

Zero-Shot Test-Time Canonicalization using Out-of-Distribution Scoring

Dominik Lindner¹ (✉), Johann Schmidt¹, Tom Siegl², Martin Becker³, and
Sebastian Stober¹

¹ Artificial Intelligence Lab, Otto-von-Guericke University Magdeburg, Germany
{dominik.lindner, johann.schmidt, stober}@ovgu.de

² University of Rostock, Germany tom.siegl@uni-rostock.de

³ Becker Lab, University of Marburg, Germany martin.becker@uni-marburg.de

Abstract. Pretrained vision models often misclassify inputs that are rotated, scaled, or sheared, even though these affine transformations leave the object class unchanged. Robustness is usually restored either by building equivariance into the architecture or by retraining with augmentation, both of which require changing or retraining the model. Test-time canonicalization instead leaves the classifier untouched. It undoes the transformation of each input, mapping it to a canonical form near the training distribution before classification. Existing canonicalizers, however, rely on a narrow set of logit-based energy scores and bespoke search procedures, leaving the design space of scoring functions and optimizers unexplored. We reframe canonicalization as out-of-distribution (OOD) detection, which lets any OOD score serve as the energy minimized over transformations. Across benchmarks ranging from handwritten characters and sketches to natural images and 3D point clouds, we systematically evaluate around twenty OOD scores and nine search algorithms, finding that distance-based scores paired with random search and local refinement perform best overall. Because canonicalizing an already-aligned input can hurt accuracy, we add a gated mechanism that transforms an input only when its OOD score indicates this is needed, preserving most in-distribution accuracy while retaining the robustness gains on transformed inputs. Code is available at github.com/johschm/its.

Keywords: Canonicalization · Out-of-Distribution Detection · Robustness · Group Invariance.

1 Introduction

For a model to be reliable, its predictions should remain stable under class-preserving transformations of the input. A rotated, rescaled, or sheared image of an object still depicts the same object. Such transformations are common at deployment time, for instance in microscopy, aerial imagery, or sketch and document analysis, where the upright pose seen during training cannot be assumed.

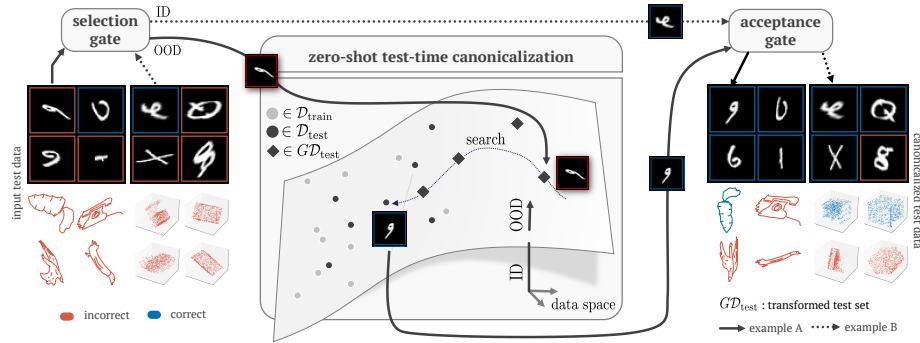


Fig. 1: At test time, in-distribution (ID) samples go directly to the downstream model, while out-of-distribution (OOD) samples enter a zero-shot canonicalization module that searches over transformations to minimize the OOD score. The result is passed on only if it lowers the score.

Yet many vision models are not robust to them [17,60], and retraining a deployed model for every such setting is often impractical. A strong form of robustness is invariance, where the predicted class remains unchanged under input transformations [26]. The most widely used method to increase robustness without altering the model architecture is *data augmentation*, which adds semantically redundant, transformed samples to the training set to encourage learning robust filters [24]. Invariance can be guaranteed by equivariant models [8] using group-pooling [7], which constrain the architecture through explicit inductive biases. This, unfortunately, is very constraining, as such an inductive bias has to be designed for the specific transformation and the model has to be trained from scratch. The downstream prediction can be made invariant without such constraints by averaging over all transformations [57]. Subsequent works reduce this to an average over a subspaces [42], but the computational cost remains prohibitive. This is closely related to Test-time Augmentation (TTA), where predictions are also averaged over transformed inputs. However, they assume a heuristic distribution over transformations and do not guarantee invariance [29].

An alternative approach to spatial robustness is *canonicalization*, which seeks to transform inputs into a canonical form by “undoing” group transformations prior to inference. Crucially, canonicalization does not constrain the downstream architecture, which can therefore be treated as fixed. Early canonicalizers learned to regress transformation parameters directly from data [25]. However, this can lead to fragile predictions even with modern backbones [46]. Later approaches formalized canonicalization as the selection of a canonical representation for an input sample by minimizing an equivariant scoring function over group transformations [28,40]. These methods typically rely on equivariant features or architectures for the canonicalizer [7] to ensure consistent scoring across transformed inputs. More recent work relaxes the requirement of equivariant canonicalizers by introducing training procedures for non-equivariant networks [41], but such approaches still require substantial training effort. Many learning-based canon-

icalizers further assume or benefit from training data that exhibits consistent alignment or a canonical pose distribution [40]. This assumption is often violated in real-world datasets, where object pose may vary significantly. Recent work addresses this by progressively re-aligning the training data during learning [48], but at the cost of additional training complexity. This raises a practical question: can canonicalization be obtained without training at all? Recent methods therefore formulate canonicalization as a test-time optimization problem, where a transformation is selected by minimizing a score defined directly on the downstream classifier’s outputs [47]. However, existing canonicalization methods are limited by their narrow choice of energy functions with costly optimization and a tendency to degrade in-distribution (ID) accuracy when inputs are already aligned. Existing approaches employ logit-based energies [47] or heuristic criteria [50] that favor transformations producing confident ID predictions.

