UNDERSTANDING AND MITIGATING CATASTROPHIC INHERITANCE VIA DISCRIMINATIVE SPECTRUM DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-training on large-scale data has become a standard paradigm for visual transfer learning, but noisy supervision in pre-training data can be inherited by downstream models and degrade their generalization. While prior studies have shown that noisy pre-training affects downstream transfer, its structural effect on downstream representations remains insufficiently understood. In this work, we recast catastrophic inheritance as *discriminative spectrum drift*, and study how noisy pre-training reshapes class-relevant discriminative information in spectral space. Across multiple commonly used cross-domain visual datasets and controlled noisy pretrained checkpoints, we show that the inherited effect of noisy pre-training is not merely reflected in performance variation, but manifests as a consistent drift of the downstream discriminative spectrum. To characterize this phenomenon, we introduce two continuous descriptors that capture the overall deviation from the clean reference and the average spectral position of discriminative information. Motivated by these findings, we further propose a lightweight spectral adaptation framework for the realistic setting where no clean reference is available and the pretrained backbone remains frozen. Experiments show that our method mitigates discriminative spectrum drift and improves downstream adaptation performance.

1 INTRODUCTION

Pre-training followed by downstream adaptation has become a standard recipe in modern visual learning. Instead of training a separate model from scratch for each task, current practice is to first pre-train a large backbone on large-scale data and then transfer it to diverse downstream problems (Kornblith et al., 2019). This paradigm underlies many widely used model families, including ResNets (He et al., 2016), Vision Transformers (Dosovitskiy, 2020), and CLIP-style vision–language models (Radford et al., 2021), and has substantially improved scalability, transferability, and data efficiency in real applications (Bommasani et al., 2021). As foundation-style models continue to grow in size and scope (Bommasani et al., 2021), the pre-training stage increasingly relies on web-scale corpora such as ImageNet (Russakovsky et al., 2015), YFCC100M/YFCC15M-style resources (Thomee et al., 2016), and LAION-scale collections (Schuhmann et al., 2022), making the quality of pre-training data an increasingly important factor in downstream behavior.

Yet the data that make large-scale pre-training possible are often weakly curated, weakly annotated, or only loosely aligned, especially when collected from web-scale sources (Thomee et al., 2016; Schuhmann et al., 2022). Such imperfections may not be immediately visible during pre-training, but they can be absorbed into the learned representation and continue to influence downstream transfer after adaptation (Chen et al., 2023; 2025). As illustrated in Figure 1, increasing pre-training noise already leads to clear degradation on standard downstream transfer benchmarks, and this inherited effect becomes particularly pronounced when the target distribution differs from the pre-training source (Koh et al., 2021; Gulrajani & Lopez-Paz, 2020). Recent studies have begun to characterize this phenomenon as *catastrophic inheritance*, emphasizing that noise or bias introduced at pre-training time can persist into downstream models rather than being removed by transfer alone (Chen et al., 2024; 2025).

Despite the growing attention to pre-training data quality and catastrophic inheritance, existing studies still do not fully answer the question we address here. Prior work has shown that noise or bias in pre-training data can alter downstream accuracy and robustness (Nguyen et al., 2022; Entezari et al.,

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

2023; Recht et al., 2019; Vasudevan et al., 2022), and recent analyses have further linked such effects to coarse-grained changes in representation structure (Chen et al., 2023; 2024; 2025). However, these analyses mostly remain at the level of end-task performance or global representation statistics, such as singular value spectra and overall feature geometry (Chen et al., 2023). While informative, they do not directly reveal how the discriminative information that downstream tasks actually rely on is reorganized after noisy pre-training. In other words, what remains unclear is not only whether noisy pre-training affects downstream transfer, but how it structurally reshapes the task-relevant discriminative information inside the representation.

To address the above gap, this paper is organized around three tightly connected questions: whether noisy pre-training systematically affects downstream transfer, how such influence manifests structurally, and how it can be mitigated under realistic adaptation constraints.

Influence. We first examine whether noisy pre-training has a systematic inherited effect on downstream transfer across commonly used cross-domain visual benchmarks. Our results show that this effect is not merely anecdotal or task-specific, but consistently observable under transfer settings.

