{42, 7, 21\}$ are nested within each pretrained encoder and are not independent SSL replications. For primary ranking analyses, their probabilities are averaged within encoder; metrics are computed per pretraining seed and summarized as mean $\pm$ sample SD across the three. Macro-AUPRC is primary because positives are rare; macro-AUROC is secondary. For a random ranker, the macro-AUPRC chance reference is the mean of the ID, OD, and TW prevalences, 0.047, and the sev20 subset has a separate macro chance reference of 0.0146 (Appendix B). The test-set any-defect prevalence, 0.107, is not a macro-AUPRC chance level.

CNN-PhysAug is the fair supervised reference: the identical backbone trained end-to-end with the same measurement-valid PhysAug transformations, over seeds $\{42, 7, 21\}$ at both budgets, with mean and sample SD across those seeds. Matching augmentation across the supervised and self-supervised arms keeps the comparison from being confounded by augmentation strength. Checkpoint selection uses clean I004 validation, which never enters pretraining or training.

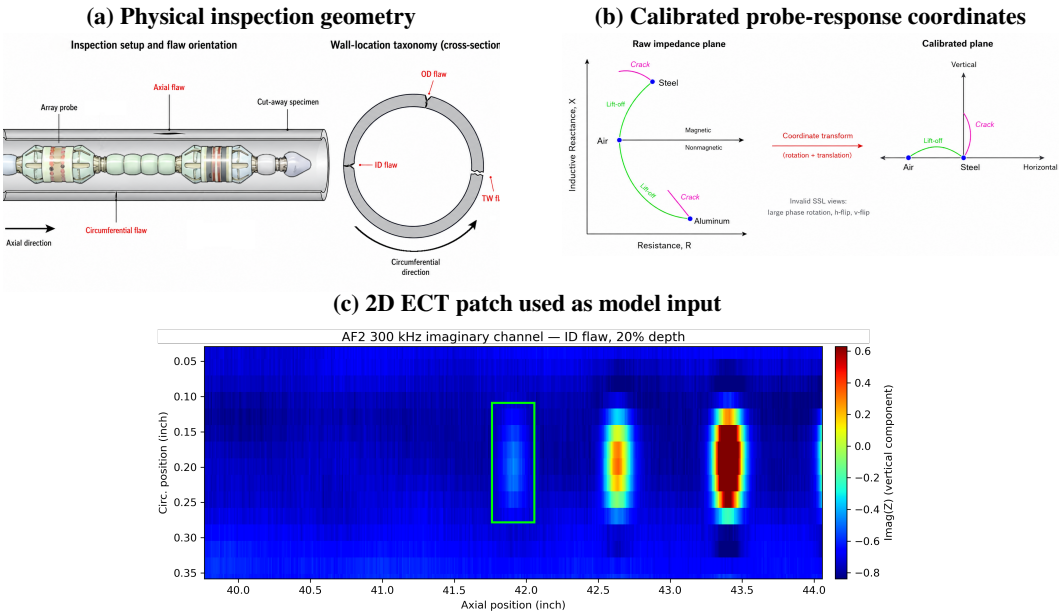


Figure 2: ECT measurement representation. **(a)** An array probe scans the tube axially while coil rows index circumference; flaws may be ID, OD, or TW. **(b)** Calibration makes lift-off approximately horizontal and flaw response vertical, so large phase rotations and flips alter physical semantics. **(c)** Model input: a calibrated 2D patch indexed by coil row and axial position.

Because ECT windows overlap axially, inference uses a grouped bootstrap over 18 positive flaw groups and 37 non-overlapping background blocks (five positive groups in sev20), not patches. Across 10,000 paired replicates we report percentile 95% CIs, two-sided plus-one p-values, and within-family BH correction (Appendix H); pretraining-seed SD and grouped CIs separately quantify stochastic learning and finite physical-unit uncertainty.

3 Measurement-Valid SSL View Design

A positive pair defines what the representation should ignore, so Table 1 records the measurement-model role assigned to each candidate view. The taxonomy specifies which physical variables the encoder is asked to treat as invariant; cross-frequency pairing is the experimentally isolated non-exchangeable intervention. PhysAug uses axial and bounded circumferential rolls, amplitude scaling, and noise to model known acquisition nuisances. Its Generic counterpart serves as an image-style baseline, replacing circumferential roll with horizontal/vertical flips and narrowing the amplitude range while otherwise matching PhysAug. In calibrated ECT coordinates (Figure 2(b)), these are not cosmetic changes: flips reverse probe travel or coil ordering, large phase rotations mix lift-off and flaw-response axes, and cross-frequency pairing changes the depth response itself.

Holding backbone and objective fixed while changing whether views respect these measurement coordinates isolates view validity from architecture and loss design. PhysAug is also the control that keeps nuisance construction fixed in the same- versus cross-frequency intervention, so the headline comparison isolates the channel equivalence being imposed. More generally, a calibrated physical view is admissible when the induced invariance preserves downstream-relevant physical information and inadmissible when it suppresses it. The audit uses measurement physics, not labels or an encoder.

Table 1: Physics-based view-design taxonomy. The cross-frequency row is isolated experimentally; the remaining statuses encode measurement-model assumptions rather than component-wise empirical ablations.

Transformation	**Design status**	**Physical rationale**
Axial roll (± 32 samples)	Modeled nuisance	probe start-position / axial registration
Circumferential roll (± 3 rows)	Modeled nuisance	bounded circumferential registration offset
Amplitude scaling ($U[0.75, 1.33]$)	Modeled nuisance	gain and sensitivity variation
Gaussian noise ($\sigma = 0.02$)	Modeled nuisance	electronics and acquisition noise
Horizontal flip	Non-preserving	reverses probe travel direction
Vertical flip	Non-preserving	reverses coil ordering; confounds ID/OD geometry
Large phase rotation ($> \lvert 15^\circ \rvert$)	Non-preserving	mixes calibrated lift-off and flaw-response coordinates
Cross-frequency pairing	Tested non-exchangeable	channels encode non-exchangeable depth response

Table 2: Cross-frequency penalty at $f = 10\%$, $\Delta = \mathrm{AUPRC}(\mathrm{AF2} + \mathrm{AF2}) - \mathrm{AUPRC}(\mathrm{AF2} + \mathrm{AFx})$. Seed-$\Delta$ SD is across pretraining seeds; CIs use grouped physical-unit inference and all nine comparisons survive BH correction. Same-frequency references are 0.7147/0.7022/0.6896 versus 0.047 random ranking.

Objective	**Pairing**	Δ	**Seed Δ SD**	**95% CI**	p	**BH** q
VICRegL	AF2+AF1	0.0653	0.0367	[0.0277, 0.1036]	0.0028	0.0042
VICRegL	AF2+AF3	0.0579	0.0179	[0.0211, 0.0938]	0.0040	0.0051
VICRegL	AF2+AF4	0.0702	0.0329	[0.0290, 0.1085]	0.0022	0.0040
DenseCL	AF2+AF1	0.0437	0.0032	[0.0124, 0.0775]	0.0052	0.0059
DenseCL	AF2+AF3	0.1048	0.0274	[0.0325, 0.2416]	0.0016	0.0036
DenseCL	AF2+AF4	0.1381	0.0388	[0.0576, 0.2467]	0.0008	0.0024
SimCLR	AF2+AF1	0.0676	0.0618	[0.0360, 0.0950]	0.0008	0.0024
SimCLR	AF2+AF3	0.0393	0.0316	[0.0210, 0.0645]	0.0002	0.0018
SimCLR	AF2+AF4	0.0323	0.0242	[0.0057, 0.0526]	0.0226	0.0226

4 RESULTS

4.1 THE CROSS-FREQUENCY POSITIVE-PAIR PENALTY

The central intervention changes only whether the positive pair is same-frequency (AF2+AF2) or cross-frequency (AF2+AF1/AF3/AF4), holding PhysAug, backbone, SSL objective, downstream protocol, and test set fixed. Because the augmentation policy is identical in both arms, any difference is attributable to the pairing and not to augmentation quality. At $f = 10\%$, the same-frequency encoder has higher macro-AUPRC in all nine objective-by-pairing comparisons under the three-pretraining-seed estimator. All nine grouped 95% confidence intervals exclude zero and all nine survive BH correction within the nine-comparison family (Table 2; Figure 3). The adverse direction also repeats for every observed pretraining seed in all nine comparisons; we treat this as directional replication, not additional significance tests.

