MIDAS: Manifold-Inspired Diverse Feature Acquisition for Simplicity Bias Mitigation

Bhartendu Kumar, Vivek B S, Jayavardhana Gubbi, Arpan Pal
Tata Consultancy Services, India
bhartenduk@iisc.ac.in, {vivekb.s31, jay.gubbi, arpan.pal}@tcs.com

***Abstract*—A comprehensive understanding of what Deep Neural Networks (DNNs) learn remains elusive. While the Manifold Hypothesis offers a compelling geometric lens, existing theories often assume smooth manifolds, frequently overlooking non-smooth components such as ReLU and max-pooling. In this work, we present a theoretical framework modeling DNN layers as piecewise-smooth mappings between *stratified spaces*. Our derivation suggests that activations from piecewise-smooth nonlinearities naturally admit a stratified-space structure, extending standard smooth manifold assumptions. Theoretically, we establish a non-increasing layerwise upper bound on the Intrinsic Dimensionality (ID) of representations. We identify *excessive geometric compression* as the mechanism behind Simplicity Bias (SB): networks shed complexity, latching onto simple, low-dimensional features (low ID) at the expense of task-relevant ones. Leveraging these insights, we propose MIDAS (Manifold-Inspired Diverse Feature Acquisition). First, we introduce an annotation-free bias metric based on the *ID gap* (analogous to the Generalization Gap), quantifying the deficit between learned representation dimensionality and the data's intrinsic complexity. Second, we propose a regularization objective that explicitly counteracts ID collapse to promote high-complexity feature learning. Empirically, MIDAS achieves significant gains (up to 10%) on CelebA, Waterbirds, Colored MNIST, and BAR over label-free baselines, approaching fully-supervised performance. Our framework establishes intrinsic dimension as both a diagnostic and a corrective tool for representation quality in deep networks.**

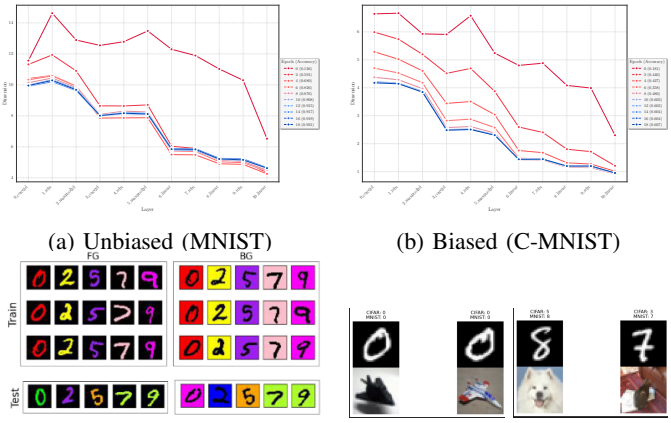
***Index Terms*—Simplicity Bias, Shortcut Learning, Manifold Learning, Intrinsic Dimension, Bias Mitigation, Deep Learning**

I. Introduction

Deep Neural Networks (DNNs) exhibit a critical vulnerability: *Simplicity Bias* (SB) [1], prioritizing simple correlations (e.g., color) over task-relevant structure (e.g., shape) to minimize loss [2]. Existing explanations [3] typically lack a unified geometric framework linking this complexity to representation structure.

We model layers as mappings between *stratified spaces* (Sec. III), identifying SB as *excessive geometric compression*: biased networks collapse data onto lower-dimensional manifolds (e.g., 1D color) insufficient for the task, unlike robust models (Figure 1(a-b)). This collapse poses risks: e.g., medical models latching onto scanner tags, or autonomous vehicles prioritizing road texture over geometry.

We propose **MIDAS** (Manifold-Inspired Diverse Feature Acquisition) to reverse this by regulating layerwise Intrinsic Dimension ($\mathcal{L}_{\text{dim}}$) and enforcing utilization via weight entropy ($\mathcal{L}_{\text{WE}}$). Supported by an **annotation-free bias metric** ($\mathcal{B}$) and diagnostic validation (Figure 1(c-d)), we contribute:



(a) Unbiased (MNIST) (b) Biased (C-MNIST)

(c) C-MNIST Data (d) MNIST-CIFAR

Fig. 1: **Simplicity Bias as Geometric Collapse.** (a-b) Tracking Intrinsic Dimension (ID): Unbiased models maintain high ID (shape), while biased models collapse to ID $\approx$ 1 (color). (c-d) Diagnostic datasets.

- **Theoretical Unification**: We frame Simplicity Bias as stratified manifold collapse, extending the Manifold Hypothesis to ReLU and Pooling singularities.
- **Diagnostic Metric**: MIDAS introduces a label-free complexity metric for early bias detection, quantifying geometric degradation.
- **Robust Methodology**: We propose a dimensionality-preserving regularizer, MIDAS, preventing collapse. It achieves competitive label-free performance (Waterbirds, BAR), recovering up to 84% of the supervision gap.

Due to institutional policies, code cannot be publicly released; however, Algorithm 1 and detailed hyperparameter settings ensure reproducibility.

II. Related Work

Our work intersects with three distinct research areas: (a) Debiasing Methods (without, with partial, or with full bias labels); (b) Manifold Learning in DNNs; and (c) Simplicity Bias. We unify these by linking SB to representation geometry and repurposing Intrinsic Dimension (ID) for both diagnosis and mitigation.

A. Simplicity Bias and Spurious Correlations

Deep networks often latch onto easy-to-learn, *low-complexity features* (e.g., color, texture) at the expense of *high-complexity,*

task-relevant ones (e.g., shape) [1], [2]. Recent work has further characterized how shortcut cues are selected from a parameter-space perspective [4] and how multi-scale feature learning dynamics interact with simplicity bias [5]. While mechanisms like *Gradient Starvation* [3] offer optimization-centric perspectives, they often treat the representation space as an unstructured black box. Our work provides the missing geometric counterpart: we link SB directly to *manifold collapse*, showing that low-complexity features tend to collapse onto lower-dimensional strata in the activation space [6].