Analysis. We then move beyond surface-level performance variation and reinterpret the inherited effect as *discriminative spectrum drift*. Instead of only measuring end-task accuracy or coarse global statistics, we focus on how task-relevant discriminative information is reorganized in spectral space after noisy pre-training, and characterize this process through the overall deviation from the clean reference and the spectral position of discriminative information.

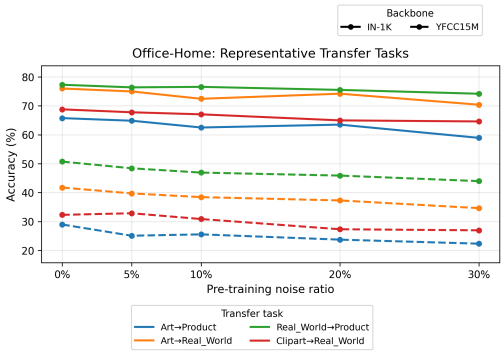

Figure 1: Downstream transfer accuracy on Office-Home with frozen `r50_in1k` and `r50_yfcc15m` backbones. Accuracy generally decreases as pre-training noise increases.

Mitigation. Motivated by this analysis, we finally develop a lightweight mitigation strategy for the practical setting where no clean reference model is available and the pretrained backbone remains frozen. Rather than treating downstream adaptation as generic feature tuning, our method directly regularizes the adaptation toward the structural drift revealed by the analysis, enabling more reliable transfer under noisy pre-training.

More broadly, we view this problem as an important yet still underexplored aspect of foundation-model transfer: imperfections introduced at pre-training time may persist into downstream adaptation rather than being washed out by transfer. This issue is particularly relevant in realistic settings where users only access a pretrained checkpoint, the backbone remains frozen, and adaptation must be lightweight. We hope our study can encourage more structure-aware investigation of inherited representation distortion under noisy pre-training.

2 Understanding Label Noise in Pretrained Models

In this section, we systematically investigate how noisy labels in supervised pre-training affect learned representations. Our evaluation and analysis are built on realistic motivating experiments with ResNet-50 (He et al., 2016) pre-trained on synthetically corrupted ImageNet-1K (Russakovsky et al., 2015) and YFCC15M (Thomee et al., 2016).

2.1 Problem Setup and Experimental Protocol

Noisy pre-training datasets. We study noisy supervised pre-training under controlled corruption. Let $D = \{(x_i, y_i)\}_{i=1}^N$ denote a clean pre-training dataset and $\hat{D} = \{(x_i, \hat{y}_i)\}_{i=1}^N$ its noisy counterpart, where a fraction γ of supervision is corrupted. We consider two representative large-scale pre-training sources: ImageNet-1K (Russakovsky et al., 2015) for standard supervised pre-training and YFCC15M, a subset of YFCC100M (Thomee et al., 2016), for CLIP-style pre-training (Radford et al.,

2021). Following prior work on noisy pre-training (Chen et al., 2023; 2025), we construct controlled noisy checkpoints with noise ratios $\gamma \in \{0\%, 5\%, 10\%, 20\%, 30\%\}$, where $\gamma = 0\%$ corresponds to clean pre-training.

Pre-trained models. We consider two families of pretrained ResNet-50 backbones (He et al., 2016) obtained under different noisy pre-training regimes. The first family, denoted as `r50_in1k`, is trained on ImageNet-1K with standard supervised classification. The second family, denoted as `r50_yfcc15m`, is the visual encoder of a CLIP-style model pretrained on YFCC15M with image–text supervision (Radford et al., 2021). These two families represent two widely used pre-training paradigms in modern transfer learning (Kornblith et al., 2019; Bommasani et al., 2021), and allow us to examine whether the inherited effect of noisy pre-training is consistent across different backbone families.

Experiment setup. We conduct experiments on Office31 (Saenko et al., 2010), which contains three domains with moderate domain shift, and Office-Home (Venkateswara et al., 2017), which is more challenging due to its larger category set and more diverse domain gap. For each pretrained backbone family, we evaluate checkpoints trained with five pre-training noise ratios, from clean to 30%. Unless otherwise specified, the pretrained backbone remains frozen throughout adaptation, and only the downstream classifier is trained. This setup isolates the inherited effect of noisy pre-training from the effect of full fine-tuning, while matching the realistic downstream scenario in which users only access pretrained checkpoints rather than the original pre-training data.