We propose the hypothesis that OOD scores constitute suitable energy functions for canonicalization. The underlying assumption is that a canonical form lies in a high-density region of the training distribution and therefore attains a low OOD score, while transformed inputs move off this manifold and score higher. Minimizing an OOD score over the transformation group should thus recover a canonical form. This is a necessary rather than a sufficient condition. A low score favors transformations that appear typical, which need not coincide with the correct semantic pose, especially for classes with inherent pose ambiguity. We therefore treat it as an empirical hypothesis and test it across a broad range of scores and datasets. If it holds, the extensive body of OOD detection literature can be leveraged directly, broadening the design space for canonicalization. Moreover, the optimization strategy used to search over transformations is equally important. Prior work relies on bespoke search procedures tailored to specific scoring functions [47,51], including Lie-algebra-based methods for continuous transformation groups [50,29]. While effective, these approaches introduce additional complexity. Based on prior work, it remains unclear whether simpler global optimization strategies, such as simulated annealing or random search, which have shown benefits in other domains [45], can achieve competitive performance for common transformations such as affine transformations. Another issue of existing methods is the degradation of ID accuracy for aligned inputs [47], as incorrect or unnecessary transformations can distort features. This presents a practical challenge, since most real-world inputs are typically close to the training distribution and only a minority exhibit large transformations.

Contributions. Our main contributions to these problems are: (i) We systematically benchmark OOD scores as energy functions for canonicalization, addressing the limited exploration of energy functions in prior work. We find that the best-performing score varied across datasets. (ii) We show that prior search procedures can be replaced by simpler global optimizers (e.g., random search, simulated annealing) without sacrificing performance, with random search with local refinements working best overall. (iii) We introduce a lightweight OOD-based gating mechanism for canonicalization, resolving the accuracy trade-off in current frameworks by preserving ID and OOD performance.

2 Preliminaries

Let $(x, y) \sim \mathcal{D}$ denote input-label pairs, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Many vision tasks admit class-preserving input transformations: a rotated image of a fish remains a fish. To study such transformations formally, we use group theory. The following introduces the key concepts needed (see Sec. A and [4,20] for more details). A *group* G is a set equipped with an associative binary operation, an identity element, and inverses for every element. We restrict to matrix Lie groups [20], so every $g \in G$ is a matrix, and the group operation is matrix multiplication. A *group action* of G on $\mathcal{X}$ is a map $(g, x) \mapsto g \cdot x$. A prominent example is the affine group $\text{Aff}_2(\mathbb{R})$, whose elements (any combination of rotation, scaling, shearing, and translation) act on pixel coordinates in homogeneous coordinates via 3×3 matrix multiplication. The *orbit* of x under G is $Gx = \{g \cdot x \mid g \in G\}$, which collects all transformed versions of x . A function $f: \mathcal{X} \rightarrow \mathcal{Y}$ is G -equivariant if $f(g \cdot x) = g \cdot f(x)$ for all $g \in G$ and $x \in \mathcal{X}$. In classification, the group acts trivially on labels, so equivariance reduces to G -invariance: $f(g \cdot x) = f(x) \quad \forall g \in G, x \in \mathcal{X}$. A G -invariant classifier assigns a single label to every orbit, which is precisely the robustness property we seek. We assume the classifier f to be not G -invariant, i.e. whose predictions can change under transformations in G , and we aim to render it G -invariant at test time. We do this by introducing a G -invariant function $h: \mathcal{X} \rightarrow \mathcal{X}$ such that $f(h(gx)) = f(h(x))$. We gain G -invariance without restricting f . This process is called *canonicalization* [28], where h maps input data to a specific $x \in Gx$ (the *canonical form*).

Definition 1 (Orbit Canonicalization [15]). *A canonicalization is a function $h: \mathcal{X} \rightarrow \mathcal{X}$ fulfilling the following invariance property for all $g \in G$ and $x \in \mathcal{X}$: $h(g \cdot x) = h(x)$ and $h(x) \in Gx$.*

As there is no trivial choice for $h(x)$, the most popular approaches [28,40] use optimization to find the canonical form of x . In this setting, each $g \cdot x \in Gx$ is assigned an energy and an optimizer seeks to find the point of least energy g^* , corresponding to the canonical form. The choice of energy function $E: \mathcal{X} \rightarrow \mathbb{R}$ is essential, as its minima should correspond to the canonical form of x and should be unique up to symmetries. The resulting $(g^*)^{-1} \cdot x$ is “undoing” the transformation applied to x , mapping it to its canonical form. Such that

$$h(x) = (g^*)^{-1} \cdot x \quad \text{with} \quad g^* = \arg \min_{g \in G} E(g^{-1} \cdot x). \quad (1)$$

For small groups this minimization can be done exactly, while continuous groups require methods to find approximate solutions. In principle E can be any function. What it does not guarantee is a unique minimizer per orbit: a generic OOD score may attain its minimum at several transformations that are not related by the stabilizer, in which case the canonical form is not well-defined and consistency across the orbit can fail [28]. We nonetheless forgo this formal uniqueness guarantee, since suboptimal solutions without such guarantees often perform well in practice [47], and we verify the resulting canonical forms empirically.

Table 1: Tested OOD scores, grouped by family (logit-based, activation-rectification, prototype-based, kNN-based; separated by rules). For each score we mark whether it requires stored training samples (TS), hyperparameter tuning (HP), or an auxiliary logit-based score (AS). The bottom row visualizes the relative inference-time increase (RT) of the full search relative to the Energy score, with bar height proportional to the increase. Bars below the axis indicate scores faster than Energy. Values are averaged across MNIST, EMNIST, TU Berlin and ModelNet10. Exact numbers are given in Tab. 1.

