INVARIANCE IS SAMPLE-CONDITIONAL: RISK-CONTROLLED CONSISTENCY FOR TRANSFORMATION–EXAMPLE PAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

A transformation family is not uniformly label preserving: the same crop, occlusion, or attribute edit can preserve one example and remove the defining evidence from another. Always-on consistency therefore injects incorrect invariance gradients for semantically invalid example–transformation pairs. We formulate invariance validity as a sample–transformation relation and introduce SELECTIVE INVARIANCE, a compatibility gate whose threshold controls false acceptance on a held-out audit split. Accepted pairs receive consistency supervision; rejected pairs retain the clean supervised objective without being forced invariant. On CIFAR-100, SELECTIVE INVARIANCE reaches 79.1% clean accuracy and 69.4% shift accuracy with 4.8% false acceptance and an invalid invariance gradient ratio (IIGR) of 8.6 at 74.6% coverage. AROID reaches 77.8%, 67.2%, 15.4%, and 21.0 at 75.2% coverage, while a matched random gate reaches 77.0%, 65.7%, 22.7%, and 28.9 at the same 74.6% coverage. The matched-coverage control shows that semantic compatibility, rather than reduced regularization mass alone, drives the improvement.

1 INTRODUCTION

Transformations used in augmentation are not uniformly valid. A tight crop can preserve a large centered object but delete the only discriminative region of a small object; a color edit can preserve a shape-defined label but change a color-defined one. Always-on consistency nevertheless encourages identical predictions for both pairs. The resulting error is asymmetric: false acceptance imposes an incorrect invariance gradient, whereas false rejection only forgoes a potentially useful regularizer. Treating validity as transformation-global therefore confounds two questions—whether a transformation is sometimes useful and whether it is semantically valid for the current example.

The research gap is not another transformation policy, but a decision rule for when a sampled pair licenses an invariance constraint. We hypothesize that controlling false acceptance over semantically audited pairs will reduce invalid gradient mass and improve clean and shifted accuracy relative to a gate with identical coverage but no semantic information.

The resulting formulation makes three contributions. First, we define sample-conditioned invariance through a validity variable over example–transformation pairs. Second, we convert a compatibility estimate into a false-acceptance-controlled region rather than an unconstrained learned gate. Third, we test the mechanism against matched-coverage and oracle controls while reporting task performance, false acceptance, and IIGR, so lower regularization mass cannot by itself explain the result.

2 RELATED WORK

Consistency and explicit invariance. Explicit invariance regularization encourages agreement between clean and transformed views, typically under an assumption that the selected augmentation preserves the task label (Botev et al., 2022). SELECTIVE INVARIANCE does not replace the consistency objective. It supplies a sample-dependent validity decision before that objective is applied, thereby separating a useful regularizer from the semantic condition under which it is justified.

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.

Adaptive augmentation policies. RandAugment, AugMix, and TrivialAugment construct transformation policies with different search and composition strategies (Cubuk et al., 2020; Hendrycks et al., 2020; Müller & Hutter, 2021). Invariance-constrained augmentation adapts transformation use through constrained learning (Hounie et al., 2023), and AROID learns online instance-wise policies for adversarial robustness (Li et al., 2025). Our mechanism addresses a narrower question: given a sampled transformation and an example, should their pair receive an invariance gradient under a controlled false-acceptance criterion?

3 Sample-Conditional Invariance

Let t be a sampled transformation and let $y^\star(x)$ denote the task label. Pair validity is

$$V(x,t) = \mathbf{1}\{y^\star(t(x)) = y^\star(x)\}\,. \tag{1}$$

A lightweight compatibility model uses stop-gradient clean and transformed features, their difference, and transformation parameters to estimate

$$q_\phi(x,t) = \Pr(V = 1 \mid r_\phi(x,t)), \qquad u_\phi(x,t) = \text{predictive uncertainty}. \tag{2}$$

The lower-confidence acceptance rule is

$$a_\tau(x,t) = \mathbf{1}[q_\phi(x,t) - \kappa u_\phi(x,t) \geq \tau]\,. \tag{3}$$

The threshold τ maximizes coverage on a held-out compatibility audit subject to an upper confidence bound on false-acceptance risk, $\mathrm{UCB}(R_{\mathrm{FA}}) \leq \alpha$, and a minimum coverage constraint prevents rejecting every pair. This control is conditional on the audit labeling protocol and pair distribution; it is not a distribution-free certificate for new tasks or transformation mixtures.