Table 1: Average baseline downstream accuracy (%) under different pre-training noise ratios on Office31 and Office-Home. Each entry is averaged over all transfer tasks and both backbone families.

Dataset	0%	5%	10%	20%	30%	Avg.
Office31	57.21	56.98	54.76	54.14	50.54	54.72
Office-Home	47.01	46.04	45.02	43.94	42.89	44.98

2.2 Results

Table 1 shows that noisy pre-training already leads to clear degradation at the downstream transfer level. As the pre-training noise ratio increases, the average baseline accuracy generally decreases on both Office31 and Office-Home. The trend is especially clear on Office-Home, where the accuracy drops monotonically from 47.01% under clean pre-training to 42.89% at 30% noise. This result suggests that the effect of noisy pre-training is inherited by downstream models rather than eliminated by transfer, motivating a more fine-grained structural analysis of the learned representations.

2.3 From Feature Space to Discriminative Spectrum Drift

The performance results above establish that noisy pre-training leaves a measurable inherited effect on downstream transfer, but they do not explain its structural origin. Accuracy variation only shows that the model is affected; it does not reveal how noisy pre-training reshapes the downstream representation itself. For the problem studied here, the key question is not merely whether transfer performance drops, but whether the discriminative information used by downstream tasks is systematically reorganized.

A natural direction is to analyze the feature space in spectral coordinates. However, ordinary feature spectra mainly reflect variance or energy distribution, which does not necessarily correspond to downstream discriminability. To better align the analysis with downstream classification, we consider a *discriminative spectrum*. Let $h(x) \in \mathbb{R}^D$ denote the penultimate-layer feature of a frozen backbone. After centering all features and performing SVD, we obtain spectral directions $\{v_i\}_{i=1}^D$ and define the coefficient of sample x on the i-th direction as

$$c_i(x) = v_i^\top (h(x) - \mu), \tag{1}$$

where μ is the mean feature vector. We then quantify how much each direction contributes to class discrimination and normalize the resulting direction-wise strengths $\{d_i\}$ as

$$q_i = \frac{d_i}{\sum_{j=1}^D d_j}, \tag{2}$$

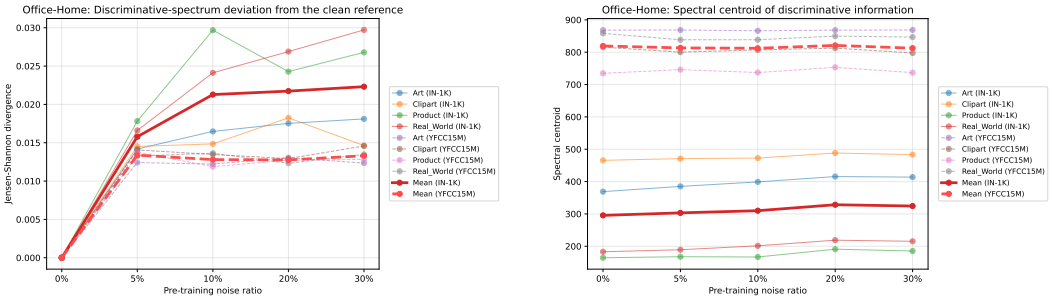

Figure 2: Discriminative-spectrum drift on Office-Home under different pre-training noise ratios. Left: the overall deviation of the discriminative spectrum from the clean reference, measured by Jensen–Shannon divergence. Right: the spectral centroid of discriminative information, which reflects its average spectral position. The red curves denote the mean trends across domains.

which yields a distribution $q = (q_1, \ldots, q_D)$ over spectral directions. We refer to q as the discriminative spectrum.

From this perspective, catastrophic inheritance can be understood as a structured drift of the discriminative spectrum relative to its clean counterpart. To quantify this drift, we use two complementary descriptors. The first measures the overall discrepancy between the discriminative spectrum of a noisy model and that of the clean model. Specifically, let $q^{(\gamma)} = (q_1^{(\gamma)}, \ldots, q_D^{(\gamma)})$ denote the discriminative spectrum of a model pretrained with noise ratio γ, and let $q^{(0)} = (q_1^{(0)}, \ldots, q_D^{(0)})$ denote the corresponding spectrum of the clean model. We characterize their overall difference by the Jensen–Shannon divergence

$$D_{\mathrm{JS}}\left(q^{(\gamma)}, q^{(0)}\right) = \frac{1}{2} D_{\mathrm{KL}}\left(q^{(\gamma)} \,\|\, m\right) + \frac{1}{2} D_{\mathrm{KL}}\left(q^{(0)} \,\|\, m\right), \quad m = \frac{1}{2}\left(q^{(\gamma)} + q^{(0)}\right), \tag{3}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the Kullback–Leibler divergence. A larger value indicates that noisy pre-training has caused a stronger overall deformation of the downstream discriminative spectrum.