B. Manifold Learning and Intrinsic Dimension (ID)

The Manifold Hypothesis—that high-dimensional data lies on low-dimensional structures—is central to understanding DNN generalization [7], [8]. Notably, [9] also link bias to global effective rank. We generalize this by lifting the analysis to *Stratified Spaces* [10], [11], a rigorous topological setting encompassing disjoint unions of non-linear manifolds and the critical singularities naturally induced by ReLU-like activations, capturing structure that Riemannian methods overlook. Thus, MIDAS extends the rank-bias hypothesis to generic piecewise-smooth geometry using local estimates of Intrinsic Dimension (ID) [12] to address dimension collapse where global rank measures fail. Unlike prior work, we repurpose ID tracking as a real-time *diagnostic tool for spurious correlations*, connecting low ID to spurious feature reliance. We uniquely introduce the ID-gap diagnostic (using a shuffled-label baseline) and layer-wise ID regularization tied to an input-ID target.

C. Bias Mitigation Strategies

Solutions follow a spectrum of supervision requirements: **No Bias Labels (Label-Free):** Our focus. Methods like *LfF* [13] and *Spectral Decoupling* [3] rely on heuristics (e.g., "bias is learned first"). While effective, these heuristics can brittle. MIDAS instead uses a precise *geometric* inductive bias (ID optimization), proving more robust across diverse tasks. **Partial/Full Bias Labels:** Methods like *Group DRO* [14], *IRM* [15], and *JTT* [16] require expensive group annotations. While serving as upper bounds, MIDAS aims to approach their robustness without the supervision cost. Similarly, works like DFR [17] and CNC [18] show ERM learns useful features but fails to select them, often needing second-stage retraining; MIDAS solves this selection *online* via geometric regularization, avoiding multi-stage pipelines. Notably, DFR requires a group-balanced validation set for last-layer retraining, and CNC relies on group-labeled contrastive sampling, placing both in the partial/full supervision regime rather than the label-free setting we target.

D. Theoretical Connections

Disentanglement and Invariance. Our geometric perspective aligns with *disentangled representation learning* [19]. Standard disentanglement aims to separate independent factors (e.g., shape vs. color) into orthogonal subspaces; Simplicity Bias is thus an extreme failure where the "hard" factor is discarded. MIDAS acts as a *coarse-grained disentanglement enforcer*: by demanding high Intrinsic Dimension, it forces the encoding of *redundant* factors (both shape and color). This redundancy is safer than invariance-based nuisance suppression when the distinction between signal and noise is ambiguous.

Neural Collapse (NC). Recent work on *Neural Collapse* (NC) [20] shows features collapse to class means terminally. While desirable for the final layer (separability), premature collapse in early layers causes Simplicity Bias by erasing fine-grained variations. MIDAS distinguishes "good collapse" (task-relevant compression) from "bad collapse" (spurious dimension reduction). We balance these forces: $\mathcal{L}_{dim}$ prevents premature backbone collapse, while $\mathcal{L}_{WE}$ facilitates readout separability, resolving the compression-expansion trade-off by depth localization.

III. Theoretical Framework

The Manifold Hypothesis [7] posits that high-dimensional real-world data concentrates near a lower-dimensional manifold. While this perspective has been fruitful, standard smooth manifold theory (assuming C^∞ everywhere) is insufficient for modern Deep Neural Networks (DNNs). Common architectural components such as ReLU and max-pooling are only piecewise smooth/piecewise linear and introduce non-differentiable singularities ("kinks"), rendering the resulting feature spaces *piecewise* smooth rather than globally smooth [10]. (Average pooling is linear and thus smooth.) To rigorously model this, we propose modeling both data and representations as **Stratified Spaces**. This formulation allows us to rigorously handle the singularities induced by deep networks while preserving the tools of differential geometry. This departure from standard theory fills a critical gap: prior works typically assume manifold-to-manifold transformations [8], which fails to account for the dimensionality collapse induced by non-differentiable activations.

A. Preliminaries and Definitions

To rigorously model the non-smooth geometry of deep networks, we rely on the theory of o-minimal structures and stratified spaces.

Definition 1 (O-minimal Structure)**.** *An* o-minimal structure *on $\mathbb{R}$ is a collection of sets $\mathcal{S} = \{S_n\}_{n\in\mathbb{N}}$ where each S_n is a Boolean algebra of subsets of $\mathbb{R}^n$, closed under products and projections, and containing all algebraic sets. Crucially, the sets in S_1 are exactly the finite unions of points and intervals. This "tameness" axiom excludes pathological behaviors (e.g., infinite oscillations like $\sin(1/x)$) while encompassing the piecewise-linear and piecewise-smooth geometries defined by standard neural network activations [21].*

Definition 2 (Stratified Space)**.** *A* stratified space $\mathcal{X}$ *is a topological space admitting a locally finite partition $\mathcal{X} = \bigsqcup_{\alpha\in\Lambda} S_\alpha$, where each* stratum S_α *is a connected smooth manifold. This decomposition must satisfy the* frontier condition*: For any two strata S_α, S_β, if $S_\alpha \cap \overline{S_\beta} \neq \emptyset$, then $S_\alpha \subset \overline{S_\beta}$ and $\dim(S_\alpha) < \dim(S_\beta)$ unless $\alpha = \beta$. This condition ensures*

that the "kinks" or singularities where smooth pieces join are structurally well-behaved [11].