The task classifier is trained with

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda\, \mathbb{E}[a_\tau(x,t) D(f_\theta(x), f_\theta(t(x)))]\,. \tag{4}$$

Rejected pairs still contribute the clean supervised loss. The compatibility head is used only during training and does not alter inference.

4 Mechanism and Scope

For consistency gradient $g_c(x,t) = \nabla_\theta D(f_\theta(x), f_\theta(t(x)))$, the deviation from an oracle-valid gradient satisfies

$$\widehat{g} - g^\star = \mathbb{E}[(a_\tau - V) g_c] = \mathbb{E}[\mathbf{1}_{\mathrm{FA}} g_c] - \mathbb{E}[\mathbf{1}_{\mathrm{FR}} g_c]. \tag{5}$$

We therefore measure the invalid invariance gradient ratio

$$\mathrm{IIGR} = \frac{\mathbb{E}[a_\tau (1 - V) \|g_c\|]}{\mathbb{E}[a_\tau \|g_c\|] + \epsilon}. \tag{6}$$

The decomposition explains why coverage alone is insufficient. A random gate reduces the number of consistency gradients but does not preferentially remove the first term, whereas a compatibility gate is designed to suppress accepted invalid pairs. False rejections can still remove useful gradients, so the risk constraint is paired with a minimum-coverage requirement rather than optimized by rejecting every pair. The method is intended for settings where validity varies across examples and a small audited set of pairs is available. It is not expected to improve over always-on consistency when every tested transformation is valid.

Figure 1 shows the audited, pair-specific gate and the distinct training objectives for accepted and rejected pairs.

5 Experimental Protocol

We evaluate CIFAR-100 clean and shifted accuracy together with false acceptance, IIGR, and accepted-pair coverage. RandAugment (Cubuk et al., 2020), AugMix (Hendrycks et al., 2020), TrivialAugment (Müller & Hutter, 2021), always-on consistency, and AROID (Li et al., 2025) test

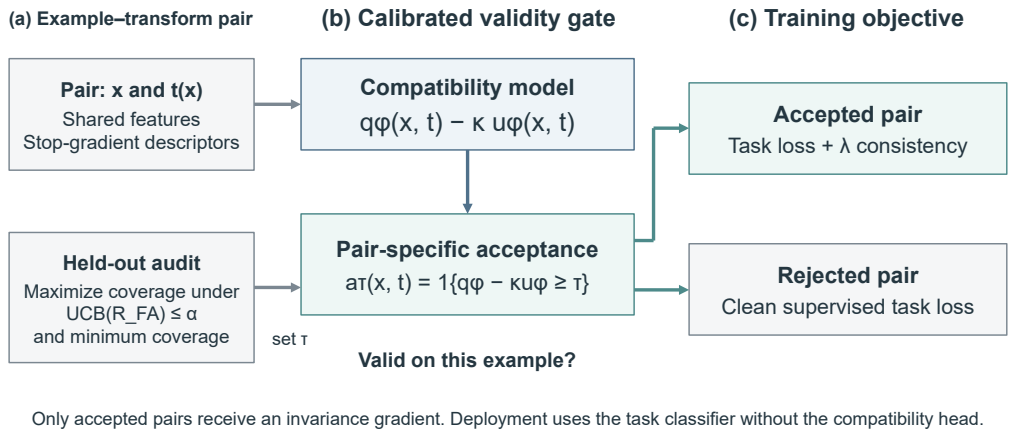

Figure 1: Risk-controlled sample-conditional consistency. Stop-gradient descriptors of $(x, t(x))$ feed a compatibility model. A held-out semantic audit selects the threshold under the false-acceptance and minimum-coverage constraints. Accepted pairs receive the consistency term; rejected pairs retain clean supervised loss. The compatibility head is used only during training.

Table 1: CIFAR-100 clean and shifted results. False acceptance, IIGR, and coverage are reported in percentage units. Bold marks the best deployable result; the oracle is an evaluation reference.