The second descriptor captures where discriminative information is concentrated along the spectrum. We compute the spectral centroid of the discriminative spectrum as

$$R_{\mathrm{disc}} = \sum_{i=1}^{D} i\, q_i, \tag{4}$$

which can be interpreted as the average spectral location of class-relevant discriminative information. Smaller values indicate that discriminative mass is concentrated in leading spectral directions. In our analysis, the divergence to the clean reference serves as the primary measure of overall spectral drift, while the spectral centroid provides an interpretable view of whether discriminative information is redistributed toward less dominant directions.

2.4 Empirical Findings of Discriminative Spectrum Drift

Observed phenomena. We first examine the overall deviation of the discriminative spectrum from its clean reference. On both Office31 and Office-Home, the divergence between noisy and clean spectra increases consistently with the pre-training noise ratio, with a particularly clear and nearly monotonic trend for the `r50_in1k` family. This pattern is stable across domains and indicates that noisy pre-training systematically perturbs the downstream discriminative structure rather than introducing random fluctuations. A plausible explanation is that noise in pre-training data disrupts the alignment between feature directions and class-relevant signals, causing discriminative information to be redistributed across spectral directions. As a result, the downstream discriminative spectrum deviates progressively from that of the clean model, leading to an accumulated structural mismatch that is reflected in the increasing divergence.

We further analyze where discriminative information is located along the spectrum. Under the representative `r50_in1k` setting, the spectral centroid shifts steadily toward later directions as the

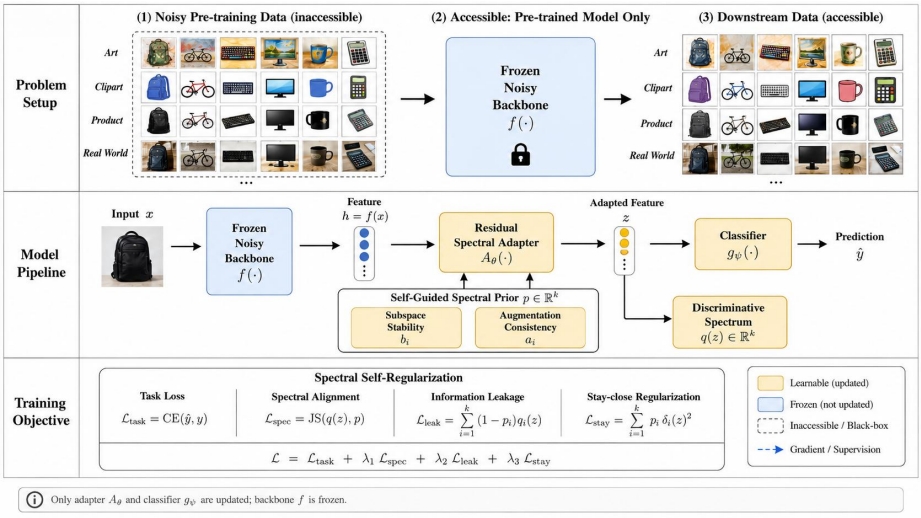

Figure 3: Overview of Self-Guided Spectral Adaptation (SGSA). A lightweight residual adapter adjusts features extracted by a frozen pretrained backbone, and a downstream classifier predicts class labels from the adapted features. Subspace stability and augmentation consistency provide a self-guided spectral prior. The adapter and classifier are trained with the classification objective, spectral alignment, leakage suppression, and stay-close regularization, without requiring a clean reference model.

pre-training noise ratio increases on both datasets. This indicates that discriminative information, which is more concentrated in leading directions under clean pre-training, becomes progressively spread toward less dominant directions when noise is introduced. Such a shift suggests that noisy pre-training weakens the compactness and efficiency of discriminative representation, forcing downstream tasks to rely on a broader and less structured set of feature directions. In this sense, the observed increase in spectral divergence can be interpreted structurally as a backward redistribution of discriminative information, providing a concrete manifestation of discriminative spectrum drift.