	<i>Energy</i>	<i>Entropy</i>	<i>GEN</i>	<i>MSP</i>	<i>MaxLogit</i>	<i>OpenMax</i>	<i>VIM</i>	<i>Laplace</i>	<i>ReAct</i>	<i>ASH</i>	<i>DICE</i>	<i>MD</i>	<i>RMD</i>	<i>SHE</i>	<i>Proto</i>	<i>PC-Proto</i>	<i>kNN</i>	<i>Trust Score</i>	<i>kNN Mix</i>	<i>PC-kNN Mix</i>
TS	-	-	-	-	-	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
HP	-	-	✓	-	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AS	-	-	-	-	-	-	-	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-
RT	0.0%	-3.0%	-1.5%	-3.5%	-1.7%	32.5%	1.8%	53.9%	-1.0%	2.2%	6.0%	13.3%	7.8%	2.2%	3.8%	9.3%	14.3%	52.8%	37.5%	43.5%

3 OOD-based Canonicalization

3.1 Searching for Canonical Forms

An OOD score $S : \mathcal{X} \rightarrow \mathbb{R}$ measures alignment with the training distribution, with lower values indicating that x is drawn from the distribution the classifier f was trained on. Our central insight is that, since canonical representatives reside near the data manifold [51], OOD scores are natural candidates for the energy E in Equation (1). This is a necessary but not sufficient condition. An OOD score rewards transformations that look typical, which may differ from the correct semantic pose for classes with inherent pose ambiguity. Whether it is sufficient in practice is the empirical question we study. We set $E(x) = S(x)$ and aim to find the lowest OOD score over all target transformations G . Throughout, we assume the transformation group G is specified in advance and contains the transformations the input may have undergone. This means that $g^* \cdot x$ is the most ID point. This connection opens the full landscape of OOD scores for canonical form estimation, whereas prior work was confined to very limited options [28,40,47,51]. Similarly, the choice of optimization strategy over G has received little attention, with prior work largely restricted to gradient-based optimization [40] or tree-searches [47] or Bayesian Optimization [51]. In this work, we study various choices for algorithms to find the minimum in Equation (1). We focus on heuristics rather than optimal solvers as for continuous groups this would be intractable. This leads to potentially suboptimal results (i.e., local minima g with $S((g^*)^{-1} \cdot x) \leq S(g^{-1} \cdot x)$). Even when S is convex in x , it is generally non-convex as a function of g , necessitating multi-start optimization rather than a single descent run [50]. These local minima can yield predicted canonical forms

that are themselves OOD [47,51]. This sub-optimality is particularly harmful when the found $g^{-1} \cdot x$ (local minima) is far from the true canonical form (global minima), degrading the classifier’s accuracy on standard inputs [47]. We address this with a gating mechanism for canonicalization functions h .

3.2 Gated Canonicalization

A practical failure mode of continuous canonicalization is the degradation of accuracy on data that is already in-distribution (which we show in the experiments). Unnecessary transformations perturb features and lower downstream confidence. To prevent this, we introduce *Gated Canonicalization*, which comprises two gates: a selection gate $\alpha_\tau(x)$ and an acceptance gate $\alpha(x)$. The former ensures that canonicalization is only applied if x is OOD. The latter ensures that the output of the canonicalization function reduces the OOD score, i.e., $S(\alpha_\tau(x)) \leq S(x)$. This is equivalent to always including the identity in the search, though stronger acceptance criteria may break this equivalence (see Sec. E.7). We define two canonicalization gates:

$$\text{selection gate } \alpha_\tau(x) = \begin{cases} (g^*)^{-1} \cdot x & \text{if } S(x) > \tau, \\ x & \text{otherwise,} \end{cases} \quad (2)$$

$$\text{acceptance gate } \alpha(x) = \begin{cases} (g^*)^{-1} \cdot x & \text{if } S((g^*)^{-1} \cdot x) < S(x), \\ x & \text{otherwise,} \end{cases} \quad (3)$$

where $\tau \in \mathbb{R}$ is a user-defined OOD threshold. In practice, τ is selected on a held-out validation set augmented with the same transformations, requiring no labels for the transformed test data. We study its effect in Sec. 5 and report validation-selected operating points in Sec. E.7. These mechanisms decouple the task of ID preservation from OOD robustness, treating the canonicalization process as a conditional corrective mapping. A visual overview is illustrated in Fig. 1.

4 OOD Scores and Search Methods

OOD Scoring Functions. We limit our focus to training-free OOD scores that require no architectural changes. Our central premise is that canonical representatives lie near the data manifold, so any reliable OOD score is a natural energy function for canonicalization. We evaluate four families of detectors. *Logit-based methods* operate on classifier outputs and require no stored features. We include Maximum Softmax Probability (MSP) [23], MaxLogit [22], Energy [36], Entropy, Generalized Entropy (GEN) [37], OpenMax [3], Virtual-logit Matching (VIM) [58], and the Laplace Approximation [9] with energy, entropy, and mutual information (MI) scores [33]. *Activation-rectification methods* modify intermediate features before computing a logit-based score, amplifying distributional differences. We include Rectified Activations (ReAct) [54], Activation Shaping (ASH) [12], and Directed Sparsification (DICE) [55]. *Prototype-based methods* measure

Table 2: Tested search algorithms. (*left*) For each method we mark whether it is a global search (GS), gradient-based (GB), and parallelizable (PL), and visualize the relative inference-time increase (RT) over Random Search as a bar (height proportional to the increase, averaged across MNIST, EMNIST, TU Berlin and ModelNet10). (*right*) Worst-case time complexity in Big-O.