Definition 3 (Intrinsic Dimension (ID))**.** *For any point x lying in a smooth stratum S_α of $\mathcal{X}$, the local Intrinsic Dimension is defined as $ID(x) = \dim(T_x S_\alpha)$, where $T_x S_\alpha$ is the tangent space. This value proxies representational complexity. We estimate it empirically using the maximum likelihood estimator (MLE) over local neighborhoods [12].*

The stratified structure is fundamental to DNNs: the "linear regions" of a ReLU network form the high-dimensional strata, while the activation boundaries form lower-dimensional strata.

B. Theoretical Modeling of Simplicity Bias

We characterize the learning process of DNNs through the lens of mappings between stratified spaces.

Assumption 1 (Data Structure)**.** *The input data support $\mathcal{M} \subset \mathbb{R}^d$ is a **compact definable set** within an o-minimal structure. Definability ensures $\mathcal{M}$ admits a finite Whitney stratification into smooth manifolds [11], avoiding pathological geometries. Compactness ensures the stability of these stratified structures under continuous transformations.*

Assumption 2 (Network Definability)**.** *The neural network F is a composition of definable functions (e.g., affine, ReLU, Sigmoid). This guarantees F is a definable map, preserving the stratified structure of the domain.* Remark: Standard components like BatchNorm are almost-everywhere smooth (definable) and treated as fixed affine maps during inference, satisfying this condition.

Theorem 1 (Preservation of Stratified Structure)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}^k$ be a neural network constructed from definable components (e.g., affine transformations, convolution, ReLU, max/average pooling, Sigmoid, Tanh). If the input support $\mathcal{M}$ is a definable set (and thus a stratified space), then the output feature space $\mathcal{Y} = F(\mathcal{M})$ is also a definable stratified space.*

Proof Sketch. The proof relies on Tame Topology within an O-minimal structure (e.g., $\mathbb{R}_{\text{alg}}$ for ReLU, $\mathbb{R}_{\text{an,exp}}$ for Sigmoid/Tanh) [21]. 1) **Definability:** Since standard activations and affine transformations are definable functions, and definability is preserved under finite composition, the entire network F is a definable map. 2) **No Pathological Explosion:** Crucially, O-minimal structures exclude pathological sets like space-filling curves (Peano curves) that could artificially inflate dimension. The *Monotonicity Theorem* ensures that definable functions are piecewise continuous and monotonic, forbidding infinite oscillations. 3) **Stratified Decomposition:** The *Definable Stratification Theorem* [11] guarantees that both the domain $\mathcal{M}$ and the map F admit a compatible finite Whitney stratification. Specifically, $\mathcal{M}$ partitions into strata $\{S_\alpha\}$ such that restricted maps $F|_{S_\alpha}$ are C^p-smooth with constant rank. The image $F(\mathcal{M})$ is thus a finite union of the images of these strata (immersed submanifolds), forming a well-defined stratified space without dimensional pathologies. □

Intuition: Any neural network built from standard components (affine maps, ReLU, pooling) preserves the piecewise-smooth geometric structure of its input—the output space is always well-behaved and never pathologically complex.

Theorem 2 (ID Monotonicity)**.** *Let F be a composition of layers. For almost every point $x \in \mathcal{M}$ (specifically, points within smooth strata interiors), the local intrinsic dimension of the representation satisfies a non-increasing bound:*

$$ID(F(x)) \leq ID(x). \tag{1}$$

Proof Sketch. We model the network components using *O-minimal structures*, which enforce "tame topology". Let $\mathcal{S}$ be a stratification of $\mathcal{M}$ compatible with F. For any stratum $S \in \mathcal{S}$, the restriction $F|_S : S \to \mathbb{R}^k$ is smooth. For any regular point $x \in S$, $ID(F(x)) = \operatorname{rank}(d(F|_S)_x)$. Since $\operatorname{rank}(d(F|_S)_x) \leq \dim(T_x S) = ID(x)$, local dimension is non-increasing. Crucially, the O-minimal framework excludes pathological dimension explosions (e.g., Peano curves). Thus, $ID(F(x)) \leq ID(x)$ almost everywhere. □

Intuition: Each network layer can only maintain or reduce local geometric complexity—no layer creates new independent dimensions of variation that were not already present in its input.

C. Mechanism: Simplicity Bias as Geometric Compression

Theorem 2 implies a critical property of deep networks: **information irreversibility**. If an early layer collapses the intrinsic dimension of the representation to r, subsequent layers—being smooth or piecewise-smooth mappings—cannot easily recover the lost degrees of freedom (technically, $ID(F(x)) \leq r$). This characterizes the network as a sequence of *dimension filters*. This geometric filtering aligns with the implicit regularization of Gradient Descent, which prefers low-rank solutions [6]. When a dataset contains both a "simple" feature (e.g., color, living on a 1D manifold) and a "complex" feature (e.g., shape, living on a high-dimensional manifold), the optimizer converges to the solution that utilizes the minimal number of active dimensions to satisfy the loss. Consequently, the network "turns off" the orthogonal directions corresponding to the complex feature, projecting the data onto the lower-dimensional spurious subspace. Once this projection occurs, the complex feature is effectively nulled, and no downstream operation can retrieve it. We interpret Simplicity Bias as a *structural reduction of representational capacity via excessive geometric compression*. This necessitates methods like MIDAS (Sec. IV) that explicitly maintain the manifold's volume against this compressive pressure.