3 Mitigating Discriminative Spectrum Drift via Self-Guided Spectral Adaptation

3.1 Motivation

Section 2.4 shows that the effect of noisy pre-training manifests as a structural drift in the downstream discriminative spectrum, rather than a simple classifier mismatch. As the noise ratio increases, the discriminative spectrum deviates from its clean counterpart and, in representative cases such as `r50_in1k`, exhibits a backward spread of discriminative information toward later spectral directions. This implies that training only a downstream classifier is insufficient, since it leaves the underlying feature structure unchanged. Meanwhile, practical adaptation often operates under a no-clean-reference, frozen-backbone setting, where neither explicit alignment to a clean model nor full fine-tuning is feasible. These observations motivate a lightweight adaptation strategy that directly targets the discriminative spectrum drift while preserving the pretrained backbone.

3.2 Spectral Adapter

Let $f(\cdot)$ denote the pretrained feature extractor, whose parameters remain frozen throughout adaptation. For an input sample x, the penultimate-layer feature produced by the backbone is denoted by

$$h = f(x), \qquad h \in \mathbb{R}^D. \tag{5}$$

As shown in Section 2, the effect of noisy pre-training is not merely a shifted decision boundary, but a structural drift already embedded in the discriminative spectrum of the representation. Therefore,

correcting such drift requires more than training a new classifier on top of h; the frozen feature itself must be allowed to change in a limited manner. To this end, we introduce a lightweight spectral adapter $A_\theta(\cdot)$ after h, and define the adapted feature z as

$$z = h + A_\theta(h). \tag{6}$$

The residual form keeps the adaptation centered around the original pretrained feature, rather than replacing it with a completely new representation.

In practice, we instantiate A_θ as a two-layer bottleneck MLP:

$$A_\theta(h) = W_2\, \phi(W_1\, \mathrm{LN}(h))\,, \tag{7}$$

where $\mathrm{LN}(\cdot)$ denotes LayerNorm, $\phi(\cdot)$ is a nonlinear activation function, $W_1 \in \mathbb{R}^{r\times D}$, $W_2 \in \mathbb{R}^{D\times r}$, and $r \ll D$. This bottleneck structure provides sufficient flexibility to adjust the feature geometry, while keeping the number of additional parameters small under the frozen-backbone setting. The adapted feature z is then fed into a downstream classifier $g_\psi(\cdot)$ to produce the prediction

$$\hat{y} = g_\psi(z). \tag{8}$$

The overall adaptation pipeline can thus be written as

$$x \xrightarrow{f} h \xrightarrow{A_\theta} z \xrightarrow{g_\psi} \hat{y}, \tag{9}$$

where only the adapter parameters θ and classifier parameters ψ are updated, while the backbone f remains frozen. This design enables limited but effective correction of the downstream representation without modifying the pretrained model itself, and provides a suitable feature space for the spectral regularization introduced next.

3.3 Self-Guided Spectral Prior

Section 2.4 shows that noisy pre-training induces a stable deviation of the downstream discriminative spectrum from its clean reference, often accompanied by a shift of discriminative information toward later spectral directions. This suggests that mitigation should not treat all feature directions equally, but instead distinguish more reliable directions from those likely to absorb unstable, noise-induced structure. Since no clean reference is available at adaptation time, we extract internal reliability signals from the current model to construct a *self-guided spectral prior*, which encourages discriminative information to remain in more stable spectral directions and thus counteracts the observed drift.