Relative to the level they are measured from, these effects are large. Same-frequency encoders reach 0.7147 (VICRegL), 0.7022 (DenseCL), and 0.6896 (SimCLR) macro-AUPRC against a random-ranking reference of 0.047 (Table 3), so the nine contrasts correspond to relative reductions of roughly 5% to 20%. Cross-frequency pairing degrades an encoder that is otherwise working well; it does not rescue a failing one.

Nor is the penalty specific to one objective family: it spans SimCLR instance discrimination, DenseCL dense correspondence, and VICRegL variance–covariance regularization, whose variance term explicitly discourages dimensional collapse. Avoiding trivial collapse therefore does not guarantee preservation of operator-dependent information that the view distribution encourages the encoder to ignore.

The sev20 incipient-flaw stratum rests on only five independent positive physical groups, so we treat it as low-powered supporting evidence: seven of nine linear-probe sev20 contrasts are BH-supported (Appendix C), with DenseCL AF2+AF1 and SimCLR AF2+AF4 unresolved.

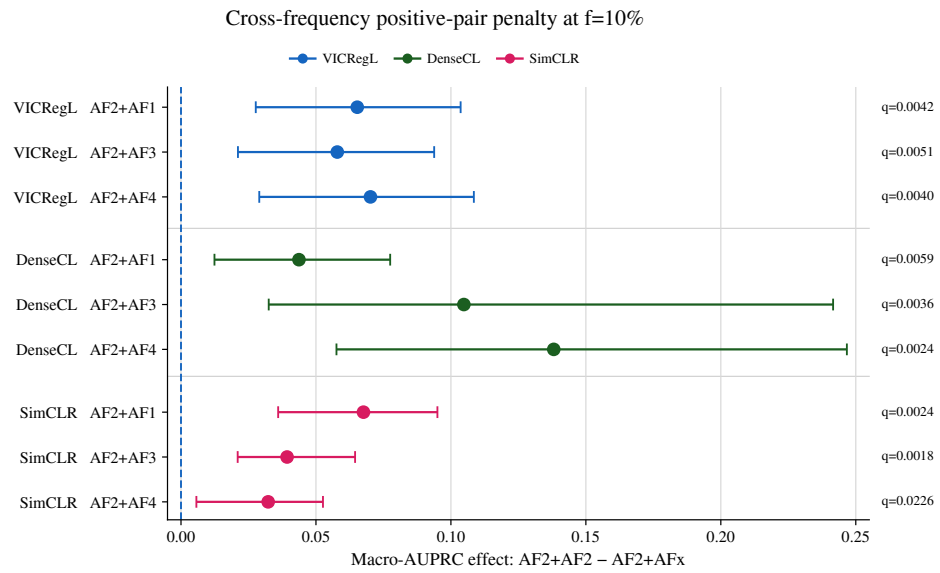


Figure 3: Headline $f = 10\%$ same-minus-cross macro-AUPRC effects. Points are three-pretraining-seed means; intervals are grouped-bootstrap 95% CIs over physical test units; q-values are BH-adjusted. Effects are approximately 5%–20% of the matched same-frequency level.

4.2 Representation Signature of the Imposed Invariance

A pretrained encoder should carry a signature of the invariance it was trained to satisfy. For a given encoder, r_{neg} is the mean cosine similarity over eligible unrelated cross-view pairs. A pair (i, j) is eligible only if $i \neq j$ and the patches come from different source files or have non-intersecting rectangles within a source:

$$r_{\text{neg}} = \frac{1}{|\mathcal{U}|} \sum_{(i,j)\in\mathcal{U}} \cos\Big(f(v_i^{(1)}), f(v_j^{(2)})\Big) . \tag{1}$$

Computation is label-free and uses all 2,343 SSL-training patches. A high value means that unrelated patches drawn from the two view channels have become geometrically hard to tell apart. Absolute values are not comparable across SSL objectives, so the diagnostic is defined as a within-objective contrast: each cross-frequency encoder is compared with the same-objective, same-pretraining-seed AF2+AF2 reference,

$$\Delta r_{\text{neg}} = r_{\text{neg}}(\text{AF2} + \text{AFx}) - r_{\text{neg}}(\text{AF2} + \text{AF2}), \tag{2}$$

alongside the drop in the positive-versus-unrelated separation margin under the model's own view protocol. Every one of the 27 evaluated cross-frequency objective/pairing/pretraining-seed configurations has positive Δr_{neg} and a smaller separation margin than its matched reference (Appendix E).

Because these encoders are trained toward cross-view invariance, elevated unrelated-pair similarity partly confirms that the intervention took effect. Two features make the quantity informative beyond that confirmation. Eligibility excludes related pairs, so the statistic is not the training objective read back on its own training pairs. The separation margin also falls, indicating loss of sample-specific cross-view structure rather than uniform similarity inflation. We therefore read Δr_{neg} as confirmatory corroboration of a known intervention. Establishing it as a prospective screen would require collapses arising independently of a designed pairing, which we do not test here.

4.3 The Recoverability Boundary

Controlled LP-FT asks how much of the deficit survives stronger downstream adaptation. The mean same-minus-cross macro-AUPRC gap falls from 0.0688 under the frozen linear probe to 0.0424 after LP-FT at $f = 10\%$, and to 0.0222 at $f = 100\%$. At $f = 10\%$ all nine contrasts remain BH-supported and the adverse direction occurs in all 27 objective/pairing/pretraining-seed comparisons; at $f = 100\%$ no contrast remains BH-supported. The complete tables are in Appendix D.

Table 3: Same-frequency clean macro-AUPRC and fair supervised reference. SSL rows report three-pretraining-seed means plus paired PhysAug-minus-Generic inference; CNN-PhysAug is mean $\pm$ sample SD across three training seeds.

Method	Setting	Macro-AUPRC	Δ	Seed Δ SD	95% CI	BH q
VICRegL	Generic, $f = 10\%$	0.6802	—	—	—	—
VICRegL	PhysAug, $f = 10\%$	0.7147	+0.0345	0.0203	[-0.0046, 0.1046]	0.0909
DenseCL	Generic, $f = 10\%$	0.6531	—	—	—	—
DenseCL	PhysAug, $f = 10\%$	0.7022	+0.0490	0.0053	[0.0009, 0.1053]	0.0696
SimCLR	Generic, $f = 10\%$	0.5359	—	—	—	—
SimCLR	PhysAug, $f = 10\%$	0.6896	+0.1537	0.0885	[0.0622, 0.2123]	0.0006
CNN-PhysAug	$f = 10\%$	0.6235 ± 0.0790	—	—	—	—
CNN-PhysAug	$f = 100\%$	0.6294 ± 0.0534	—	—	—	—

That gradient is what bounds the claim: cross-frequency collapse is a supervision-dependent transferability failure that stronger adaptation partially repairs, not an irreversible destruction of the depth cue. The full-label result combines a smaller point estimate with a loss of statistical resolution, and we do not read it as evidence that the effect is zero.

