Improving Invariant Risk Minimization with Self-Supervised Learning for OOD Robustness

Laxmiraman Gudewar
Saarland University
lagu00003@stud.uni-saarland.de

Soham Tolwala
Saarland University
soto00002@stud.uni-saarland.de

Abstract

Empirical Risk Minimization (ERM) has long been the default training paradigm for deep neural networks. However, ERM tends to exploit spurious correlations in the training distribution, resulting in sharp failures under distribution shift. Invariant Risk Minimization (IRM) was introduced to mitigate this issue by encouraging predictors that remain optimal across different training environments. While elegant in theory, IRM as commonly implemented (through the IRMv1 surrogate) is difficult to train in practice, often unstable and highly sensitive to hyperparameters.

In this work, we study a simple but powerful recipe: combining self-supervised pretraining *(DINOv2, ViT-S/14) with IRM applied only to the linear head. We document the stability choices that enable IRM to work in this setup, such as penalty scheduling and training setup. Further, we present an extensive experimental study on the Waterbirds dataset, analyzing performance gains of SSL+IRM compared to ERM and vanilla IRM, supported by both quantitative results and qualitative Grad-CAM analysis.*

Our key contributions are:

1. *A two-stage training pipeline combining DINOv2 self-supervised features with IRM applied only to the linear classifier head.*

2. *Study of stability choices critical for IRM optimization, including penalty scheduling, gradient clipping, and optimizer strategies.*

3. *Experimental validation on Waterbirds, showing significant improvements in worst-group accuracy and stability compared to ERM baseline.*

4. *Qualitative Grad-CAM analysis demonstrating stronger causal localization of object features when using SSL+IRM.*

1. Introduction

Modern deep learning models trained under Empirical Risk Minimization (ERM) have achieved remarkable performance across benchmarks in vision, language, and multimodal domains. Yet, this method has a crucial weakness: models trained with ERM tend to exploit superficial correlations that are predictive in the training data, but which do not generalize under distribution shift. This fragility becomes visible in benchmarks like the Waterbirds dataset [3], where the background strongly correlates with the label (waterbirds on water, landbirds on land). ERM-trained models often overfit to the background, failing catastrophically on rare cases such as a waterbird standing on land or landbird on water. By penalizing the variability of optimal classifiers across environments, IRM aims to recover predictors that focus on causal features rather than spurious correlations. Despite this theoretical appeal, IRM in practice (via the surrogate IRMv1 objective) has proven extremely difficult to optimize: the penalty term is unstable, sensitive to hyperparameters, and can overpower learning when applied to large models.

In parallel, the field has witnessed rapid progress in self-supervised learning (SSL). Models like DINOv2 [5] train large vision transformers without labels, learning general-purpose features that transfer remarkably well across diverse downstream tasks. Importantly, these features appear to already capture object-centric information, potentially reducing the reliance on spurious background correlations.

Our work explores the synergy between SSL and IRM. We propose a simple method: freeze a pretrained DINOv2 backbone and apply IRM training only to a lightweight linear head. By decoupling feature learning from invariance enforcement, we reduce the optimization burden and make IRM tractable. Our empirical study shows that this combination not only improves worst-group accuracy but also shifts model attention toward causal object regions, as visualized with Grad-CAM.

2. Related Work

Invariant Risk Minimization (IRM) [1] was introduced as a principled framework to encourage predictors that remain optimal across environments by enforcing gradient invariance through the IRMv1 surrogate (penalized version). While theoretically appealing, this objective has proven difficult to optimize in practice, with results highly sensitive to the choice of penalty weight, optimizer, and scheduling strategies [2].

An alternative line of research focuses on distributionally robust optimization (DRO). Methods such as Group DRO [3] explicitly minimize worst-case subgroup risk, often delivering strong robustness guarantees. However, these methods require group annotations at training time, which limits their applicability in broader settings. Relaxations like VREx and Fishr instead penalize variance in risks or gradients across environments, providing greater stability and scalability.

Meanwhile, self-supervised learning (SSL) has rapidly advanced as a means to acquire strong visual representations without labels. Approaches such as MoCo, SimCLR, and more recently DINOv2 [5] produce features that are transferable and less reliant on dataset-specific shortcuts, making them attractive for robustness. Complementary to this, post-hoc adaptation methods [4] demonstrate that even ERM-trained representations can be retrofitted with more robust classifiers, underscoring the central role of feature quality in subgroup generalization.