Specifically, let $\{v_i\}_{i=1}^{D}$ denote the fixed spectral basis obtained in Section 2. For each spectral direction, we construct two complementary reliability measures. The first is *subspace stability*. We randomly sample two subsets from the training set, compute their corresponding PCA bases $\{v_i^{(a)}\}$ and $\{v_i^{(b)}\}$, and measure the stability of the i-th direction by their alignment:

$$b_i = \mathbb{E}\left[\left|\left\langle v_i^{(a)},\, v_i^{(b)}\right\rangle\right|\right]. \tag{10}$$

A larger b_i indicates that this direction consistently appears across different subsets and is therefore more likely to correspond to a robust dominant structure of the data. The second is *augmentation consistency*. For each sample x, we construct a weakly augmented view x^w and a strongly augmented view x^s, and measure how much their coefficients differ along the i-th spectral direction. We define

$$a_i = \mathbb{E}_x\left[\exp\left(-\frac{|c_i(x^w) - c_i(x^s)|}{|c_i(x^w)| + |c_i(x^s)| + \varepsilon}\right)\right], \tag{11}$$

where $c_i(\cdot)$ is the spectral coefficient defined in Section 2. A larger a_i indicates that the information carried by this direction remains stable under perturbation.

After obtaining b_i and a_i, we normalize them to the same scale, denoted by $\tilde{b}_i$ and $\tilde{a}_i$, and combine them into a direction-wise reliability score

$$s_i = \alpha \tilde{b}_i + \beta \tilde{a}_i, \tag{12}$$

where α and β are weighting coefficients. Furthermore, motivated by the observation in Section 2 that noisy pre-training is often accompanied by a backward spread of discriminative information, we introduce a mild positional bias on top of the reliability score and define the final spectral prior as

$$p_i = \frac{\exp(s_i/T)\, \exp(-\lambda i/D)}{\sum_{j=1}^{D} \exp(s_j/T)\, \exp(-\lambda j/D)}, \tag{13}$$

where T is a temperature parameter and $\lambda \geq 0$ controls the positional decay. The resulting $p = (p_1, \ldots, p_D)$ satisfies $p_i \geq 0$ and $\sum_i p_i = 1$, and can be viewed as a direction-wise reliability prior induced by the current model itself. In the next subsection, we use this prior to regularize the discriminative spectrum of the adapted feature, so that discriminative information is preferentially retained in reliable directions and suppressed from spreading into unreliable ones.

3.4 Spectral Self-Regularization

Section 2.4 shows that noisy pre-training induces a systematic drift in the discriminative spectrum, with discriminative information often spreading toward later spectral directions. This motivates spectral regularization during adaptation, which constrains the adapted representation away from unreliable directions while preserving the original pretrained structure.

Let z denote the adapted feature produced by the adapter, and let $\hat{y} = g_\psi(z)$ be the classifier output. We first retain the standard supervised objective and use the cross-entropy loss

$$\mathcal{L}_{\text{task}} = \text{CE}(\hat{y}, y). \tag{14}$$

By itself, this term only drives boundary fitting and does not guarantee that the discriminative spectrum moves toward a more stable structure. We therefore introduce additional regularization terms defined directly in spectral space.

Given the discriminative spectrum $q(z) = (q_1, \ldots, q_D)$ constructed from the current batch of adapted features z, and the self-guided spectral prior $p = (p_1, \ldots, p_D)$ defined in Section 3.3, we first introduce a spectral alignment term

$$\mathcal{L}_{\text{spec}} = D_{\text{JS}}(q(z),\, p)\,. \tag{15}$$

This term explicitly encourages the current discriminative spectrum to move toward the internally induced reliability prior. Unlike explicit clean alignment, it does not assume access to a true clean target, but only uses the direction-wise reliability preference derived from the current model itself to suppress the overall spectral drift.

To further restrict the spread of discriminative information toward unreliable directions, we define a leakage suppression term

$$\mathcal{L}_{\text{leak}} = \sum_{i=1}^{D} (1 - p_i)\, q_i(z). \tag{16}$$

If a spectral direction is assigned a small prior weight p_i, indicating that it is considered unreliable, but the current discriminative spectrum still allocates a large amount of discriminative mass $q_i(z)$ to it, then this term imposes a stronger penalty. In this sense, $\mathcal{L}_{\text{leak}}$ directly corresponds to the phenomenon identified in Section 2, namely the spread of discriminative information toward later or unreliable directions.

At the same time, although the adapter should modify the feature structure, such modification must remain limited; otherwise, the new module may degenerate into learning a new representation space rather than performing lightweight correction on top of the pretrained feature. To this end, we further introduce a stay-close regularization. Let

$$\delta_i(x) = v_i^\top (z - h), \tag{17}$$

where $h = f(x)$ is the frozen backbone feature and v_i is the fixed spectral basis obtained in Section 2. We define

$$\mathcal{L}_{\text{stay}} = \sum_{i=1}^{D} p_i\, \delta_i(x)^2. \tag{18}$$

This term constrains the adapter-induced feature change in spectral space and imposes stronger protection on directions with larger prior weights. In other words, the method allows more flexible correction on unreliable directions while preserving the original pretrained structure on reliable ones as much as possible.