4.4 Measurement-Grounded View Construction

If invalid invariances can be identified from the measurement model, valid ones can be built from it. PhysAug makes known registration, gain, and acquisition nuisances invariant while retaining frequency identity. On clean data at $f = 10\%$, PhysAug has higher mean macro-AUPRC than the Generic policy for all three objectives, though the inferential strength is objective-dependent: VICRegL is positive but unresolved, DenseCL's interval excludes zero without surviving BH correction, and SimCLR is BH-supported (Table 3).

Sharper evidence comes from stability under the nuisance PhysAug actually models. Across the circumferential-shift sweep, the macro-AUPRC range of the three-seed mean curve is 0.0176 versus 0.0680 for VICRegL PhysAug/Generic, 0.0196 versus 0.1575 for DenseCL, and 0.0312 versus 0.0679 for SimCLR (Figure 4a); Generic variation is therefore approximately $3.9\times$, $8.0\times$, and $2.2\times$ larger within objective. At the individual pretraining-seed level PhysAug has the smaller range in all 9/9 objective-by-seed comparisons. Points along a sweep are correlated and we read them descriptively.

To test whether that stability is specific, we resample axial speed. Its PhysAug/Generic ranges are 0.0324/0.0435 for VICRegL, 0.0357/0.0342 for DenseCL, and 0.0513/0.0482 for SimCLR: the large separation seen under circumferential shift does not reproduce, while PhysAug retains its higher absolute clean-performance offset. Stability is therefore tied to the modeled nuisance, not general out-of-distribution robustness. PhysAug is evaluated as a composite policy, so no individual component is credited with the effect.

A frozen-operating-point analysis corroborates the pattern under a different aggregation hierarchy: thresholds are selected per downstream model on clean validation and frozen under shift. PhysAug exceeds Generic in all 63 objective-by-seed-by-shift Macro-F_3 comparisons; these correlated directions are descriptive, not 63 independent tests (Appendix G).

4.5 Fair Supervised Reference

All three PhysAug SSL means exceed CNN-PhysAug at $f = 100\%$ (Table 3), but the advantage is not statistically established. Grouped SSL-minus-CNN differences are $+0.0853$ for VICRegL (95% CI $[-0.1137, 0.2118]$), $+0.0728$ for DenseCL ($[-0.1282, 0.2030]$), and $+0.0602$ for SimCLR ($[-0.1507, 0.1947]$), all with BH $q = 0.5201$; at matched $f = 10\%$ the intervals again include zero (Appendix D). CNN-PhysAug itself gains only $+0.0059$ macro-AUPRC between the two budgets ($p = 0.888$); the experiment does not isolate whether this limited change reflects optimization, capacity, sampling, or label-regime effects. Our cross-frequency result is an SSL-versus-SSL comparison and does not depend on this baseline.

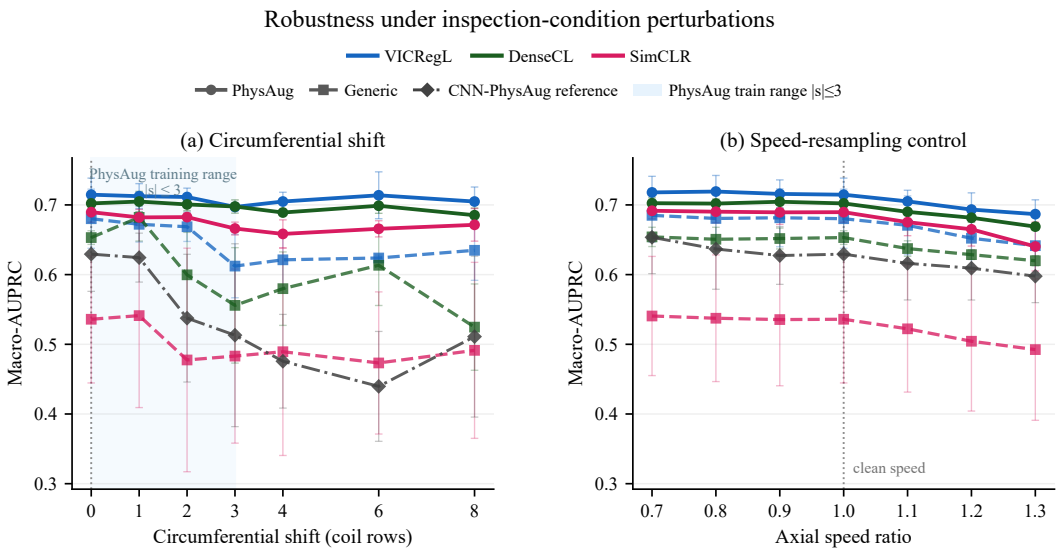


Figure 4: Targeted threshold-free robustness across three pretraining seeds. **(a)** PhysAug is markedly flatter than Generic under the modeled circumferential-registration nuisance. **(b)** Speed resampling yields much more similar within-objective sensitivity, while PhysAug retains its clean-performance offset. CNN-PhysAug $f = 100\%$ is contextual; sweep points are correlated and descriptive.

5 Related Work

View selection and harmful invariance in SSL. Modern contrastive SSL depends strongly on the view distribution (Chen et al., 2020). InfoMin-style analyses ask which shared information should be preserved across views (Tian et al., 2020), while subsequent work shows that invariance to transformations relevant to a downstream task can hurt transfer (Xiao et al., 2021) and that contrastive learning can suppress class-relevant or harder features (Xue et al., 2023). We build directly on this established lineage. Our contribution is a controlled failure mode in which the candidate positives are not synthetic transformations at all: they are physically real, correctly co-registered measurements of the same state, with identical labels, yet their channel-dependent measurement operators are non-exchangeable for the downstream variable. The correspondence is not approximately right; it is exactly right, and the invariance is still harmful.

Domain-informed positive construction in scientific sensing. Scientific and medical SSL commonly replaces generic image assumptions with domain knowledge. Medical-image and physiological approaches use domain-specific transformations or acquisition structure (Azizi et al., 2021; Kiyasseh et al., 2021); MedAug uses patient metadata to select positive chest-X-ray pairs (Vu et al., 2021). Remote-sensing work similarly exploits known spatiotemporal invariances, as in Seasonal Contrast (SeCo) (Mañas et al., 2021), or constructs views using atmospheric correction for multispectral imagery (Patnala et al., 2023). PhysAug belongs to this constructive lineage. These methods search for views that are semantically equivalent, so their invalid pairs arise from a mismatch: seasonal change that matters downstream, or a pathology that silently differs across studies. Our intervention asks the complementary question, when a physically natural correspondence should *not* be promoted to equivalence. The paired ECT channels are already aligned and share the same inspected state, and there is no mismatch to detect; what changes is the operator, and with it the depth-sensitive measurement content.

Multiview, unique information, and equivariance. Multimodal SSL often aligns paired observations because they share semantics (Patrick et al., 2021; Radford et al., 2021; Girdhar et al., 2023), while FactorCL models task-relevant view-unique information and SIE/E-SSL preserve equivariant structure rather than enforcing invariance indiscriminately (Liang et al., 2023; Garrido et al., 2023; Dangovski et al., 2022). Our setting isolates a distinct failure mode: the problematic views are

genuine, exactly co-registered measurements of the same physical state, yet their known measurement operators are non-exchangeable for the target.