Remark 1 (**Significance and Generality**)**.** *Why stratified spaces matter. Prior geometric analyses of DNNs assume globally smooth manifold-to-manifold mappings [7], [8]. This assumption fails precisely where it matters most: at ReLU activation boundaries and pooling discontinuities—the regions where dimensional collapse initiates. Our stratified-space framework rigorously handles these singularities, providing*

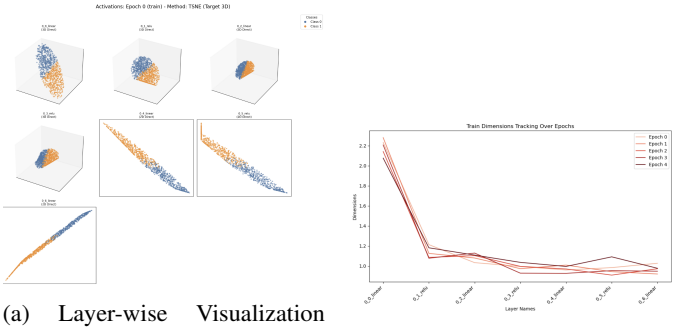
guarantees valid across the entire *representation space, not merely at smooth points. This is what allows Theorem 2 to establish ID monotonicity almost everywhere, including near the critical non-smooth boundaries that smooth theory cannot address.*

Assumptions are generically satisfied. Both assumptions impose minimal restrictions. Compactness (Assumption 1) holds for any bounded dataset—all real-world data is bounded by construction (finite pixel values, bounded sensor ranges). Definability (Assumption 2) is satisfied by every standard architectural component: ReLU, sigmoid, tanh, softmax, affine maps, convolutions, and max/average pooling are all definable in an o-minimal structure [21]. Consequently, the framework applies to virtually all practical DNNs without restricting architecture choice.

From theory to practice. The actionable consequence of Theorem 2 is that a sharp ID drop at any *layer propagates irreversibly to all downstream layers. This transforms a theoretical property into both a measurable diagnostic—the bias metric $\mathcal{B}$ (Sec. IV-A) quantifies cumulative collapse at the final layer—and a direct training objective—$\mathcal{L}_{dim}$ (Sec. IV-C) prevents collapse at its source, bridging geometric theory and practical bias mitigation without requiring bias annotations.*

D. Validation of Theory

We validated these theorems on a 3D Sphere Surface dataset using a 4-layer MLP (3-3-3-2-2 dims, ReLU).



(a) Layer-wise Visualization (Epoch 0)

(b) ID Evolution (Epochs 0–4)

Fig. 2: Empirical validation on 3D Sphere Surface. (a) Visualization of data propagation. (b) Layer-wise ID drops significantly, confirming geometric collapse.

Figure 2 confirms stratified activation structures (Theorem 1) and progressive dimensional collapse (Theorem 2) where representations shrink to a 1D line as training fits the binary labels (ID tracking in Fig. 2b).

IV. MIDAS: Proposed Methodology

Having established that Simplicity Bias manifests as ID collapse (Theorem 2), we now design regularization to prevent this collapse. By imposing geometric constraints that counteract information loss, we ensure the network retains sufficient complexity for robust generalization. To this end, we propose **MIDAS** (Manifold-Inspired Diverse Feature Acquisition), a framework to detect and mitigate this collapse without requiring spurious attribute labels.

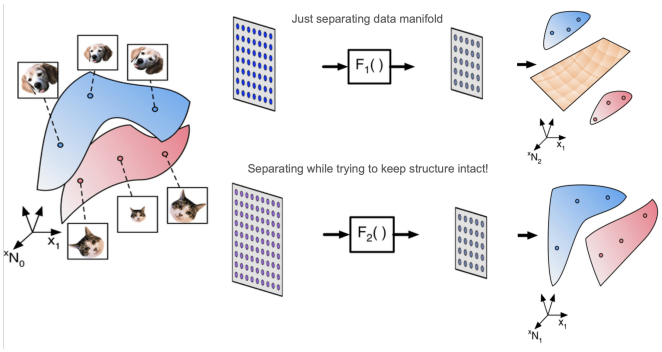


Fig. 3: **MIDAS Overview.** Standard training (top, F_1) collapses the data manifold onto low-dimensional strata, discarding complex features. MIDAS (bottom, F_2) preserves the manifold's geometric structure while still achieving class separation.

A. Bias Quantification Metric ($\mathcal{B}$)

We introduce a metric to quantify SB based on the gap between the learned representation's complexity and the data's intrinsic complexity.

Definition 4 (Bias Metric $\mathcal{B}$)**.** *Let F_{train} be trained on $\mathcal{D}$, and F_{shuff} on $\mathcal{D}_{shuffled}$ (permuted labels). Using the Levina-Bickel MLE estimator [12] for intrinsic dimension ($\widehat{ID}$), we define:*

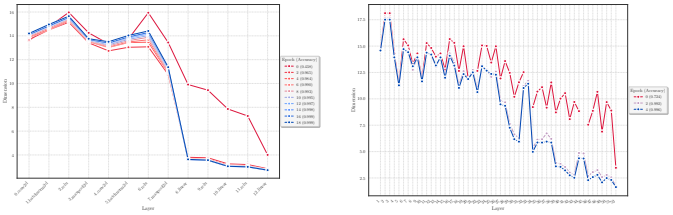
$$\mathcal{B} = \widehat{ID}(\phi_{F_{shuff}}(X)) - \widehat{ID}(\phi_{F_{train}}(X)) \tag{2}$$

where $\phi_F(X)$ are pre-softmax logits. The estimator is given by $\widehat{ID}(x_i) = [(k-1)^{-1} \sum_{j=1}^{k-1} \ln(r_k(x_i)/r_j(x_i))]^{-1}$.

Justification. Shuffling labels destroys semantic correlations, forcing the network to memorize individual samples. To achieve zero training error on noise, deep networks must utilize their maximum available capacity, leading to representations with high Intrinsic Dimension [22]. This establishes a principled, task-independent upper bound on achievable geometric complexity, against which the compression induced by genuine (potentially spurious) correlations can be measured. While our theory applies to any layer (Section III), we compute $\mathcal{B}$ specifically on the pre-softmax logits because manifold collapse is monotonic across the network (Theorem 2). Consequently, the ID reduction is maximal at the final layer, identifying this layer as the most sensitive diagnostic proxy for network-wide bias. In contrast, biased learning compresses inputs onto simple, low-rank manifolds (low ID). Thus, $\mathcal{B} \gg 0$ signals excessive compression and reliance on shortcuts.