	<i>RS</i>	<i>RS-LR</i>	<i>SA</i>	<i>GD</i>	<i>CD-S</i>	<i>CD</i>	<i>WCD</i>	<i>WCD-L</i>	<i>ITS</i>	Method	Complexity	
GS	✓	✓	✓	✓	—	—	—	—	—	RS	$\mathcal{O}(N)$	N : samples
GB	—	✓	—	✓	—	—	—	—	—	RS-LR	$\mathcal{O}(N + Lk)$	L : local pts
PL	✓	✓	✓	✓	—	—	—	—	—	SA	$\mathcal{O}(Rk)$	k : iterations
RT	0.0%	5.7%	34.8%	6.7%	11.2%	13.1%	19.3%	27.3%	24.76%	GD	$\mathcal{O}(Rk)$	R : runs/restarts
										WCD	$\mathcal{O}(M(d+W)c)$	M : samples/dim
										ITS	$\mathcal{O}(MHd)$	d : dims
												W : first-dim weight
												c : cycles
												H : hypotheses

distance to class representatives in feature space. We include Mahalanobis Distance (MD) [35], Relative Mahalanobis Distance (RMD) [43], Simplified Hopfield Energy (SHE) [61], and prototype distances using Euclidean, cosine, or a convex combination of both, applied either globally or per predicted class (PC-Proto). *kNN-based methods* score inputs by their average distance to the k nearest in-distribution embeddings [56]. We include variants using Euclidean and cosine distance, a convex mixture (kNN Mix), per-class versions (PC-kNN, PC-kNN Mix), and Trust Score [27]. These methods require storing training embeddings but no additional training. Tab. 1 summarizes the requirements of each detector. Further implementation details are in Sec. B.

Search Algorithms. We evaluate algorithms from global exploration to local exploitation. *Random Search (RS)* [52] samples the parameter space via Sobol sequences for uniform coverage. *RS with Local Refinement (RS-LR)* applies Adam-based gradient descent to the best candidate from global sampling. *Simulated Annealing (SA)* [31] uses neighborhood sampling [53] with multiple parallel runs to reduce latency. *Multi-start Gradient Descent (GD)*, as used in [50], runs Adam from several random initializations. As a non-global baseline inspired by ITS [47], we study *Coordinate Descent (CD)*, which sequentially optimizes each transformation parameter via uniform random sampling, iterating over multiple cycles. Variants include *CD-S* (one cycle), *Weighted CD (WCD)* (increased sampling density in the first dim.), and *WCD-L* (equally spaced samples). Tab. 2 summarizes properties and complexities with full details in Sec. C.

5 Experiments

We make our source code publicly available for the reproducibility. We refer the reader to Sec. D or our source code for more details on the used hyperparameters.

Finding. Across all benchmarks, random search with local refinement (**RS-LR**) combined with a distance-based OOD score performs best (Secs. 5.2 and 5.3). The single best score is dataset dependent but always distance-based: kNN for MNIST, PC-kNN for EMNIST, kNN Mix for ModelNet10, and PC-Proto for TU Berlin. On the validation set, the exact same detectors achieve the top scores across all datasets except ModelNet10, where **PC-kNN Mix** performs best. We refer to using the best configuration(RS-LR plus score) on the validation set as **OODC** (OOD-based Canonicalization) and use it throughout unless stated otherwise.

Benchmarks, Models, and Metrics. We evaluate on a deliberately diverse set of benchmarks: MNIST and EMNIST Balanced [6] (handwritten characters), TU Berlin [16] (single-line sketches), ModelNet10 [62] (3D point clouds), and SI-Score (rotation) [60], a robustness benchmark of ImageNet [11] foreground objects transformed onto varying backgrounds. For MNIST, EMNIST, and TU Berlin we construct transformed test sets by applying rotation, shear, and scaling. For ModelNet10 only rotation is applied, as scale and translation are normalized a priori. Crucially, the search draws candidate transformations from the same family and parameter ranges used to generate these transformed sets. This aligns the search space with the true corruption and likely benefits informed and random search differently. Under natural OOD, where the transformation family is unknown or only approximate, the advantage of any particular search strategy may shrink. Full augmentation ranges, downstream architectures, and training details are given in Sec. D. We report average top-1 test accuracy (mean over multiple runs with $1.96\times$ standard error to indicate statistical significance). To measure how close the search comes to the canonical form (the energetic minimum), we avoid the degenerate distance between predicted and ground-truth affine matrices and instead use a score-based distance,

$$\text{Scoring Distance (Err.): } \frac{S((g^*)^{-1}gx) - S((g_{\min}^*)^{-1}gx)}{S((g_{\max}^*)^{-1}gx) - S((g_{\min}^*)^{-1}gx)} \in [0, 1], \quad (4)$$

where g^* is the found solution, gx the transformed input, and $g_{\min}^*, g_{\max}^*$ the transformations with minimal and maximal score across all runs and methods.

5.1 How competitive is our OOD-based canonicalization?

We compared our OOD-based canonicalization framework on various benchmarks (including ID and OOD test sets) against multiple baselines. Results for MNIST, EMNIST, TU Berlin, and ModelNet10 are shown in Tab. 3. All retraining-based models were trained on augmented data. The STN [25] used a smaller version of the main model, with one-quarter of the channels. Test-time Augmentation averaged predictions over eight transformation samples, always including the identity. Equivariant architectures perform best among approaches requiring retraining, while the vanilla model leads among training-free variants but lacks robustness. OODC achieves the highest accuracy on transformed test

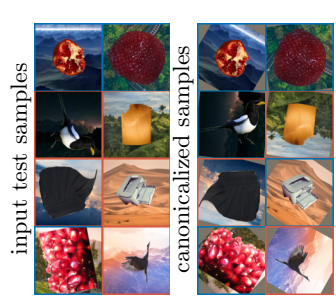
Table 3: Accuracy (%) on the original (top row, $\times$) and transformed (bottom row, $\checkmark$) test sets, indicated in column **Trf.** Methods are grouped into those that require retraining and those that do not. Retraining baselines: augmented training (**Aug.**), an equivariant SteerableCNN [8] or Vector Neuron model [10] (**Equi.**), a spatial transformer trained on augmented data (**STN**), and canonicalization with a learned energy model (**Learned**, Sec. E.3). Training-free methods: the vanilla classifier (**Van.**), Test-time Augmentation (**TTA**) [34], vanilla **ITS** [47], ITS with its logit energy (**ITS^a**), **ITS^b** with the same OOD score as OODC, and **OODC**. The best result within each group is shown in bold per row.