Machine learning for NDT and ECT. Work in NDT/ECT typically interprets inspection signals through sensing physics and supervised or feature-engineered pipelines (Taheri et al., 2022; Munir et al., 2025; Xiang et al., 2000; Zhu et al., 2019; Park et al., 2019; Sahu et al., 2025); SSL is emerging in other NDT modalities (McKnight et al., 2025; Hoxha et al., 2025). We do not propose another ECT detector; here ECT is the controlled sensing system for studying which physical measurements should define SSL invariance.

6 Discussion

Correspondence is not equivalence. The cross-frequency comparison is the primary inferential result: the views are co-registered measurements of the same physical region, with the same labels, yet all nine low-label same-minus-cross macro-AUPRC effects are BH-supported, corresponding to relative reductions of 5% to 20%. Correct physical correspondence can itself define a harmful invariance when the channel-dependent information is task relevant, so view validity cannot be settled by checking that two measurements describe the same object.

The transfer penalty is supervision dependent. Under LP-FT with scarce labels the mean same-minus-cross gap falls from 0.0688 to 0.0424, where all nine penalties remain BH-supported, and to 0.0222 with full labels, where none does. Stronger supervised adaptation compensates for part of a pretraining-induced transfer penalty, bounding "collapse" to a partially recoverable transferability failure. It does not follow that the full-label effects are exactly zero.

The same measurement model that invalidates a view can construct valid ones. PhysAug makes known registration, gain, and acquisition nuisances invariant while retaining frequency identity, and it is markedly more stable than the generic policy under exactly the nuisance it models, with the speed-resampling control showing that the stability is nuisance-specific rather than general. This suggests a practical heuristic: audit before pairing. If the operator difference is nuisance for the target, pairing can learn the intended invariance; if it carries target signal, preserve it instead. Structured alternatives such as conditioning, shared–private representations, or fusion could do so without full equivalence, but fair comparison—especially for fusion—requires matched representation dimension and model capacity.

Objective-matched diagnostics reveal a consistent representation signature. Across all 27 cross-frequency objective–pairing–pretraining-seed settings, Δr_{neg} is positive and the positive-versus-unrelated separation margin decreases relative to the matched same-objective, same-seed AF2+AF2 reference. Because absolute r_{neg} scales differ across objectives, these matched contrasts provide the relevant representation-level comparison. We use them as supporting evidence that cross-frequency pairing changes representation geometry in the expected direction, rather than as a universal threshold or calibrated predictor of downstream performance.

Limitations. Evidence comes from one held-out tube/acquisition setting, AF2 as the anchor frequency, and one probe/geometry regime; three pretraining seeds establish stochastic rather than condition-level robustness. Our claim concerns equivalence-based pairing rather than multi-frequency representation learning generally; structured frequency-aware alternatives are not evaluated here. Full-label LP-FT shows that the penalty is partially recoverable with stronger supervision. Deployment in safety-relevant inspection still requires application-specific validation and human oversight.

Conclusion. Multi-frequency ECT shows that physical co-registration is insufficient for SSL positive-pair equivalence when channel-dependent measurement operators preserve different task-relevant information. These results motivate a broader design principle for multi-channel scientific sensing: *enforce invariance only when operator-dependent variation is nuisance for the downstream target; otherwise preserve operator-specific information through conditioning, shared–private structure, fusion, or another structured representation.* Together, the controlled ECT results establish a reproducible operator-dependent positive-pair failure mode and motivate a directly testable criterion for auditing view design in other multi-channel scientific sensors.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and presentation, including the preparation and refinement of some figures and visual materials. The authors performed the experiments, independently verified all reported numerical results and scientific claims against the underlying records, and take responsibility for the final content.

ETHICS STATEMENT

This study uses non-human, non-personal eddy-current calibration data and involves no human subjects or personally identifiable information. The studied application is nuclear steam-generator tube inspection, a safety-relevant NDT setting; application-specific validation and human oversight remain necessary before deployment.

REPRODUCIBILITY STATEMENT

Sections 2–4 and Appendices A–H document the data splits, view construction, objective-specific training settings, nested seed hierarchy, supervised and LP-FT protocols, grouped physical-unit inference, and robustness evaluation used for the reported results. The appendix also separates confirmatory analyses from descriptive robustness checks and specifies the physical resampling units and multiplicity families. The controlled calibration-standard steam-generator-tube ECT records are not released publicly with this submission, so the manuscript does not claim full open-data reproducibility.

REFERENCES

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3478–3488, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegL: Self-supervised learning of local visual features. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0640.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2022.

Quentin Garrido, Laurent Najman, and Yann LeCun. Self-supervised learning of split invariant equivariant representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10975–10996. PMLR, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Ejup Hoxha, Jinglun Feng, Agnimitra Sengupta, David Kirakosian, Yang He, Bo Shang, Ardian Gjinofci, and Jizhong Xiao. Contrastive learning for robust defect mapping in concrete slabs using impact echo. *Construction and Building Materials*, 461:139829, 2025. doi: 10.1016/j.conbuildmat.2024.139829.

Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5606–5615. PMLR, 2021.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1430.

Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vázquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9414–9423, 2021.

Shaun McKnight, Vedran Tunukovic, Amine Hifi, S. Gareth Pierce, Ehsan Mohseni, Charles N. MacLeod, and Tom O'Hare. Three-dimensional ultrasonic self supervised segmentation. *Engineering Applications of Artificial Intelligence*, 154:110870, 2025. doi: 10.1016/j.engappai.2025.110870.

Nauman Munir, Jingyuan Huang, Chak-Nam Wong, and Sung-Jin Song. Machine learning based eddy current testing: A review. *Results in Engineering*, 25:103724, 2025. doi: 10.1016/j.rineng.2024.103724.

Jinhyun Park, Seong-Jin Han, Nauman Munir, Yun-Taek Yeom, Sung-Jin Song, Hak-Joon Kim, and Segon Kwon. MRPC eddy current flaw classification in tubes using deep neural networks. *Nuclear Engineering and Technology*, 51(7):1784–1790, 2019. doi: 10.1016/j.net.2019.05.011.

Ankit Patnala, Scarlet Stadtler, Martin G. Schultz, and Jürgen Gall. Generating views using atmospheric correction for contrastive self-supervised learning of multispectral images. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. doi: 10.1109/LGRS.2023.3274493.

Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9577–9587, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Tikesh Kumar Sahu, S. Thirunavukkarasu, and Anish Kumar. Intelligent flaw detection in eddy current inspection data through machine learning model. *Journal of Nondestructive Evaluation*, 44:88, 2025. doi: 10.1007/s10921-025-01229-2.

Hossein Taheri, Maria Gonzalez Bocanegra, and Mohammad Taheri. Artificial intelligence, machine learning and smart technologies for nondestructive evaluation. *Sensors*, 22(11):4055, 2022. doi: 10.3390/s22114055.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839, 2020.

Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y. Ng, and Pranav Rajpurkar. MedAug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 755–769. PMLR, 2021.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3024–3033, 2021.

P. Xiang, S. Ramakrishnan, X. Cai, P. Ramuhalli, R. Polikar, S. S. Udpa, and L. Udpa. Automated analysis of rotating probe multi-frequency eddy current data from steam generator tubes. *International Journal of Applied Electromagnetics and Mechanics*, 12(3–4):151–164, 2000. doi: 10.3233/JAE-2000-216.

Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021.

Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? On the role of simplicity bias in class collapse and feature suppression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38938–38970. PMLR, 2023.