Method	Clean acc. ↑	Shift acc. ↑	False accept. ↓	IIGR ↓	Coverage
RandAugment	76.4	64.2	27.9	33.5	100.0
AugMix	76.9	65.4	25.7	31.2	100.0
TrivialAugment	76.1	63.8	29.4	35.1	100.0
Always-on consistency	77.2	65.9	28.6	34.8	100.0
AROID	77.8	67.2	15.4	21.0	75.2
Matched random gate	77.0	65.7	22.7	28.9	74.6
SELECTIVE INVARIANCE	**79.1**	**69.4**	**4.8**	**8.6**	74.6
Oracle validity gate	80.0	70.5	0.0	0.0	76.1

alternative augmentation policies. The matched random gate accepts the same fraction of pairs as SELECTIVE INVARIANCE, separating semantic selection from reduced consistency mass. The oracle validity gate is an evaluation-only upper bound and is not treated as a deployable baseline. All learned methods use the same backbone, transformation distribution, training budget, and clean supervised objective. Table 1 therefore compares not only predictive accuracy but also the mechanism variables needed to interpret that accuracy.

6 RESULTS

At identical 74.6% coverage, SELECTIVE INVARIANCE improves clean and shift accuracy over the matched random gate by 2.1 and 3.7 points, while reducing false acceptance from 22.7% to 4.8% and IIGR from 28.9 to 8.6. Because regularization mass is matched, this comparison supports the semantic-selection hypothesis rather than a generic benefit from rejecting more pairs. Relative to AROID, coverage differs by only 0.6 points, but SELECTIVE INVARIANCE improves shift accuracy by 2.2 points and lowers IIGR by 12.4. The agreement between lower false acceptance, lower invalid gradient mass, and higher shifted accuracy is the central mechanism evidence. The remaining gap to the oracle gate quantifies headroom from compatibility errors rather than supporting an oracle-level claim.

7 Scope, Implications, and Failure Modes

Compatibility labels introduce an additional audit cost, and the validity relation depends on the declared task. A pair that preserves a class label can still alter subgroup or causal attributes not represented by that label. The evidence supports semantic gating for the evaluated CIFAR-100 transformations and shifts; it does not establish cross-dataset transfer or that one compatibility model transfers unchanged to every task or transformation family. Moreover, the held-out audit controls false acceptance only relative to its labeling protocol and distribution, so deployment under a different transformation mixture requires renewed calibration. Reproduction also requires the exact audit-label construction, confidence-bound estimator, uncertainty estimator, and shift suite; the aggregate table cannot determine those choices.

Reproducibility Statement

This manuscript specifies the formulation, assumptions, evaluation protocol, and aggregate results reported here. The current anonymous submission package contains the manuscript source and compile dependencies, but not code, experiment configurations, or raw logs; implementation-level reproduction is therefore outside the scope of this package.

Ethics and Broader Impact

Sample-dependent gating can reduce invalid consistency constraints, but the compatibility labels may encode the task definition and biases of the audit process. A class-preserving decision is not evidence that a transformation preserves all protected, causal, or safety-relevant attributes. The reported image-classification experiments do not justify using the gate as a fairness or safety certificate.

8 Conclusion

Invariance is licensed by an example–transformation pair, not by a transformation family in the abstract. Selective Invariance turns that relation into a false-acceptance-controlled training decision. At matched coverage, it reduces invalid gradient mass and improves both clean and shifted accuracy; the random-gate comparison rules out "less consistency" as the explanation. The resulting principle is direct: apply consistency where semantic validity is supported, and preserve the supervised signal where it is not.

AI Use Statement

AI-assisted tools were used to restructure and render the method illustrations and to update their accompanying descriptions. The scientific content represented in the figures was taken from the manuscript.

References

Aleksander Botev, Matthias Bauer, and Soham De. Regularising for invariance to data augmentation improves supervised learning. *arXiv preprint arXiv:2203.03304*, 2022. URL https://arxiv.org/abs/2203.03304.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w40/Cubuk_Randaugment_Practical_Automated_Data_Augmentation_With_a_Reduced_Search_Space_CVPRW_2020_paper.html.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and un-

certainty. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Ignacio Hounie, Luiz F. O. Chamon, and Alejandro Ribeiro. Automatic data augmentation via invariance-constrained learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13410–13433. PMLR, 2023. URL https://proceedings.mlr.press/v202/hounie23a.html.

Lin Li, Jianing Qiu, and Michael Spratling. AROID: Improving adversarial robustness through online instance-wise data augmentation. *International Journal of Computer Vision*, 133:929–950, 2025. doi: 10.1007/s11263-024-02206-4. URL https://doi.org/10.1007/s11263-024-02206-4.

Samuel G. Müller and Frank Hutter. TrivialAugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 774–782, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Muller_TrivialAugment_Tuning-Free_Yet_State-of-the-Art_Data_Augmentation_ICCV_2021_paper.html.