Dataset	Trf.	Requires Retraining				No Retraining					
		Aug.	Equi.	STN	Learned	Van.	TTA	ITS	ITS ^a	ITS ^b	OODC
MNIST	$\times$	99.2	99.3	99.3	98.6 $\pm$ 0.1	99.6	83.0	76.3	93.8	98.1	95.1 $\pm$ 0.1
	$\checkmark$	98.7	99.2	99.2	98.1 $\pm$ 0.1	37.7	49.8	60.4	75.5	85.0	89.8$\pm$0.3
EMNIST	$\times$	83.2	84.1	83.6	76.7 $\pm$ 0.2	88.4	44.2	54.1	71.2	78.0	64.4 $\pm$ 0.2
	$\checkmark$	82.1	83.9	82.8	75.6 $\pm$ 0.1	17.1	17.2	33.4	46.6	51.4	56.1$\pm$0.4
TU Berlin	$\times$	42.4	—	30.5	35.0 $\pm$ 0.8	51.6	18.9	36.6	39.7	40.5	34.0 $\pm$ 0.4
	$\checkmark$	40.7	—	28.5	31.2 $\pm$ 1.4	8.2	8.2	19.7	22.0	25.7	28.1$\pm$0.3
ModelNet10	$\times$	80.1	87.8	82.5	70.1 $\pm$ 0.5	91.2	47.7	33.6	32.4	86.7	60.6 $\pm$ 0.7
	$\checkmark$	83.9	87.8	82.9	68.2 $\pm$ 1.6	20.5	28.3	27.1	27.6	54.0	59.6$\pm$0.9

data among all training-free methods. On untransformed data, however, its accuracy falls well below the vanilla model. This is inherent to always-on canonicalization: every input is transformed, so already-aligned inputs are needlessly perturbed and their features distorted. The same effect appears for the STN, which also transforms unconditionally, indicating that the drop is the cost of unconditional robustness rather than a weakness of the OOD score. In Sec. 5.4 we show that gating restores most of the in-distribution accuracy while retaining the gains on transformed inputs, converting this apparent trade-off into a controllable one. Next, we evaluate our canonicalization framework on SI-Score (Rotation) [60] as a large-scale robustness benchmark using a ResNet-50 [21] and a ViT-B16 [13] backbone. For optimization, we use 17 equally spaced samples as we only consider rotation. Due to computational cost, we restricted OODC to PC-kNN here and additionally report the energy of the logits, without exploring other scores. Hyperparameters were optimized on rotated ImageNet[11] images (distinct from data used for embeddings). Feature-based OOD scores significantly outperform existing approaches, including ITS [47] (using their logit-based energy detector and iterative search), foreground fine-tuning for ViT [5], and Test Time Augmentation (TTA) over 16 equally spaced rotations (Fig. 2).

5.2 What Search Algorithm to choose for Canonicalization?

As testing all classifiers and combinations is computationally not feasible, we opted to evaluate the search on several selected score functions first. As scores



Method	ViT-B16	ResNet50
Vanilla Classifier	38.8	48.0
RS Augm. [47]	41.0	—
Chefer et al. [5]	44.8	—
Chefer et al. + AS [5]	46.2	—
ITS [47]	49.0	—
TTA [34]	49.5	54.9
Energy of Logits	51.4 ± 0.2	48.3 ± 0.2
PC-kNN	57.9 ± 0.1	60.7 ± 0.1

Fig. 2: Performance on the SI-Score (rotation) with canonicalized images on the left (blue correct, red incorrect classification), and test accuracies on the right.

Table 4: Comparison of search strategies on different datasets with 60 evaluations, showing random search with local refinement performing best overall.

Method	MNIST		EMNIST		ModelNet10		TU Berlin	
	Acc ↑	Err ↓	Acc ↑	Err ↓	Acc ↑	Err ↓	Acc ↑	Err ↓
CD	52.2 ± .2	0.42	80.2 ± .1	0.33	37.5 ± .4	0.52	24.4 ± .3	0.37
CD-S	52.2 ± .1	0.42	80.1 ± .2	0.33	36.0 ± .3	0.59	24.5 ± .3	0.37
GD	54.0 ± .2	0.57	85.6 ± .2	0.42	39.6 ± .5	0.55	—	—
ITS	55.7 ± .1	0.44	83.6 ± .1	0.34	39.2 ± .1	0.59	25.3 ± .1	0.34
SA	56.3 ± .2	0.56	85.0 ± .3	0.47	40.2 ± .4	0.54	21.3 ± .3	0.47
RS	57.4 ± .2	0.56	87.0 ± .2	0.46	40.3 ± .4	0.53	21.1 ± .3	0.47
RS-LR	59.9 ± .2	0.36	88.0 ± .2	0.27	41.9 ± .4	0.44	—	—
WCD	56.0 ± .2	0.32	82.6 ± .2	0.24	38.4 ± .5	0.52	26.1 ± .2	0.36
WCD-L	57.0 ± .1	0.30	83.0 ± .1	0.22	40.2 ± .4	0.48	28.7 ± .5	0.69