Peipei Zhu, Yuhua Cheng, Portia Banerjee, Antonello Tamburrino, and Yiming Deng. A novel machine learning model for eddy current testing with uncertainty. *NDT & E International*, 101: 104–112, 2019. doi: 10.1016/j.ndteint.2018.09.010.

A Experimental Details and Reproducibility

Seed hierarchy and estimator. SSL pretraining is repeated independently with seeds $\{42, 7, 21\}$. For every pretrained encoder, downstream seeds $\{42, 7, 21\}$ are nested classifier/adaptation runs and are not independent representation-learning replications. For threshold-free ranking analyses, downstream probability matrices are aligned and averaged *within* the pretrained encoder; macro-AUPRC/macro-AUROC is computed separately for each pretraining seed and summarized as mean $\pm$ sample SD across the three pretraining seeds. CNN-PhysAug is trained independently with seeds $\{42, 7, 21\}$. Figure 7 uses the distinct threshold hierarchy specified in Appendix G.

Table 4: Objective-specific SSL pretraining configurations. All objectives use the same four-block ECTBackbone with channels $[8, 16, 32, 64]$, 64-dimensional pooled representation, 300 pretraining epochs, and a 10-epoch warmup followed by cosine decay.

Objective	Optimizer	LR	WD	Batch	Projection/objective	Key constants
VICRegL	AdamW	3×10^{-4}	10^{-4}	128	global/local projection dim. 256	$\lambda = 25, \mu = 25, \nu = 1, \alpha = 0.5$
DenseCL	Adam	3×10^{-4}	10^{-6}	32	global/dense $64 \to 64 \to 64$	queue 512, $\tau = 0.2$, momentum 0.99, dense wt. 0.5
SimCLR	Adam	3×10^{-4}	10^{-6}	128	$64 \to 256 \to 128$	$\tau = 0.07$

Table 5: Augmentation policies used for SSL view construction. Each listed transform is applied independently with the shown probability. PhysAug is held fixed across the same- versus cross-frequency intervention; no component-level causal ablation is claimed.

Transform	PhysAug	Generic
Gaussian noise	$\sigma = 0.02, p = 0.5$	$\sigma = 0.02, p = 0.5$
Axial circular roll	$\pm32, p = 0.5$	$\pm32, p = 0.5$
Circumferential circular roll	$\pm3, p = 0.5$	excluded
Horizontal flip	excluded	$p = 0.5$
Vertical flip	excluded	$p = 0.3$
Amplitude scale	$U[0.75, 1.33], p = 0.5$	$U[0.80, 1.20], p = 0.5$

For linear probes, LP-FT, and CNN-PhysAug, checkpoint selection uses $(\mathrm{AUPRC}_{ID} + \mathrm{AUPRC}_{OD})/2$ on I004 validation. The CNN-PhysAug protocol requires at least 20 epochs before checkpoint eligibility and uses early-stopping patience 15. The shared compact backbone has approximately 24.6K parameters (24,819 in the supervised CNN implementation).

Positive-budget split definition. The split parameter f denotes the target effective positive sampling rate in downstream supervised training, not a random fraction of unique patch labels. With the full

Table 6: Downstream and supervised protocols. Linear-probe and LP-FT values are shared across the three SSL objectives in the replicated pipelines. CNN-PhysAug is listed separately because its supervised training/checkpoint schedule is not the SSL pretraining schedule.

Stage	Optimization	Training details	Selection/regularization
Frozen linear probe	AdamW, LR 10^{-3}, WD 10^{-4}	50 epochs, batch 32, warmup 5	label smoothing 0.1; positive weights ID/OD/TW $= 3/3/5$; sampler ratio 4
LP-FT stage 1	AdamW, head LR 10^{-3}, WD 10^{-4}	50 epochs, batch 32, encoder frozen	same label/sampler setup; warmup 5
LP-FT stage 2	AdamW, head LR 10^{-5}, WD 10^{-4}	50 epochs, batch 32, encoder unfrozen	LLRD 0.70; BN running statistics frozen; warmup 5
CNN-PhysAug	AdamW, LR 10^{-3}, WD 10^{-4}	max 80 epochs, batch 32, warmup 5	PhysAug; smoothing 0.1; sampler ratio 4; positive weights $3/3/5$

background pool $n_{\text{neg}} = 2{,}175$ and positive oversampling ratio $r = 4.0$, the number of selected positive patches is

$$n_{\text{pos}}(f) = \min\left(168, \left\lceil \frac{f\, n_{\text{neg}}}{r} \right\rceil\right).$$

Thus $f = 10\%$ selects 55 unique positives and, under the weighted sampler, yields an effective positive draw fraction $4 \times 55/(4 \times 55 + 2175) = 9.2\%$. The $f = 100\%$ setting is capped at all 168 positives, giving $4 \times 168/(4 \times 168 + 2175) = 23.6\%$. The full background pool is retained in both settings, and the same split/sampler definition is used for SSL linear probes, LP-FT, and CNN-PhysAug. The 55-versus-168 cross-budget comparison is descriptive.

Reproducibility/provenance. The linear analysis contains 15 SSL configurations $\times$ 3 pretraining seeds $\times$ 3 downstream seeds $\times$ 2 label budgets $= 270$ aligned prediction runs, and all formal ranking comparisons use the same 441 I003 rows.

B Dataset Composition, Prevalence, and Incipient-Flaw Support

The split is tube-level: CAL14008 I003 is the fixed held-out test set and CAL14008 I004 is validation; neither enters SSL pretraining or supervised training. Table 7 distinguishes positive patches from per-class marginal counts, and Figure 5 summarizes the split structure, the positive-budget protocol, and the test-set severity distribution. Random-ranking macro-AUPRC is the mean per-class prevalence, not the any-defect prevalence.

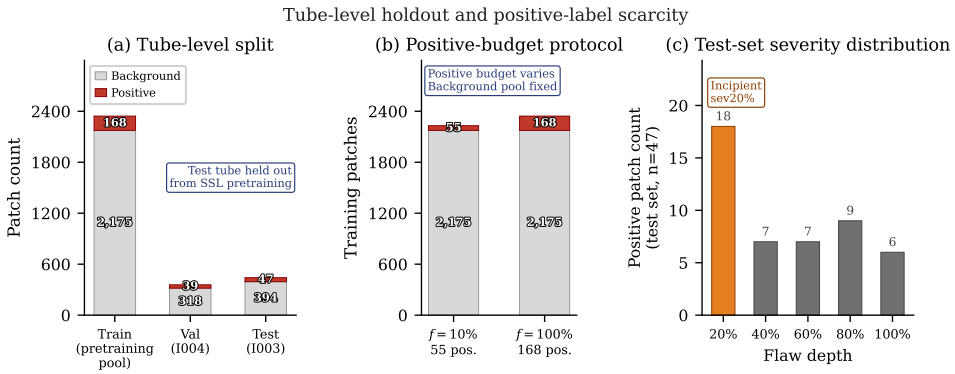


Figure 5: Tube-level holdout and positive-label scarcity. **(a)** All patches from each physical tube belong to exactly one split. **(b)** SSL pretraining uses the full training pool without labels; downstream adaptation uses a positive-budget protocol in which $f = 10\%$ selects 55 positives with the full background pool and $f = 100\%$ uses all 168 training positives with the same background pool. **(c)** The held-out test positives are severity-imbalanced, with 20% wall-loss incipient flaws forming the critical low-amplitude regime.

Table 7: Patch and class counts. ID/OD/TW are marginal counts and therefore need not sum to the number of positive patches. The held-out test any-defect prevalence is $47/441 = 0.107$, whereas the correct overall random-ranking macro-AUPRC reference is 0.046863 (reported as 0.047).