B. Empirical Validation of the Metric

To confirm that $\mathcal{B}$ accurately reflects manifold collapse, we validate it on two diagnostic datasets with *known* bias structures: **MNIST-CIFAR** (concatenation of simple MNIST and complex CIFAR-10) and **Colored MNIST (CMNIST)**.



(a) Biased Concept (ID ≈ 5) (b) Unbiased/Shuffled (ID ≈ 10)

Fig. 4: **Validation of $\mathcal{B}$ on MNIST-CIFAR.** (a) Biased training collapses ID to ≈ 5, the complexity of simpler MNIST digits. (b) Unbiased training maintains ID ≈ 10, the complexity of CIFAR-10. The gap $\mathcal{B} \approx 5$ flags the bias.

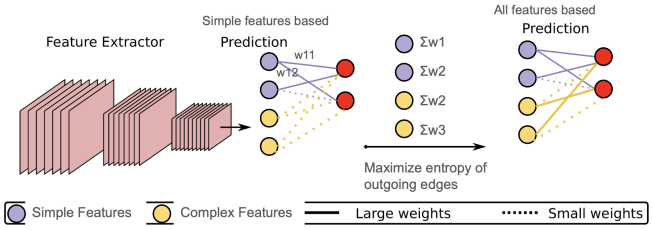


Fig. 5: **Weight Entropy Regularization.** Without $\mathcal{L}_{\text{WE}}$ (left), the classifier assigns large weights only to simple features, ignoring complex ones. $\mathcal{L}_{\text{WE}}$ (right) maximizes the entropy of outgoing edge weights, forcing the classifier to utilize all available feature channels.

As shown in Fig. 4 (for MNIST-CIFAR) and Fig. 6 (for F-CMNIST), biased models consistently collapse. **MNIST-CIFAR:** Collapses to ID ≈ 5 (MNIST). Unbiased maintains ≈ 10 (CIFAR). **F-CMNIST:** Collapses to ID ≈ 1 (Color). Unbiased maintains ≈ 5 (Digit). In all cases, $\mathcal{B} > 0$ correctly flags the presence of simplicity bias, validating our geometric hypothesis.

C. MIDAS Regularization

To counteract the compression identified by $\mathcal{B}$, we introduce two geometric regularizers explicitly added to the standard cross-entropy loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{dim}} + \mathcal{L}_{\text{WE}}$.

1) Dimensionality Regularization ($\mathcal{L}_{dim}$): We directly penalize low-dimensional representations by forcing the layer-wise ID to approach a target complexity.

$$\mathcal{L}_{\text{dim}} = \lambda_{\text{dim}} \sum_{l \in \text{layers}} |\widehat{d}_l - d_{\text{target}}| \tag{3}$$

where $\widehat{d}_l$ is the estimated ID at layer l, and $d_{\text{target}} = \widehat{ID}(\mathcal{D}_{\text{input}})$ is the intrinsic dimension of the raw input. This dynamic target encourages preservation of the data's natural complexity. We anchor to the input ID because it represents the full geometric complexity of the data before any learned compression; targeting a lower value would permit discarding independent factors of variation that may be task-relevant.

2) Weight Entropy Regularization ($\mathcal{L}_{WE}$): While $\mathcal{L}_{\text{dim}}$ ensures that the hidden representations maintain a high intrinsic dimension, it does not guarantee that the final classifier *uses* all these dimensions. An adversary could satisfy $\mathcal{L}_{\text{dim}}$ by encoding high-frequency noise into the features while the classification head continues to rely solely on the simple spurious feature.

To prevent this *feature selection bias*, we maximize the entropy of the weight magnitude distribution in the final linear layer $W \in \mathbb{R}^{C \times D}$. We employ a combined Row-Column Entropy formulation to enforce both feature utilization and class balance:

$$\mathcal{L}_{\text{WE}} = \lambda_{\text{WE}} \left(\sum_{i=1}^{C} \hat{r}_i \log \hat{r}_i + \sum_{j=1}^{D} \hat{c}_j \log \hat{c}_j \right) \tag{4}$$

where $\hat{r}_i = \frac{\sum_j |W_{ij}|}{\|W\|_1}$ is the normalized row mass (class importance) and $\hat{c}_j = \frac{\sum_i |W_{ij}|}{\|W\|_1}$ is the normalized column mass (feature importance). Minimizing this negative entropy forces the classifier to distribute importance across many input channels (Dense Feature Hypothesis), ensuring engagement with the diverse geometric structure preserved by $\mathcal{L}_{\text{dim}}$.

D. MIDAS Algorithm

The complete training procedure is detailed in Algorithm 1. We apply $\mathcal{L}_{\text{dim}}$ across *selected intermediate layers*, as prescribed by Eq. 3, to prevent geometric collapse throughout the network.

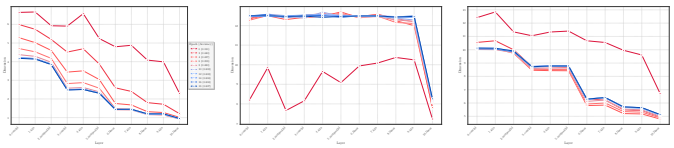
Algorithm 1 MIDAS Training Loop

Input: Dataset $\mathcal{D}$, Network F_θ, Target ID d_{target}, Period T.

for batch (x, y) in $\mathcal{D}$ **do**
  $z = F_\theta(x)$ {Forward pass}
  $\mathcal{L}_{task} = \text{CrossEntropy}(z, y)$
  $\mathcal{L}_{dim} = 0, \mathcal{L}_{WE} = 0$
  if step$\%T == 0$ **then**
    for layer $l \in$ selected layers **do**
      $h_l = \text{Features}_l(x)$ {shape $B \times D_l$}
      $\widehat{d}_l \leftarrow \text{MLE}(h_l, k = 20)$ {Levina-Bickel}
    end for
    $\mathcal{L}_{dim} = \lambda_{dim} \sum_l |\widehat{d}_l - d_{target}|$
    $W \leftarrow$ ClassifierWeights
    $\mathcal{L}_{WE} = \text{Eq.}4(W)$
  end if
  $\mathcal{L}_{total} = \mathcal{L}_{task} + \mathcal{L}_{dim} + \mathcal{L}_{WE}$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{total}$
end for