for this experiment, we use the negative energy of the logits [36], distance to the nearest embedding using cosine distance at the last layer (PC-kNN). In addition, we train a classifier to distinguish between data points that the downstream model predicts correctly and those it predicts incorrectly (called a trained energy model). The goal is to have a search algorithm that works with a wide range of possible scores, so we use the average accuracy across these scores as the metric. A main limiting factor of the searches is the computation time which is mainly decided by the number of model calls the methods make. For a fair comparison, the first experiment uses a fixed budget of 60 sample evaluations, chosen due to its divisibility. A forward pass counts as one evaluation, while a forward and backward pass counts as two to reflect the higher cost. The BiLSTM model requires unrolling, which is more expensive, so we do not apply gradient-based methods on the TU Berlin dataset. The hyperparameters for each method and dataset are optimized for 30 iterations, using the average accuracy across all scores, while ensuring that the computational budget is not exceeded. As shown in Tab. 4, the relative distance varied by at most 0.003 ($1.96 \times$ standard error),

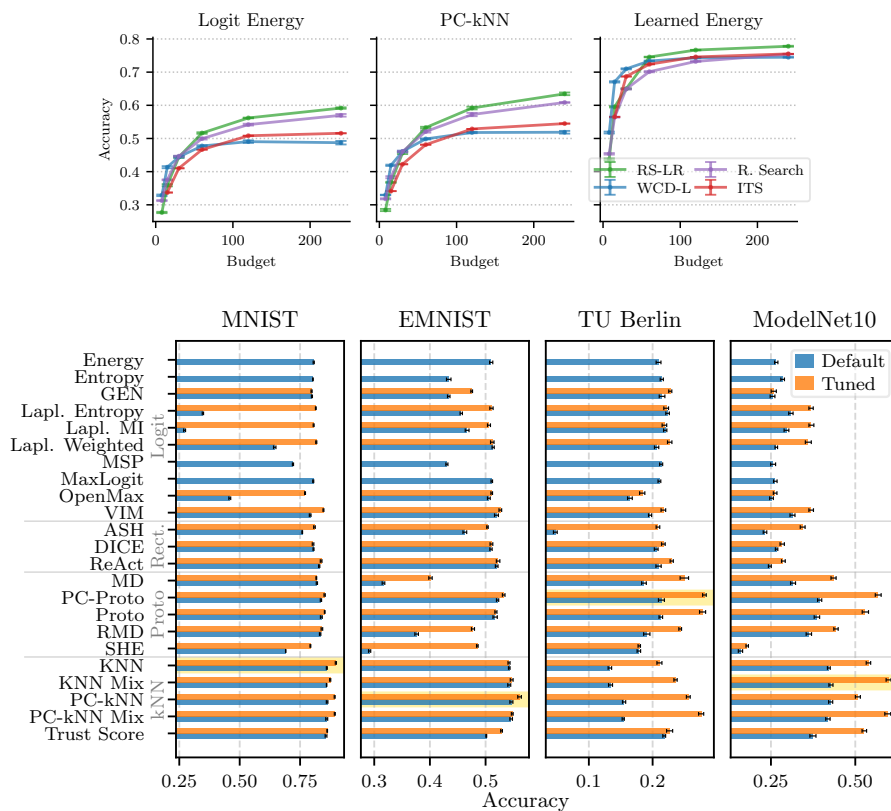


Fig. 4: Comparison of unsupervised OOD scores (best test score highlighted).

so it is omitted for readability. We found random search outperforms most other search algorithms. Coordinate descent-based methods find solutions closer to the minimum than random search, though often with lower accuracy, except on TU Berlin. Using local refinement can boost the accuracy slightly and find results closer to the minima, but typically not as close as WCD-L. As accuracy is typically prioritized, this makes it the preferable method within this budget. Next, we investigate if this holds for different budgets, re-optimizing hyperparameters for each. The results for previously well-performing methods on EMNIST are shown in Fig. 3 (cont. in Sec. E.5). On most datasets, RS-LR performs well with budgets above 30, though coordinate descent-based methods may outperform it at lower budgets. We found TU Berlin to be an exception where ITS works better. Overall, we recommend RS-LR for its simplicity, dimension-independent scaling, and strong performance at budgets where canonicalization is reliable.

5.3 What OOD Scoring Function to choose for Canonicalization?

As OOD scores exhibit varying performance depending on the dataset [59], we compare several detectors to examine whether this holds for canonicalization

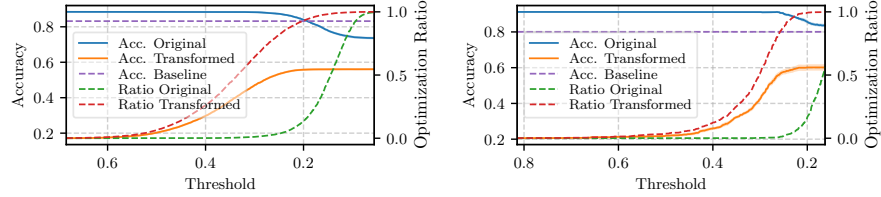


Fig. 5: Trade-off between accuracy on original and transformed data when varying the OOD threshold for EMNIST (left) and ModelNet10 (right), compared against the augmented model as the baseline.

and whether a well-chosen score can improve results. As the search strategy we use RS-LR as it worked well in previous experiments. Hyperparameters of the detectors were optimized in a two-stage process (see Sec. D.3). The results are shown in Fig. 4. Distance-based methods like kNN and prototype-based methods outperform logit-based methods, with kNN-based methods performing best overall. Laplace approximations and VIM also surpass logit-based methods but are beaten by distance-based approaches. Activation reshaping methods show similar performance to logit-based scores. In Sec. E we show example transformations by our method compared to our core baseline ITS. A concern with logit- and feature-based OOD detection is that information loss in deeper layers may hinder canonicalization. In Sec. E.3, we show that training a learned score function on later-layer features reduces accuracy, suggesting that feature-based methods may be limited compared to methods trained on the input data.