Our work sits at the intersection: we leverage SSL features and demonstrate that IRM becomes tractable and effective in this setting.

3. IRM Objective and Explanation

The IRM objective balances predictive risk and invariance:

$$L_{\text{IRM}}(\Phi, w) = \sum_{e \in \mathcal{E}_{\text{tr}}} \Big[R^e(w \circ \Phi) + \lambda\, \mathbb{D}(w, \Phi, e) \Big], \qquad (1)$$

where Φ is the feature extractor, w the classifier, R^e the risk in environment e, and $\mathbb{D}$ is the invariance penalty.

The IRM objective combines two components: the standard classification loss $R^e(w \circ \Phi)$, which measures predictive error in each environment e, and the invariance penalty $\mathbb{D}(w, \Phi, e)$, which evaluates how sensitive the risk is to perturbations of the classifier. If the same classifier w is truly optimal across all environments, this penalty becomes zero. The balance between these two terms is controlled by the hyperparameter λ, which determines the strength of the invariance constraint. A small λ effectively reduces the method to ERM, risking reliance on spurious features, while an excessively large λ can overwhelm optimization and cause underfitting. In practice, moderate values with gradual warmup (e.g., from 0 to 3–5) provide a stable trade-off between predictive accuracy and invariance. Thus, λ plays a central role in guiding the model toward representations that are both accurate and robust across environments.

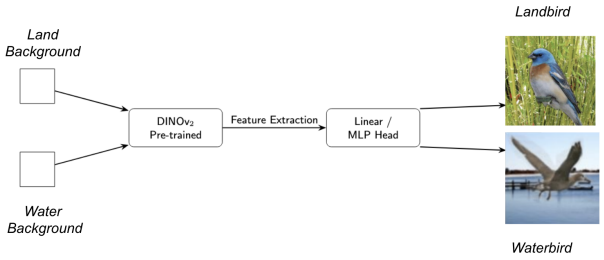

Figure 1. SSL+IRM pipeline. Stage 1: DINOv2 extracts frozen features. Stage 2: IRM trains a linear or MLP head with invariance penalty.

4. Method

4.1. Pipeline Overview

Our approach follows a two-stage pipeline designed to leverage the strengths of self-supervised pretraining while keeping IRM optimization tractable.

Stage 1: Self-supervised pretraining. We adopt the model DINOv2 ViT-S/14 [5] as a frozen feature extractor. The publicly released pretrained checkpoint is used without further finetuning. The model outputs 384-dimensional feature embeddings (CLS token), which we optionally normalize using LayerNorm to improve stability.

Stage 2: IRM head training. On top of frozen features, we train a lightweight classifier head-a linear layer using the IRM objective (Eq. 1). Training is performed across Waterbirds environments (land vs. water), with invariance enforced only at the head level. This design reduces computational complexity and avoids destabilizing the pretrained backbone.

4.2. Training Setup

We conduct experiments on the Waterbirds dataset [3], which is specifically designed to evaluate robustness to spurious correlations between bird type and background. We report two metrics: *average accuracy* (AvgAcc) across all groups and *worst-group accuracy* (WGAcc) to measure robustness under distribution shift.

The IRM head is trained using AdamW with learning rate 10^{-3} and weight decay 10^{-4}. Batch size is set to 32 and models are trained for 70 to 250 epochs depending on schedule. We use cosine learning rate decay where noted.

4.3. Stability-Oriented Design Choices

Since IRM is notoriously unstable, therefore we have used several stabilizing strategies: gradually warming up the invariance penalty λ to prevent collapse, computing penalties per environment to enforce local invariance, and using AdamW with cosine scheduling (or SGD in some runs) for smoother optimization. We further apply gradient clipping (norm $\leq$ 1) to avoid exploding updates and rely on balanced-validation model selection to ensure improvements reflect subgroup robustness rather than overfitting.

5. Results

5.1. Quantitative Results

We first compare IRM training on top of DINOv1 features versus DINOv2 self-supervised features. Table 1 shows that when using DINOv1, IRM training is highly unstable, with average accuracies around 50% and worst-group accuracies dropping below 40%. This highlights the difficulty of optimizing IRM directly from scratch, as the invariance penalty often overwhelms learning.