Split	Total	Positive	Background	ID	OD	TW
Train / SSL pool	2,343	168	2,175	82	83	39
Validation (I004)	357	39	318	17	24	13
Test (I003)	441	47	394	25	25	12

The test severity counts are 18/7/7/9/6 positive patches at 20/40/60/80/100% wall loss. The sev20 analysis uses those 18 positive patches but only **five independent positive physical groups** (two ID and three OD). Its separate random-ranking macro-AUPRC reference is 0.014563 (reported as 0.0146), and all sev20 inference is therefore treated as secondary and low-powered.

C Incipient-Flaw (sev20) Supporting Analysis

Table 8: sev20 cross-frequency supporting analysis at $f = 10\%$. Δ is AF2+AF2 minus AF2+AFx sev20 macro-AUPRC. The grouped bootstrap uses five independent sev20 positive physical groups and 10,000 valid replicates; BH correction is within the nine-comparison sev20 family. Seven of nine comparisons are BH-supported.

Objective	Pairing	Δ	95% CI	p	BH q
VICRegL	AF2+AF1	0.1813	[0.1058, 0.2377]	0.0002	0.0004
VICRegL	AF2+AF3	0.1790	[0.0907, 0.2363]	0.0004	0.0005
VICRegL	AF2+AF4	0.1955	[0.0998, 0.2447]	0.0004	0.0005
DenseCL	AF2+AF1	0.1033	[-0.0014, 0.1791]	0.0534	0.0572
DenseCL	AF2+AF3	0.1479	[0.0500, 0.1993]	0.0002	0.0004
DenseCL	AF2+AF4	0.1888	[0.0492, 0.2748]	0.0002	0.0004
SimCLR	AF2+AF1	0.0794	[0.0206, 0.1366]	0.0002	0.0004
SimCLR	AF2+AF3	0.0549	[0.0178, 0.1067]	0.0002	0.0004
SimCLR	AF2+AF4	0.0490	[-0.0009, 0.0968]	0.0572	0.0572

DenseCL AF2+AF1 and SimCLR AF2+AF4 are unresolved after BH correction. The five-group limitation cautions inference in both directions and prevents sev20 from carrying the headline conclusion.

D Extended Results

All linear and LP-FT ranking tables in this section use downstream probability consensus within pretrained encoder, metric computation per pretraining seed, and mean/sample-SD aggregation across pretraining seeds $\{42, 7, 21\}$. Grouped intervals use the physical-unit grouped bootstrap of Appendix H.

D.1 Fair Supervised Comparison and Secondary Macro-AUROC

Table 9: Fair augmentation-matched SSL versus CNN-PhysAug macro-AUPRC. Positive differences favor SSL, but every grouped primary AUPRC interval includes zero. BH correction is within each declared three-comparison family.

Comparison	SSL	SSL AUPRC	CNN AUPRC	Δ	95% CI	BH q
Matched $f = 10\%$	VICRegL	0.7147	0.6235	+0.0912	[-0.0330, 0.1954]	0.3306
Matched $f = 10\%$	DenseCL	0.7022	0.6235	+0.0787	[-0.0444, 0.1863]	0.3306
Matched $f = 10\%$	SimCLR	0.6896	0.6235	+0.0661	[-0.0660, 0.1750]	0.3306
SSL $f = 10\%$ vs CNN $f = 100\%$	VICRegL	0.7147	0.6294	+0.0853	[-0.1137, 0.2118]	0.5201
SSL $f = 10\%$ vs CNN $f = 100\%$	DenseCL	0.7022	0.6294	+0.0728	[-0.1282, 0.2030]	0.5201
SSL $f = 10\%$ vs CNN $f = 100\%$	SimCLR	0.6896	0.6294	+0.0602	[-0.1507, 0.1947]	0.5201

Macro-AUROC is secondary because positives are rare and macro-AUPRC was selected as the primary metric. For SSL $f = 10\%$ versus CNN-PhysAug $f = 100\%$, the secondary macro-AUROC differences are +0.0876 for VICRegL (95% CI [0.0308,0.1343], BH $q = 0.0036$), +0.0724 for DenseCL ([0.0167,0.1245], $q = 0.0186$), and +0.0691 for SimCLR ([0.0011,0.1218], $q = 0.0478$). These results are complementary ranking evidence and do not reinstate a primary macro-AUPRC superiority claim.

D.2 Linear-Probe $f = 100\%$ Cross-Frequency Persistence

The penalty persists when the downstream budget is not scarce. Under the frozen linear probe at $f = 100\%$, seven of nine comparisons are nominally supported and five of nine remain BH-supported (Table 10).

Table 10: Secondary $f = 100\%$ linear-probe cross-frequency persistence. Δ is AF2+AF2 minus AF2+AFx macro-AUPRC. Seven of nine comparisons are nominally supported and five of nine remain BH-supported.

Objective	Pairing	Δ	95% CI	p	BH q
VICRegL	AF2+AF1	0.0455	[0.0110, 0.0840]	0.0052	0.0162
VICRegL	AF2+AF3	0.0377	[0.0017, 0.0726]	0.0388	0.0558
VICRegL	AF2+AF4	0.0498	[0.0074, 0.0920]	0.0180	0.0324
DenseCL	AF2+AF1	0.0380	[0.0008, 0.0820]	0.0434	0.0558
DenseCL	AF2+AF3	0.0975	[0.0196, 0.2547]	0.0076	0.0171
DenseCL	AF2+AF4	0.1131	[0.0363, 0.2278]	0.0032	0.0162
SimCLR	AF2+AF1	0.0625	[0.0211, 0.0961]	0.0054	0.0162
SimCLR	AF2+AF3	0.0253	[-0.0014, 0.0672]	0.0628	0.0707
SimCLR	AF2+AF4	0.0095	[-0.0207, 0.0389]	0.4078	0.4078

D.3 LP-FT Recoverability

Table 11: Same-frequency LP-FT macro-AUPRC at $f = 10\%$. Δ is PhysAug minus Generic.

Objective	PhysAug	Generic	Δ	95% CI	p	BH q
VICRegL	0.712 ± 0.013	0.682 ± 0.023	+0.0305	[-0.0011, 0.0922]	0.0582	0.0582
DenseCL	0.713 ± 0.007	0.644 ± 0.017	+0.0690	[0.0170, 0.1245]	0.0092	0.0138
SimCLR	0.698 ± 0.008	0.658 ± 0.050	+0.0403	[0.0158, 0.0743]	0.0036	0.0108

Table 12: Cross-frequency LP-FT at $f = 10\%$. Δ is AF2+AF2 minus the indicated cross-frequency pairing; all nine comparisons are BH-supported.