E. Implementation and Computational Complexity

Manifold Estimation Cost. The intrinsic dimension estimation relies on the k-nearest neighbor (k-NN) search, which introduces a computational complexity of $O(B^2 \cdot D)$ per batch, where B is the batch size and D is the feature dimension. For large-scale training, this overhead can be non-negligible. To mitigate this impact while preserving the regularizer's effectiveness, we employ **Periodic Computation**, calculating and optimizing the $\mathcal{L}_{\text{dim}}$ term only every $T = 10$ optimization steps. Since the manifold geometry of the global representation



(a) Biased (b) MIDAS (c) Oracle

Fig. 6: **MIDAS Effect (F-CMNIST).** (a) Biased training collapses ID to ≈ 1. (b) MIDAS preserves ID ≈ 8, recovering the structural complexity seen in (c) the unbiased oracle.

space evolves slowly relative to the high-frequency weight updates, this approximation introduces negligible error. In practice, this results in approximately 15% additional training time in our experiments, which is lower than auxiliary-model methods (e.g., LfF [13]) or multi-stage pipelines (e.g., JTT [16], DFR [17]).

Training Dynamics. MIDAS is trained end-to-end in a single stage, avoiding the complexity of multi-round training. By explicitly maintaining target dimensionality, the regularizer acts as a spectral constraint on feature transformations, helping counteract the rank collapse often associated with sharp minima.

F. Validation of Regularization

We demonstrate the effectiveness of this approach on diagnostic datasets.

As seen in Fig. 6, applying MIDAS effectively restores the intrinsic dimension. Quantitative results (Table I) show MIDAS achieves 53.7% on F-CMNIST, matching the supervised IRM baseline (52.8%) without using group labels.

TABLE I: Diagnostic Validation Accuracy (%). MIDAS recovers performance on standard benchmarks.

Method	F-CMNIST	B-CMNIST	MNIST-CIFAR
ERM	11.1	11.2	11.5
MIDAS	**53.7**	**34.6**	**32.1**
IRM (Supervised)	52.8	42.8	34.2

V. Experiments

Having validated MIDAS on controlled diagnostic datasets (Sec. IV), we now demonstrate its effectiveness on challenging real-world benchmarks. We evaluate on four standard datasets spanning diagnostic (**F-CMNIST**), subpopulation shift (**Waterbirds**, **CelebA**), and real-world action recognition (**BAR**) domains. To ensure rigorous comparison, we benchmark against methods representing distinct debiasing paradigms: *optimization-based* (SD), *reweighting-based* (LfF, JTT, Group DRO), and *invariance-based* (IRM). Our focus is on the **label-free regime**, where spurious attribute annotations are unavailable—the practical setting for most real-world deployments.

A. Experimental Setup

Datasets and Metrics. We evaluate on four benchmarks: (1) **F-CMNIST** [23] (Metric: Unbiased Acc); (2) **Waterbirds** [14], [24] & (3) **CelebA** [25] (Metric: Worst Group Acc); (4) **BAR** [13] (Metric: Unbiased Acc).

Baselines and Protocol. We adhere to the **Label-Free** adaptation setting implies:

- **Training**: No access to spurious attribute labels (a) or group labels (g).
- **Validation/Test**: Attribute labels are used *only* for model selection and evaluation (e.g., computing worst-group accuracy).

Baselines are categorized by valid inputs under this protocol:

- **ERM**: Standard training (Lower Bound).
- **Label-Free (Direct Comp.)**: **SD** [3], **LfF** [13].
- **Oracle (Upper Bounds)**: **Group DRO** (uses g in train) [14], **IRM** (uses e/g in train) [15], **HEX** (uses partial a in train) [26].

Implementation. We utilize ResNet-18 backbones for CelebA, Waterbirds, and BAR, and LeNet-5 for F-CMNIST. Training employs the LAMB optimizer [27] with large batch sizes ($B \in [1024, 65536]$) to ensure stable Intrinsic Dimension estimation. We utilize the Levina-Bickel estimator [12] with $k = 20$ nearest neighbors on flattened layer representations. We found that standard SGD with smaller batches leads to high variance in ID estimates, as the local neighborhood structure is undersampled. Large-batch training, facilitated by LAMB's layer-wise adaptive moments, stabilizes the k-NN graph construction and improves the reliability of the $\mathcal{L}_{\text{dim}}$ gradient signal. Baseline methods were trained using a combination of their published protocols and re-tuned hyperparameters; MIDAS requires large batch sizes specifically for stable k-NN ID estimation, not for optimization advantage. All hyperparameters were rigorously tuned on validation sets.

B. Results and Analysis

Table II presents the comprehensive benchmarking.

1. Performance on Synthetic Benchmarks (F-CMNIST). On the diagnostic F-CMNIST dataset, where the spurious color correlation is explicitly designed to be stronger than the digit signal, MIDAS achieves **53.7%** unbiased accuracy. This is a remarkable improvement over the ERM baseline (11.2%), which essentially acts as a random classifier for digits. Furthermore, MIDAS outperforms the specialized debiasing method LfF (53.1%) and significantly surpasses Spectral Decoupling (47.8%). This confirms that explicitly enforcing manifold complexity is more effective than implicitly penalizing gradient norms when the spurious feature is low-dimensional.