The overall objective is then written as

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{spec}} + \lambda_2 \mathcal{L}_{\text{leak}} + \lambda_3 \mathcal{L}_{\text{stay}}, \tag{19}$$

where λ_1, λ_2, and λ_3 are weighting coefficients. Here, $\mathcal{L}_{\text{task}}$ preserves downstream discriminative ability, $\mathcal{L}_{\text{spec}}$ addresses the overall spectral deviation identified in Section 2, $\mathcal{L}_{\text{leak}}$ targets the spread of discriminative information toward unreliable directions, and $\mathcal{L}_{\text{stay}}$ ensures that the correction remains lightweight and controlled. As a result, the adapter update is no longer a form of generic feature tuning, but is directly driven by the structural characteristics of discriminative spectrum drift.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate our method on two commonly used cross-domain visual datasets: Office31 and Office-Home. We consider two families of ResNet-50 backbones pretrained under different noise levels: `r50_in1k`, obtained by supervised pre-training on ImageNet-1K, and `r50_yfcc15m`, the CLIP-style visual encoder pretrained on YFCC15M. For each backbone family, we use five pre-training noise ratios: clean, 5%, 10%, 20%, and 30%. All experiments are conducted under the frozen-backbone setting. We compare the proposed **SGSA** (**S**elf-**G**uided **S**pectral **A**daptation) with four baselines: **Baseline** (linear classifier only), **Adapter-only** (adapter without spectral regularization), **NML**, and **SKD**. We use downstream classification accuracy as the primary evaluation metric.

Table 2: Average downstream transfer accuracy (%) under different pre-training noise ratios. Each entry is averaged over all transfer tasks within the corresponding dataset.

Dataset	Noise	Baseline	Adapter-only	NML	SKD	SGSA
Office31	0%	74.6	76.4	81.3	80.4	**82.8**
	5%	71.1	77.2	**83.7**	81.5	83.5
	10%	70.8	76.8	81.4	80.2	**82.6**
	20%	67.1	75.0	80.1	78.9	**81.6**
	30%	63.0	72.1	77.6	76.4	**79.3**
Office-Home	0%	33.3	35.2	**38.0**	36.9	37.9
	5%	32.6	35.4	37.5	36.8	**38.1**
	10%	31.6	34.1	35.7	35.2	**36.2**
	20%	30.0	33.2	35.1	34.5	**35.7**
	30%	28.9	31.6	33.5	32.9	**34.0**

4.2 MAIN RESULTS AND NOISE-LEVEL TRENDS

Table 2 reports the average downstream transfer accuracy under different pre-training noise ratios. Overall, SGSA achieves the best or comparable-best performance across both datasets and most noise levels. On Office31, SGSA obtains the highest accuracy in four out of five noise settings, and remains competitive at 5% noise where NML performs slightly better. On Office-Home, SGSA also achieves the best results under most noisy settings, while NML is marginally better under clean pre-training. Compared with Adapter-only, SGSA brings consistent improvements, indicating that the gains do not simply come from adding extra parameters, but from the proposed spectrum-guided regularization. The advantage of SGSA is especially clear under higher noise ratios: at 30% noise, SGSA improves over the baseline by 16.3 points on Office31 and 5.1 points on Office-Home. These results show that SGSA provides a more effective and stable adaptation strategy for mitigating the inherited effect of noisy pre-training.

4.3 DOES SGSA MITIGATE DISCRIMINATIVE SPECTRUM DRIFT?

To verify that the improvement of SGSA is not merely due to the additional adapter parameters, we compare **Adapter-only** and **SGSA**. As shown in Table 3, SGSA consistently produces a smaller divergence from the clean discriminative spectrum on both Office31 and Office-Home, indicating that it more effectively suppresses the overall spectral drift induced by noisy pre-training. We

Table 3: Comparison between Adapter-only and SGSA on spectral drift metrics for Office31 and Office-Home with two pretrained backbone families. JS denotes the Jensen–Shannon divergence to the clean discriminative spectrum; Centroid denotes the average spectral position of discriminative information. Lower JS indicates less deviation from the clean reference, while a lower centroid indicates greater concentration in leading spectral directions. A-only denotes Adapter-only.