5.4 How does Gating trade ID vs OOD performance?

A working OOD detection setup should not degrade a model’s performance on ID data [59]. By thresholding the OOD score and only canonicalizing samples deemed OOD, we can control the trade-off between performance on transformed and untransformed data, accepting search results only when the score improves. Due to a lack of datasets with realistic mixtures of transformations, we only report the threshold’s effect on the standard and transformed test sets. Fig. 5 shows this trade-off for EMNIST and ModelNet10 using the best unsupervised OOD score per dataset. This shows that most performance on untransformed data can be retained while significantly improving it on transformed data. It remains below always-canonicalizing since some transformed inputs are skipped, which reduces computational costs. Other datasets show similar trends (Sec. E.7).

6 Conclusions

Zero-shot test-time canonicalization can and should take advantage of the vast space of existing out-of-distribution (OOD) scores and search heuristics. Selecting an appropriate OOD score significantly improves performance compared to

previous work [47]. We found that random search with local refinement coupled with distance-based OOD scores performs best overall, while the single best score is dataset dependent. Furthermore, existing canonicalization methods tend to degrade in-distribution (ID) performance as all inputs are always transformed. We introduce a gating mechanism that ensures that canonicalization is only executed when needed, significantly mitigating the negative effect on the ID performance. The framework requires no retraining and treats the classifier as a black box, so it can be attached to any pretrained model to increase robustness, with the inference-time search cost being the main practical consideration.

Limitations and Future Work. The 60-sample inference budget may limit real-time use. More sample-efficient optimizers such as evolutionary methods or Bayesian optimization are a natural next step. Deep features alone can be insufficient when spatial information is critical, so combining OOD scores with foundation-model energies [51] is promising. When transformations are known a priori and retraining is feasible, augmentation-based training can still be preferable, though test-time canonicalization offers a path toward closing this gap without such assumptions. Canonicalization could also be applied locally, to parts or individual objects rather than the whole frame, via part-sampling methods [49].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

7 Acknowledgments

Dominik Lindner was funded by the IB (ZS/2026/08/204393). Johann Schmidt was funded by the ZIM (KK5540302MS4). Tom Siegl was cofunded by the BMFTR (01IS22077, 01ED2507) and DFG SFB 1713.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: International Conference on Knowledge Discovery & Data Mining (ACM SIGKDD) (2019)
2. Amos, B., Xu, L., Kolter, J.Z.: Input convex neural networks. In: International Conference on Machine Learning (ICML) (2017)
3. Bendale, A., Boulton, T.E.: Towards Open Set Deep Networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Bronstein, M.M., Bruna, J., Cohen, T., Velicković, P.: Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. 2104.13478, arXiv (2021)
5. Chefer, H., Schwartz, I., Wolf, L.: Optimizing relevance maps of vision transformers improves robustness. In: Conference on Neural Information Processing Systems (NeurIPS) (2022)
6. Cohen, G., Afshar, S., Tapson, J., André van Schaik: EMNIST: Extending MNIST to handwritten letters. In: International Joint Conference on Neural Networks (IJCNN) (2017)

7. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: International Conference on Machine Learning (ICML) (2016)
8. Cohen, T.S., Welling, M.: Steerable CNNs. In: International Conference on Learning Representations (ICLR) (2017)
9. Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P.: Laplace Redux - Effortless Bayesian Deep Learning. In: Conference on Neural Information Processing Systems (NeurIPS) (2021)
10. Deng, C., Litany, O., Duan, Y., Poulenard, A., Tagliasacchi, A., Guibas, L.J.: Vector neurons: A general framework for $so(3)$ -equivariant networks. In: International Conference on Computer Vision (ICCV) (2021)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
12. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely Simple Activation Shaping for Out-of-Distribution Detection. In: International Conference on Learning Representations (ICLR) (2023)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR) (2021)
14. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data (2025)
15. Dym, N., Lawrence, H., Siegel, J.W.: Equivariant frames and the impossibility of continuous canonicalization. In: International Conference on Machine Learning (ICML) (2024)
16. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Transactions on Graphics (2012)
17. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A.: Exploring the landscape of spatial robustness. In: International Conference on Machine Learning (ICML) (2019)
18. Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds (2019)
19. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning (ICML) (2016)
20. Hall, B.C.: Lie Groups, Lie Algebras, and Representations: An Elementary Introduction. Springer International Publishing (2015)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
22. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling Out-of-Distribution Detection for Real-World Settings. In: International Conference on Machine Learning (ICML) (2022)
23. Hendrycks, D., Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In: International Conference on Learning Representations (ICLR) (2017)
24. Huang, T., Sankar, C., Amini, P., Kottur, S., Geramifard, A., Razaviyayn, M., Beirami, A.: Dair: Data augmented invariant regularization. ArXiv (2021)
25. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: International Conference on Neural Information Processing Systems (NeurIPS) (2015)