In contrast, Table 2 reports results with frozen DINOv2 features and a linear head. Here, IRM becomes much more tractable: all settings yield strong improvements, with worst-group accuracy exceeding 66% even under short training schedules. The best configuration (Exp 5, fixed $\lambda = 3$ after warmup) achieves an average accuracy of 89.2% and a worst-group accuracy of 72.6%, a significant improvement over ERM baselines on Waterbirds dataset.

5.2. Effect of Penalty Scheduling

Across experiments, the scheduling of the invariance penalty λ played a critical role. Directly applying a large penalty from the start (IRMv1-style) caused optimization collapse. Instead, gradual warmup or fixing λ to a small value after burn-in allowed stable convergence. This aligns with theory: early in training the model must learn predictive features, while invariance penalties are only beneficial once features are discriminative.

Interestingly, the long ramp schedule (Exp 4) achieved solid results but required very long training. The fixed-λ strategy (Exp 5) was more efficient, converging in fewer epochs while achieving the best subgroup robustness. This suggests that penalty scheduling not only stabilizes optimization but also directly controls the trade-off between predictive accuracy and invariance.

5.3. Grad-CAM Analysis

To gain qualitative insights into model behavior, we compute Grad-CAM on test images. The visualizations in Figure 2 highlight clear differences between ERM, and SSL+IRM. ERM tends to focus on background textures such as trees or water, confirming that it exploits spurious correlations. IRM shifts attention partially toward the bird but remains noisy and unstable, reflecting its difficulty in optimization. In contrast, SSL+IRM produces sharp, concentrated activations on the bird's body and head, indicating that the model relies on causal object features rather than contextual cues.

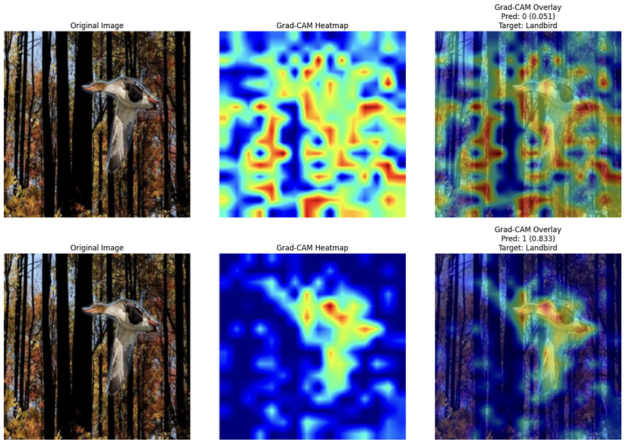

Figure 2. Grad-CAM visualizations for input image (left), ERM (middle), and SSL+IRM (right).

These results complement our quantitative findings: the consistent causal localization observed with SSL+IRM helps explain its higher worst-group accuracy. ERM overfits to spurious backgrounds, while SSL+IRM focuses reliably on causal bird regions (body and head). The method not only improves numerical performance but also aligns attention maps with semantically meaningful regions, providing stronger evidence of robustness and interpretability compared to ERM and vanilla risk minimization methods.

6. Discussion

6.1. Ablation Studies

We further analyze the effect of the invariance penalty λ scheduling strategy. Experiments show that: Our ablation studies on the penalty parameter λ highlight three findings. Applying a large λ from the start destabilizes training and often causes collapse. Gradually ramping up λ improves stability, as seen in Exp 4, though it requires long training. In contrast, fixing λ at a moderate value (e.g., 3) after a short warmup (Exp 5) achieved the best balance of accuracy, robustness, and efficiency. This supports the IRMv1 intuition that invariance penalties are only beneficial once discriminative features have been learned, making careful scheduling essential for stable IRM training.
These results confirm that stability choices are critical to making IRM practical in conjunction with SSL features.

Results with DINOv1 backbone

Exp	IRM λ Strategy	Epochs	Opt.	Sched.	Avg Acc	Worst Acc
Exp 1	Zero (1–5), Const. (6–50)	50	AdamW	None	50.71	37.98
Exp 2	Zero (1), Var. (2–6), Max (7–50)	50	AdamW	None	50.40	29.32

Table 1. IRM experiments with DINOv1 backbone. Experiments are unstable with low accuracies, showing the difficulty of optimizing IRM from scratch.