Dataset	Backbone	JS↓		Centroid↓	
		A-only	SGSA	A-only	SGSA
Office31	`r50_in1k`	0.0381	**0.0264**	67.1	**61.8**
	`r50_yfcc15m`	0.0302	**0.0249**	321.0	**318.9**
Office-Home	`r50_in1k`	0.0228	**0.0156**	330.4	**316.5**
	`r50_yfcc15m`	0.0138	**0.0124**	820.6	**816.7**

further report the spectral centroid as a structural indicator of where discriminative information is concentrated. Under the representative `r50_in1k` setting, SGSA yields lower centroid values than Adapter-only, suggesting a weaker backward spread of discriminative information toward later directions. These results show that SGSA improves downstream adaptation not only by increasing accuracy, but also by directly correcting the discriminative spectrum drift identified in Section 2.

Table 4: Ablation study of SGSA. Classification accuracy (%) is averaged over two backbone families and all pre-training noise levels. Avg. denotes the mean accuracy across Office31 and Office-Home. Higher is better.

Method	Office31 Acc.↑	Office-Home Acc.↑	Avg.↑
Baseline	69.3	31.3	50.3
Adapter-only	75.5	33.9	54.7
+ $\mathcal{L}_{\text{spec}}$	78.6	34.9	56.8
+ $\mathcal{L}_{\text{spec}} + \mathcal{L}_{\text{leak}}$	80.2	35.7	58.0
SGSA (full)	**82.0**	**36.4**	**59.2**

Ablation study. As shown in Table 4, each component of SGSA contributes to the final performance. Adapter-only improves over the baseline, indicating that lightweight feature adjustment is useful under the frozen-backbone setting. Adding $\mathcal{L}_{\text{spec}}$ brings a clear further gain, showing the importance of directly regularizing the discriminative spectrum. The additional $\mathcal{L}_{\text{leak}}$ term further improves performance by suppressing unreliable spectral directions. The full SGSA objective achieves the best accuracy on both Office31 and Office-Home, improving the average accuracy from 50.3% to 59.2%. This verifies that SGSA benefits from its spectrum-guided design rather than simply from extra adapter parameters.

5 Conclusion

In this paper, we studied catastrophic inheritance under noisy pre-training from the perspective of downstream discriminative structure. We show that noisy pre-training induces a structural drift in the discriminative spectrum, rather than only causing surface-level performance variation. Motivated by this finding, we proposed **SGSA**, a lightweight adaptation framework for the practical no-clean-reference, frozen-backbone setting. Experiments on cross-domain visual datasets show that SGSA improves downstream transfer and mitigates discriminative spectrum drift. We hope this work offers a useful step toward understanding and correcting inherited representation distortion in noisy pre-trained models.

REFERENCES

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv e-prints*, pp. arXiv–2108, 2021.

Hao Chen, Jindong Wang, Ankit Shah, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Understanding and mitigating the label noise in pre-training on downstream tasks. *arXiv preprint arXiv:2309.17002*, 2023.

Hao Chen, Bhiksha Raj, Xing Xie, and Jindong Wang. On catastrophic inheritance of large foundation models. *arXiv preprint arXiv:2402.01909*, 2024.

Hao Chen, Zihan Wang, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, Bhiksha Raj, and Jindong Wang. Impact of noisy supervision in foundation model learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Rahim Entezari, Mitchell Wortsman, Olga Saukh, Moein Shariatnia, Hanie Sedghi, and Ludwig Schmidt. The role of pre-training data in transfer learning. *CoRR*, abs/2302.13602, 2023. doi: 10.48550/ARXIV.2302.13602. URL https://doi.org/10.48550/arXiv.2302.13602.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2661–2671. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00277. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of CLIP. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/86a8a512b27f49519594ebe89f66d708-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 5389–5400. PMLR, 2019. URL http://proceedings.mlr.press/v97/recht19a.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Vijay Vasudevan, Benjamin Caine, Raphael Gontijo Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/2cd5737c59645f7ef23b2842b705edf2-Abstract-Conference.html.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.