26. Jang, H., Zaidi, S.S.A., Boix, X., Prasad, N., Gilad-Gutnick, S., Ben-Ami, S., Sinha, P.: Robustness to transformations across categories: Is robustness driven by invariant neural representations? *Neural Computation* (2023)
27. Jiang, H., Kim, B., Guan, M., Gupta, M.: To trust or not to trust a classifier. In: *Conference on Neural Information Processing Systems (NeurIPS)* (2018)
28. Kaba, S.O., Mondal, A.K., Zhang, Y., Bengio, Y., Ravanbakhsh, S.: Equivariance with learned canonicalization functions. In: *International Conference on Machine Learning (ICML)* (2023)
29. Kim, J., Kaba, S.O., Park, J., Hong, S., Ravanbakhsh, S.: Inverting Data Transformations via Diffusion Sampling. *ArXiv* (2026)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)* (2015)
31. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. *Science* (1983)
32. Kristiadi, A., Hein, M., Hennig, P.: Being bayesian, even just a bit, fixes overconfidence in ReLU networks. In: *International Conference on Machine Learning (ICML)* (2020)
33. Lafage, A., Laurent, O., Gabetni, F., Franchi, G.: Torch-uncertainty: A deep learning framework for uncertainty quantification. In: *NeurIPS Datasets and Benchmarks Track* (2025)
34. Laptev, D., Savinov, N., Buhmann, J.M., Pollefeys, M.: Ti-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
35. Lee, K., Lee, K., Lee, H., Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In: *Conference on Neural Information Processing Systems (NeurIPS)* (2018)
36. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: *Conference on Neural Information Processing Systems (NeurIPS)* (2020)
37. Liu, X., Lochman, Y., Zach, C.: GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023)
38. Ma, G., Wang, Y., Lim, D., Jegelka, S., Wang, Y.: A Canonicalization Perspective on Invariant and Equivariant Learning. In: *Conference on Neural Information Processing Systems (NeurIPS)* (2024)
39. MacKay, D.J.C.: Bayesian Interpolation. *Neural Computation* (1992)
40. Mondal, A.K., Panigrahi, S.S., Kaba, S.O., Rajeswar, S., Ravanbakhsh, S.: Equivariant adaptation of large pretrained models. In: *Conference on Neural Information Processing Systems (NeurIPS)* (2023)
41. Panigrahi, S.S., Mondal, A.K.: Improved canonicalization for model agnostic equivariance. In: *CVPR Workshop on Equivariant Vision: From Theory to Practice* (2024)
42. Puny, O., Atzmon, M., Smith, E.J., Misra, I., Grover, A., Ben-Hamu, H., Lipman, Y.: Frame averaging for invariant and equivariant network design. In: *International Conference on Learning Representations (ICLR)* (2022)
43. Ren, J., Fort, S., Liu, J., Roy, A.G., Padhy, S., Lakshminarayanan, B.: A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection. *ArXiv* (2021)
44. Rudin, W.: *Principles of Mathematical Analysis*. McGraw-Hill (1976)
45. Schmidt, J., Köhler, B., Borstell, H.: Reviving simulated annealing lifting its degeneracies for real-time job scheduling. *Hawaii International Conference on System Sciences (HICSS)* (2024)

46. Schmidt, J., Siegl, T., Becker, M., Stober, S.: Geometrically constrained and token-based probabilistic spatial transformers. *ArXiv* (2026)
47. Schmidt, J., Stober, S.: Tilt your Head: Activating the Hidden Spatial-Invariance of Classifiers. In: *International Conference on Machine Learning (ICML)* (2024)
48. Schmidt, J., Stober, S.: Robust canonicalization through bootstrapped data re-alignment. *EurIPS: UPLB Workshop* (2025)
49. Schmidt, J., Stober, S., Denzler, J., Bodesheim, P.: Saccadic vision for fine-grained visual classification. *ArXiv* (2025)
50. Shumaylov, Z., Zaika, P., Rowbottom, J., Sherry, F., Weber, M., Schönlieb, C.B.: Lie algebra canonicalization: Equivariant neural operators under arbitrary lie groups. In: *International Conference on Learning Representations (ICLR)* (2025)
51. Singhal, U., Feng, R., Yu, S.X., Prakash, A.: Test-Time Canonicalization by Foundation Models for Robust Perception. In: *International Conference on Machine Learning (ICML)* (2025)
52. Sobol', I.: On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* (1967)
53. Spall, J.C.: *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc. (2003)
54. Sun, Y., Guo, C., Li, Y.: ReAct: Out-of-distribution detection with rectified activations. In: *Conference on Neural Information Processing Systems (NeurIPS)* (2021)
55. Sun, Y., Li, Y.: DICE: Leveraging sparsification for out-of-distribution detection. In: *European Conference on Computer Vision (ECCV)* (2022)
56. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: *International Conference on Machine Learning (ICML)* (2022)
57. Tahmasebi, B., Jegelka, S.: Generalization bounds for canonicalization: A comparative study with group averaging. In: *International Conference on Learning Representations (ICLR)* (2025)
58. Wang, H., Li, Z., Feng, L., Zhang, W.: ViM: Out-Of-Distribution with Virtual-logit Matching. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022)
59. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., PENG, WENXUAN., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., Liu, Z.: OpenOOD: Benchmarking generalized out-of-distribution detection. In: *Conference on Neural Information Processing Systems (NeurIPS)* (2022)
60. Yung, J., Romijnders, R., Kolesnikov, A., Beyer, L., Djolonga, J., Houlsby, N., Gelly, S., Lucic, M., Zhai, X.: SI-Score: An image dataset for fine-grained analysis of robustness to object location, rotation and size. In: *ICML Shift Happens Workshop* (2022)
61. Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Liu, x., Han, S., Zhang, D.: Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In: *International Conference on Learning Representations (ICLR)* (2023)
62. Zhirong Wu, Song, S., Khosla, A., Fisher Yu, Linguang Zhang, Xiaoou Tang, Xiao, J.: 3D ShapeNets: A deep representation for volumetric shapes. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015)
63. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the Continuity of Rotation Representations in Neural Networks. *ArXiv* (2020)