2. Scaling to Natural Images (CelebA & Waterbirds). Moving to large-scale vision benchmarks, MIDAS demonstrates robust scalability. On **Waterbirds**, it achieves **85.1%** worst-group accuracy, outperforming all other label-free methods, including SD (82.5%) and LfF (75.2%). On **CelebA**, a challenging face attribute task, MIDAS reaches 80.1%, which is competitive with the top label-free performer SD (81.2%) and drastically superior to LfF (70.6%). The performance gap between MIDAS and ERM on these tasks underscores the universality of geometric collapse as a failure mode in CNNs.

3. Robustness on Real-World Shifts (BAR). The BAR dataset represents a "wild" distribution shift scenario involving

TABLE II: Debiasing performance across standard benchmarks. We classify methods by their supervision requirements. **Metrics:** We report *Test Accuracy* for F-CMNIST/BAR and *Worst Group Accuracy* for CelebA/Waterbirds, adhering to community standards. MIDAS achieves competitive performance among the representative label-free methods we evaluate (winning 3/4 tasks) and rivals full supervision on diagnostic datasets. Results are reported from single runs following established protocols for each method. *Note:* N/A indicates the method cannot be applied due to missing bias annotations. Specifically, IRM and Group DRO require group labels unavailable in CelebA/Waterbirds (under our protocol) or BAR; we contrast MIDAS against them only where explicit skylines are established.

	ERM	**MIDAS (Ours)**	**SD**	**LfF**	**HEX**	**JTT**	**Group DRO**	**IRM**
Dataset	(Baseline)	(No Bias Labels)			(Partial Labels)		(Full Labels)	
Metric: Unbiased Test Accuracy (↑)								
F-CMNIST	11.2	**53.7**	47.8	53.1	35.1	37.9	50.7	52.8
MNIST (Ref)	99.2	99.3	99.1	99.2	N/A	N/A	N/A	N/A
BAR	35.3	**50.1**	43.7	48.1	N/A	N/A	N/A	N/A
Metric: Worst Group Accuracy (↑)								
CelebA	40.1	80.1	81.2	70.6	50.7	81.1	**87.7**	N/A
Waterbirds	72.6	85.1	82.5	75.2	80.3	86.0	**91.4**	N/A

complex action recognition biases. Here, MIDAS achieves **50.1%** accuracy, surpassing the competing label-free methods in our evaluation. It outperforms LfF (48.1%) and SD (43.7%) by a significant margin. Since BAR lacks explicit group annotations, supervised methods (JTT, Group DRO) are inapplicable, highlighting the critical value of MIDAS's unsupervised geometric approach in real-world deployments where bias labels are often nonexistent.

4. Recovering Fully-Supervised Performance. MIDAS rivals methods requiring costly annotation. On **F-CMNIST**, it outperforms Group DRO (**53.7%** vs 50.7%). On **CelebA** and **Waterbirds**, it closes $\approx 84\%$ and $\approx 66\%$ of the gap between ERM and fully-supervised Group DRO, respectively, effectively recovering robustness at zero annotation cost.

5. Single-Stage Computational Efficiency. Existing methods often trade efficiency for robustness—LfF trains two concurrent networks (biased and debiased), while JTT requires two sequential training stages. In contrast, MIDAS operates as a **single-model training pass** with only periodic geometric-loss computation. This makes it uniquely suitable for large-scale foundational model pre-training, where multi-stage or multi-model approaches are computationally prohibitive.

C. Ablation Analysis

To understand the specific contributions of our geometric components, we conduct a detailed ablation study on the F-CMNIST benchmark. This dataset provides a controlled environment where the "simple" feature (color, ID ≈ 1) and "complex" feature (digit, ID ≈ 5) are precisely defined.

1. Impact of Dimensionality Regularization ($\mathcal{L}_{\text{dim}}$). This term is the primary engine of bias mitigation. **Without $\mathcal{L}_{\text{dim}}$:** Baseline models collapse to $\approx 11\%$ accuracy with ID ≈ 1. **With $\mathcal{L}_{\text{dim}}$ Only:** By enforcing a target dimension $d_{target} \approx 10$ (estimated from raw input $\mathcal{D}_{\text{input}}$), we strongly discourage the model from relying solely on the 1D color feature. It is *driven* to encode additional variations (like shape) to satisfy the geometric constraint, achieving plateau ID ≈ 8.

2. Impact of Weight Entropy ($\mathcal{L}_{\text{WE}}$). While $\mathcal{L}_{\text{dim}}$ ensures that diverse features exist, it does not guarantee their utilization. We observed that standard linear classifiers often assign near-zero weights to "hard" features even when present. Adding $\mathcal{L}_{\text{WE}}$ corrects this by maximizing the entropy of weight magnitudes, forcing the classifier to aggregate evidence from *all* available channels. This yields the final performance boost to **53.7%**.

Sensitivity to Hyperparameters. We extensively tuned the regularization strength λ_{dim} and the estimator neighbor count k. **Regularization Strength (λ_{dim}):** MIDAS is stable for a range of λ_{dim}. Setting λ_{dim} too high degrades performance as models encode noise to artificially inflate dimension. Conversely, very low λ_{dim} cannot overcome the shortcut's strong gradient attraction. **Estimator Neighbors (k):** The estimator is robust to choices of k. Very small k leads to high variance estimates, while very large k oversmooths local structures. We standardized k for all main experiments without per-dataset tuning.

D. Qualitative Analysis: Manifold Recovery

Visualizing the learned manifolds (t-SNE) confirms our theory. (1) **ERM (Collapsed):** Features cluster primarily by color, with significant overlap between digit classes within color groups. The manifold is topologically equivalent to a set of disjoint color blobs. (2) **MIDAS (Unfolded):** Representations clearly separate digit classes even within the same color group. The clusters are topologically distinct, forming a higher-dimensional structure where 'shape' defines the neighborhood. ID tracking (Fig. 6) shows dimension plateauing at a higher value (≈ 8), indicating the manifold has physically "unfolded" to accommodate shape adaptation. This expansion enables the linear classifier to form decision boundaries robust to randomized color correlations.