Results with DINOv2 features + MLP head.

Exp	IRM λ Strategy	Epochs	Opt.	Sched.	Avg Acc	Worst Acc
Exp 3	Ramp-up to $\lambda = 10$ (10→50)	50	SGD	None	85.69	66.92
Exp 4	Long ramp to $\lambda = 5$ (50→200)	250	AdamW	Cosine (T=250)	86.50	69.17
Exp 5	Fixed $\lambda = 3$ after warmup (5)	250	AdamW	Cosine (T=60)	**89.23**	**72.63**

Table 2. Summary of IRM experiments with DINOv2 + Linear Head. Pretraining stabilizes IRM and yields strong OOD robustness.

IRM's penalty drives gradients of the per-environment risk (w.r.t. the classifier) toward zero at the same classifier, thereby approximating the constrained optimum where w is jointly optimal across all environments [1]. This formulation reflects the principle of causal invariance: if a predictor remains optimal simultaneously in multiple environments, it is more likely to capture underlying causal features rather than spurious correlations. Our results support that, when strong pretrained features (such as DINOv2 representations) are provided, this surrogate penalty becomes both *optimizable* and *useful*.

6.2. Limitations.

Despite the gains, several challenges remain. First, our experiments are limited to the Waterbirds dataset; it is unclear if the improvements transfer seamlessly to larger, more complex benchmarks such as DomainNet or ImageNet-R. Second, although self-supervised pretraining stabilizes IRM training, the method is still sensitive to the choice of the invariance penalty λ, and reliable validation heuristics are necessary. Our pipeline freezes pretrained DINOv2 features, which respects the two-stage design but may underutilize the backbone capacity.

6.3. Future Work

- **Scaling to richer datasets:** Extend experiments from Waterbirds to larger, multi-domain benchmarks such as DomainNet, ImageNet-R, or WILDS to evaluate whether SSL+IRM maintains robustness when spurious correlations are more complex and less obvious.
- **Adaptive invariance penalties:** Investigate new formulations or dynamic scheduling of the IRM penalty λ (e.g., gradient-based adaptation, learning strategies) to reduce sensitivity to hyperparameters and improve stability across training regimes.
- **Refining backbone integration:** Move beyond frozen features by selectively unfreezing late-stage transformer blocks in DINOv2, or exploring causality-inspired SSL objectives, to better align pretrained invariances with downstream subgroup robustness.

7. Conclusion

In this work, we explored the integration of IRM with self-supervised pretraining using DINOv2 to improve OOD shift robustness. Our experiments on the Waterbirds dataset demonstrate that combining DINOv2 features with IRM training leads to substantial improvements in worst-group accuracy compared to both ERM and standard IRM, while also stabilizing the optimization process. The results further show that self-supervised pretraining reduces sensitivity to hyperparameter choices, allowing IRM to focus more reliably on causal features rather than spurious correlations.

Beyond accuracy, qualitative analysis with Grad-CAM reveals that SSL+IRM consistently shifts attention from irrelevant background regions to object-specific regions, providing stronger evidence of causal learning. Ablation studies confirm that the benefits arise from the synergy between SSL and IRM: pretraining supplies invariant representations, and IRM refines them to enforce subgroup robustness. While this pipeline reduces dependence on large labeled datasets and extensive hyperparameter tuning, it remains limited to a single benchmark dataset, motivating future work on more generalizable SSL objectives, larger-scale evaluations, and multimodal extensions.

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2, 4

[2] Y. J. Choe, J. Ham, and K. Park. An empirical study of invariant risk minimization. in *Proc. ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. 2

[3] 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. 1, 2

[4] Y. Lu, X. Jin, M. Zhang, L. Li, X. Zhang, C. Tao, X. Ren, and H. Zhao. Understanding post-hoc adaptation for improving subgroup robustness. in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[5] H. Touvron, M. Cord, H. Jégou, F. Massa, J. Kahn, P. Fernandez, J. Beaumont, R. Texier, G. Synnaeve, and A. Joulin. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2