Objective	Pairing	AF2+AF2	Cross	Δ	95% CI	p	BH q
VICRegL	AF2+AF1	0.712 ± 0.013	0.669 ± 0.008	+0.0433	[0.0200, 0.0710]	0.0002	0.0018
VICRegL	AF2+AF3	0.712 ± 0.013	0.685 ± 0.008	+0.0274	[0.0008, 0.0576]	0.0430	0.0430
VICRegL	AF2+AF4	0.712 ± 0.013	0.668 ± 0.019	+0.0437	[0.0104, 0.0805]	0.0066	0.0093
DenseCL	AF2+AF1	0.713 ± 0.007	0.667 ± 0.012	+0.0455	[0.0129, 0.0800]	0.0018	0.0081
DenseCL	AF2+AF3	0.713 ± 0.007	0.665 ± 0.028	+0.0479	[0.0130, 0.0912]	0.0048	0.0086
DenseCL	AF2+AF4	0.713 ± 0.007	0.642 ± 0.031	+0.0706	[0.0188, 0.1290]	0.0072	0.0093
SimCLR	AF2+AF1	0.698 ± 0.008	0.663 ± 0.023	+0.0358	[0.0138, 0.0655]	0.0046	0.0086
SimCLR	AF2+AF3	0.698 ± 0.008	0.662 ± 0.031	+0.0367	[0.0147, 0.0651]	0.0032	0.0086
SimCLR	AF2+AF4	0.698 ± 0.008	0.668 ± 0.029	+0.0306	[0.0030, 0.0661]	0.0314	0.0353

The average same-minus-cross macro-AUPRC gap falls from 0.0688 (frozen linear, $f = 10\%$) to 0.0424 after LP-FT; at $f = 100\%$ it falls to 0.0222 and no comparison remains BH-supported. The gradient establishes partial recoverability rather than irreversible information destruction, and the loss of resolution at $f = 100\%$ is not read as a zero effect.

E Objective-Matched r_{neg} Diagnostic

r_{neg} is computed on the full 2,343-patch unlabeled SSL-training pool from L2-normalized 64-dimensional pooled backbone representations. Unrelated pair eligibility requires $i \neq j$ and either

Table 13: Cross-frequency LP-FT at $f = 100\%$. With stronger/full-label adaptation, no comparison remains BH-supported.

Objective	Pairing	AF2+AF2	Cross	Δ	95% CI	p	BH q
VICRegL	AF2+AF1	0.702 ± 0.011	0.683 ± 0.003	+0.0189	[-0.0008, 0.0434]	0.0592	0.2061
VICRegL	AF2+AF3	0.702 ± 0.011	0.700 ± 0.010	+0.0023	[-0.0219, 0.0284]	0.7741	0.7741
VICRegL	AF2+AF4	0.702 ± 0.011	0.678 ± 0.023	+0.0238	[-0.0097, 0.0666]	0.1603	0.2061
DenseCL	AF2+AF1	0.709 ± 0.002	0.677 ± 0.010	+0.0323	[-0.0047, 0.0760]	0.0965	0.2061
DenseCL	AF2+AF3	0.709 ± 0.002	0.676 ± 0.022	+0.0340	[-0.0116, 0.1011]	0.1549	0.2061
DenseCL	AF2+AF4	0.709 ± 0.002	0.673 ± 0.039	+0.0363	[-0.0099, 0.0979]	0.1379	0.2061
SimCLR	AF2+AF1	0.695 ± 0.011	0.673 ± 0.024	+0.0217	[-0.0016, 0.0457]	0.0640	0.2061
SimCLR	AF2+AF3	0.695 ± 0.011	0.679 ± 0.021	+0.0161	[-0.0056, 0.0397]	0.1335	0.2061
SimCLR	AF2+AF4	0.695 ± 0.011	0.681 ± 0.025	+0.0144	[-0.0100, 0.0464]	0.2365	0.2661

different source files or non-intersecting two-dimensional patch rectangles within a source. Each cross-frequency encoder is compared with the same-objective, same-pretraining-seed AF2+AF2 reference; the companion margin quantity is the drop in positive-versus-unrelated separation under the model's own view/augmentation protocol. These are descriptive representation diagnostics and are not accompanied by grouped downstream confidence intervals.

Table 14: Objective-level representation diagnostic. "Same" and "Cross" are unrelated-pair cosine similarities averaged at objective level; the diagnostic is the within-objective contrast between them.

Objective	Same r_{neg}	Cross mean	Mean Δr_{neg}	Margin drop
VICRegL	0.743	0.867	0.125	0.126
DenseCL	0.598	0.747	0.148	0.155
SimCLR	0.729	0.782	0.053	0.070

Table 15: Pairing-level objective-matched diagnostic. Values are mean ± sample SD across the three independent pretraining seeds.

Objective	Pairing	Δr_{neg}	Margin drop
VICRegL	AF2+AF1	0.115 ± 0.016	0.116 ± 0.018
VICRegL	AF2+AF3	0.118 ± 0.017	0.119 ± 0.019
VICRegL	AF2+AF4	0.141 ± 0.013	0.142 ± 0.014
DenseCL	AF2+AF1	0.158 ± 0.007	0.158 ± 0.006
DenseCL	AF2+AF3	0.164 ± 0.009	0.165 ± 0.009
DenseCL	AF2+AF4	0.123 ± 0.021	0.143 ± 0.020
SimCLR	AF2+AF1	0.059 ± 0.023	0.071 ± 0.024
SimCLR	AF2+AF3	0.045 ± 0.017	0.056 ± 0.018
SimCLR	AF2+AF4	0.054 ± 0.018	0.084 ± 0.014

Across the evaluated grid, the directional increase in Δr_{neg} and the separation-margin drop occur in 27/27 cross-frequency objective/pairing/pretraining-seed comparisons. The two quantities are closely coupled (Table 15) and should be read as two views of one signature rather than as independent lines of evidence. Together they support a within-objective warning signature. Their magnitude does not track the size of the downstream penalty, so they are not read as a continuous predictor of macro-AUPRC.

F Threshold-Free Robustness Pointwise Results

For Figure 4, downstream probabilities are averaged within each pretrained encoder before macro-AUPRC is computed; the pointwise values below are then means across pretraining seeds. Circumferential shift A1 uses $s \in \{0, 1, 2, 3, 4, 6, 8\}$; axial-speed control A2 uses ratios $r \in \{0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3\}$. Points along a curve are correlated and are interpreted descriptively.

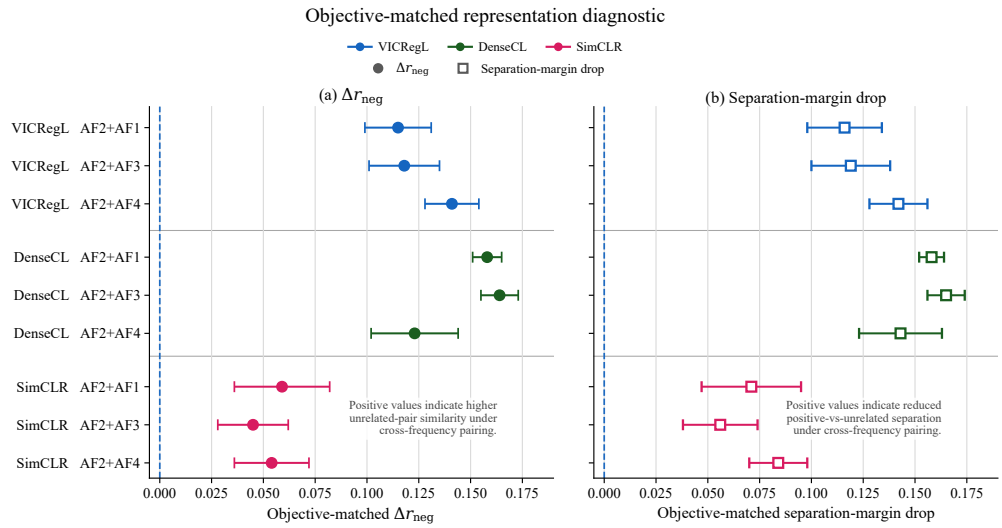


Figure 6: Objective-matched representation diagnostic. Points show three-pretraining-seed means of Δr_{neg} and positive-versus-unrelated separation-margin drop; error bars are sample SD across pretraining seeds. Both quantities are positive throughout the evaluated grid. No threshold is drawn.