Robustness-Accuracy Trade-off. Robust optimization often degrades "clean" accuracy. Uniquely, MIDAS maintains standard MNIST accuracy (99.3%, Table II) comparable to ERM (99.2%). This suggests geometric regularization encourages *redundancy*—learning *both* simple and complex features—avoiding the exclusive reliance typical of simplicity bias.

E. Limitations and Discussion

The Validity of the Geometric Assumption. MIDAS posits that task-relevant features (e.g., shape) possess a higher ID than spurious correlates (e.g., color). This aligns with

the physical generation of data: shape involves multiple independent factors unlike global biases. However, if a spurious feature is high-dimensional (e.g., complex adversarial noise), geometric regularization might inadvertently preserve it. In such regimes, MIDAS must be coupled with causality-based independence tests.

Computational Scalability. A limitation is the $O(B^2)$ complexity of the k-NN graph for ID estimation. While periodic computation (every $T = 10$ steps) makes this negligible for standard benchmarks, scaling to massive Foundation Models would require linear-time approximations, such as anchor-based estimation relative to a fixed set of reference points.

Geometric Regularization vs. Loss Landscapes. Maintaining high intrinsic dimension may relate to avoiding collapsed Hessian spectra and favoring flatter minima, though this connection remains a direction for future investigation.

VI. Conclusion

We proposed a rigorous stratified geometric framework characterizing *monotonic manifold collapse* as a driver of Simplicity Bias. Our analysis shows layers acting as dimension filters, discarding complex task features. To counter this, we introduced **MIDAS**, a unified bias mitigation suite comprising a label-free ID-gap diagnostic and a dual-objective geometric regularizer. By enforcing effective dimensionality and weight entropy, MIDAS prevents collapse onto low-rank spurious manifolds. Empirically, MIDAS demonstrates competitive label-free debiasing (Waterbirds, BAR), approaching fully supervised methods.

Unlike prior label-free methods that rely on heuristics about training dynamics, MIDAS operates on a *geometric invariant*—intrinsic dimension—that is architecture- and task-agnostic. More broadly, the ID-monotonicity principle is not specific to simplicity bias: it positions intrinsic dimension tracking as a general-purpose diagnostic for representation health, capable of detecting premature information collapse regardless of its cause. Future work extends this perspective to Transformer architectures and curvature-based analysis.

Acknowledgments

Portions of this manuscript were refined with the assistance of AI-based language tools for grammar and style editing. All technical content, experimental results, and scientific claims were developed and verified by the authors.

References

[1] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli, "The pitfalls of simplicity bias in neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9573–9585, 2020.

[2] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018.

[3] M. Pezeshki, O. Kaba, Y. Bengio, A. C. Courville, D. Precup, and G. Lajoie, "Gradient starvation: A learning proclivity in neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1256–1272, 2021.

[4] L. Scimeca, S. J. Oh, S. Mishra, S. Sahoo, and S. Kak, "Which shortcut cues will DNNs choose? A study from the parameter-space perspective," in *International Conference on Learning Representations*, 2024.

[5] M. Pezeshki, A. Mila, Y. Bengio, and G. Lajoie, "Multi-scale feature learning dynamics: Insights for double descent," in *International Conference on Machine Learning*, 2023.

[6] M. Huh, H. Mobahi, R. Zhang, B. Cheung, P. Agrawal, and P. Isola, "The low-rank simplicity bias in deep networks," *arXiv preprint arXiv:2103.10427*, 2021.

[7] M. Chen, H. Jiang, W. Liao, and T. Zhao, "Efficient approximation of deep relu networks for functions on low dimensional manifolds," *Advances in neural information processing systems*, vol. 32, 2019.

[8] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, "The intrinsic dimension of images and its impact on learning," *arXiv preprint arXiv:2104.08894*, 2021.

[9] G. Y. Park, C. Jung, S. Lee, J. C. Ye, and S. W. Lee, "Self-supervised debiasing using low rank regularization," *arXiv preprint arXiv:2210.05248*, 2022.

[10] J. Schmidt-Hieber, "Deep relu network approximation of functions on a manifold," *arXiv preprint arXiv:1908.00695*, 2019.

[11] H. Whitney, "Tangents to an analytic variety," *Annals of mathematics*, pp. 496–549, 1965.

[12] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," *Advances in neural information processing systems*, vol. 17, 2004.

[13] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin, "Learning from failure: Training debiased classifier from biased classifier," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 673–20 684, 2020.

[14] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *arXiv preprint arXiv:1911.08731*, 2019.

[15] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[16] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn, "Just train twice: Improving group robustness without training group information," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6781–6792.

[17] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," in *International Conference on Machine Learning*. PMLR, 2022, pp. 11 184–11 202.

[18] M. Zhang, N. S. Sohoni, H. R. Zhang, C. Finn, and C. R'e, "Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 484–26 516.

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[20] V. Papyan, X. Y. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, pp. 24 652–24 663, 2020.

[21] L. Van den Dries, *Tame topology and o-minimal structures*. Cambridge university press, 1998, vol. 248.

[22] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[23] S. Addepalli, A. Nasery, R. V. Babu, P. Netrapalli, and P. Jain, "Learning an invertible output mapping can mitigate simplicity bias in neural networks," *arXiv preprint arXiv:2210.01360*, 2022.

[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," California Institute of Technology, Tech. Rep., 2011.

[25] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[26] H. Wang, Z. He, Z. C. Lipton, and E. P. Xing, "Learning robust representations by projecting superficial statistics out," *arXiv preprint arXiv:1903.06256*, 2019.

[27] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large batch optimization for deep learning: Training BERT in 76 minutes," *arXiv preprint arXiv:1904.00962*, 2019.