Table 16: Figure 4(a) pointwise macro-AUPRC under circumferential shift. PA = PhysAug; Gen = Generic.

Objective/model	Aug.	0	1	2	3	4	6	8
VICRegL	PA	.715	.713	.711	.697	.705	.714	.705
VICRegL	Gen	.680	.672	.669	.612	.621	.624	.635
DenseCL	PA	.702	.705	.701	.698	.689	.699	.685
DenseCL	Gen	.653	.682	.600	.556	.580	.613	.525
SimCLR	PA	.690	.682	.683	.666	.658	.666	.671
SimCLR	Gen	.536	.541	.478	.483	.489	.473	.491
CNN-PhysAug $f = 100\%$	PA	.629	.624	.538	.513	.476	.440	.511

The A1 macro-AUPRC ranges are 0.0176/0.0680 (VICRegL PA/Gen), 0.0196/0.1575 (DenseCL), and 0.0312/0.0679 (SimCLR). PhysAug has the smaller A1 range in all 9/9 objective-by-pretraining-seed comparisons, a descriptive replication pattern.

Table 17: Figure 4(b) pointwise macro-AUPRC under the axial-speed resampling control.

Objective/model	Aug.	0.7	0.8	0.9	1.0	1.1	1.2	1.3
VICRegL	PA	.718	.719	.716	.715	.705	.693	.687
VICRegL	Gen	.685	.681	.682	.680	.671	.652	.642
DenseCL	PA	.702	.702	.705	.702	.690	.682	.669
DenseCL	Gen	.654	.651	.652	.653	.637	.629	.620
SimCLR	PA	.692	.690	.689	.690	.675	.665	.640
SimCLR	Gen	.541	.537	.535	.536	.522	.504	.492
CNN-PhysAug $f = 100\%$	PA	.654	.637	.627	.629	.616	.609	.598

The corresponding A2 ranges are 0.0324/0.0435 (VICRegL PA/Gen), 0.0357/0.0342 (DenseCL), and 0.0513/0.0482 (SimCLR). Thus the large targeted A1 stability separation is not reproduced under A2; PhysAug retains its higher absolute clean-performance offset. This is evidence for nuisance-specific stability, not universal OOD robustness or component-level causality.

G Fixed-Threshold Operating-Point Robustness

This analysis deliberately answers a different question from threshold-free AUPRC. For every individual downstream checkpoint, class-specific ID/OD/TW thresholds are selected on clean I004 validation by exhaustive search over unique predicted probabilities, maximizing F_3. That downstream

model's thresholds are frozen for every shifted I003 evaluation. Precision, recall, and F_3 are computed per downstream model, averaged over the three downstream models within a pretrained encoder, and then summarized as mean $\pm$ sample SD across pretraining seeds $\{42, 7, 21\}$. No downstream probability consensus occurs before threshold selection.

Table 18: Fixed-threshold circumferential-shift summary over $s \in \{0, 1, 2, 3, 4, 6, 8\}$. Range is max–min Macro-F_3 over all seven shifts; CNN-PhysAug $f = 100\%$ is the fair supervised reference.

Model	Clean F3	Worst F3	Range	Recall@8	Prec.@8	F3@8
VICRegL-PhysAug	0.776	0.748	0.031	0.886	0.426	0.767
VICRegL-Generic	0.686	0.609	0.097	0.777	0.413	0.634
DenseCL-PhysAug	0.779	0.761	0.026	0.871	0.480	0.784
DenseCL-Generic	0.722	0.559	0.163	0.664	0.395	0.559
SimCLR-PhysAug	0.724	0.711	0.021	0.828	0.425	0.711
SimCLR-Generic	0.574	0.513	0.107	0.699	0.330	0.572
CNN-PhysAug, $f = 100\%$	0.602	0.464	0.138	0.836	0.137	0.519

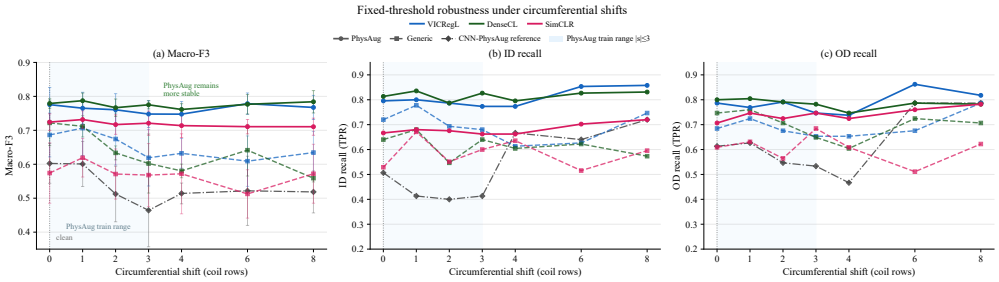


Figure 7: Three-pretraining-seed fixed-threshold robustness under circumferential shifts. Each downstream model independently selects class-specific F_3-optimal thresholds on clean I004 validation and freezes them for all shifted I003 evaluations; operating metrics are averaged within pretrained encoder and then summarized across pretraining seeds. The 63/63 PhysAug-over-Generic Macro-F_3 directional observations are descriptive, not 63 independent significance tests.

PhysAug exceeds its matching Generic model in all 63 objective $\times$ pretraining-seed $\times$ shift Macro-F_3 point comparisons. The fair CNN reaches recall 0.836 at $s = 8$ but precision 0.137, illustrating why recall alone is not an SSL-superiority result; the joint precision/recall/F_3 operating point is the relevant descriptive summary.

H Physical-Unit Grouped Bootstrap

ECT windows overlap axially (width 256, stride 64), so independent patch resampling would be anti-conservative and is not used as the inferential procedure. The grouped bootstrap assigns positive patches to 18 first-parsed primary physical flaw-UID groups and background patches to 37 non-overlapping 256-sample axial blocks; the sev20 subset has five positive physical groups. Each replicate resamples positive and background groups separately with replacement and applies the identical physical-unit resample to all compared methods and all observed pretraining seeds. After downstream probability consensus within encoder, macro-AUPRC is computed separately per pretraining seed and the paired effect is averaged across the three pretraining seeds. We use 10,000 paired replicates, percentile 95% CIs, two-sided plus-one p-values, and BH correction within each declared family. Full-test macro-AUPRC comparisons have 9,993 valid replicates; sev20 has 10,000. Pretraining-seed sample SD and grouped-bootstrap CI quantify different uncertainty sources and are reported separately.

The main $f = 10\%$ cross-frequency family has 9/9 BH-supported contrasts. The sev20 family has 7/9 BH-supported contrasts and is secondary because only five independent positive physical groups are available. Figure 7's 63/63 directions and Figure 4's perturbation sweeps are descriptive and are not reinterpreted as additional significance families.

Table 19: Declared inferential families. BH correction is applied within, not across, each listed family.

Analysis	**Family size**	**Role**
Linear cross-frequency, $f = 10\%$	9	primary headline family
Linear cross-frequency, $f = 100\%$	9	secondary persistence family
Linear cross-frequency sev20, $f = 10\%$	9	secondary low-power family
Same-frequency PhysAug–Generic, $f = 10\%$	3	augmentation-support family
Fair SSL–CNN, matched $f = 10\%$	3	supervised-fairness family
Fair SSL $f = 10\%$–CNN $f = 100\%$	3	cross-budget descriptive family
LP-FT same-frequency PhysAug–Generic, $f = 10\%$	3	recovery/support family
LP-FT cross-frequency, $f = 10\%$	9	primary LP-FT scope family
LP-FT cross-frequency, $f = 100\%$	9	full-label recovery family