Self Supervised Latent Representations for Robust Generalization in Unconstrained Environments

Yann Ahlgrim

I. Context & Problem Statement

Despite some deep learning models achieving impressive results on benchmark datasets, they might be struggling to generalize to real-world data [26] due to distribution shifts where a model trained on one domain is exposed to a different domain or a subpopulation during testing [20]. Such shifts reveal a significant generalization gap, particularly on rare groups [28].

Another form of a shift is the Covariate shift which occurs when the distribution of the covariates (input features) changes between the training and testing phases, while the conditional density of the labels given the covariates remains the same [29]. This discrepancy often results in a precipitous performance drop in practical applications [18].

Fundamentally, this lack of robustness stems from the model's tendency to exploit spurious correlations—shortcuts within the training data that do not hold in the test distribution [14]. Achieving robust generalization therefore requires models to extract invariant features [2] that are causally relevant to the task rather than relying on domain-specific noise.

Recent advancements in self-supervised learning (SSL) offer potential solutions to these challenges. Current paradigms include contrastive learning [9], invariance-based methods like Barlow Twins and VICReg [5, 33], and generative approaches such as Masked Autoencoders (MAE) [16]. A particularly promising development is the Image-based Joint-Embedding Predictive Architecture (I-JEPA) [4]. I-JEPA learns by predicting the latent representations of target image blocks from a context block [4].

The I-JEPA outperforms the MAE and invariance-based architectures on ImageNet [4], however it is still an open question whether the predictive approach of I-JEPA can learn more robust representations that generalize better to unconstrained environments than the reconstructive approach of MAE or the invariance-based approaches.

II. Research Question & Objectives

The primary objective of this research is to evaluate the efficacy of predictive self-supervised architectures in learning robust representations for unconstrained environments. This goal is distilled into the following central research question and supporting sub-questions:

Main Research Question

To what extent does the predictive latent objective of the I-JEPA facilitate the learning of robust, invariant representations that generalize to unconstrained environments more effectively than generative or invariance-based SSL methods?

Sub-questions

1) *Spurious Correlations:* How effectively does I-JEPA's predictive approach in latent space decouple semantic features (e.g., animal) from nuisance variables (e.g., habitat-specific backgrounds) compared to pixel-level reconstruction in MAE and invariance-based methods?
2) *Distribution Shift:* How stable are the learned representations when subjected to distribution shifts and does this stability correlate with improved downstream performance?
3) *Label Efficiency:* To what degree does the predictive learning of image structures reduce the reliance on manually annotated data for the classification of rare subpopulations compared to traditional supervised deep learning methods?

III. Methodology and Approach

This research employs an empirical benchmarking framework to evaluate the efficacy of predictive SSL architectures in learning robust representations for unconstrained environments. The methodology is structured as follows:

1) *Theoretical Foundation:*
   - *SSL:* analysis of current SSL approaches, specifically comparing generative approaches like MAE [8, 16, 19, 21, 24, 27], invariance-based methods [5–7, 9, 10, 15, 17, 33], and predictive architectures like I-JEPA [4].
   - *Vision Transformer (ViT):* understand the architecture of Vision Transformers [11–13, 31] as the backbone of modern SSL models.
   - *Distribution Shifts:* review literature on distribution shifts, particularly covariate shifts [29] and their impact on model performance [18, 26].
   - *Quantitative Analysis:* identify appropriate metrics for assessing robust generalization under distribution shifts, such as in-distribution and out-of-distribution accuracy [22]. Research how to conduct a comparative experiment with Linear Probing [1] and changing the amount of labeled data for the downstream task [3, 9].
   - *Qualitative Analysis:* learn algorithms, like t-SNE [30] or UMAP [25] to reduce the dimensions of the latent representation vectors to visualize them in 2D.
2) *Practical Implementation:*
   a) *Pre-Training:* training an I-JEPA, MAE, and an invariance-based model on unconstrained datasets, like iWildCam [20, 32] or different other available datasets [23].
   b) *Linear Probing:* adding linear classifier on top of the frozen backbone [4, 9] which will be trained on a small subset of labeled data.
3) *Evaluation:*
   a) *Quantitative Performance Analysis:* Compare the in-distribution and out-of-distribution accuracy [22] as well as the label efficiency of the models [3, 9].
   b) *Qualitative Performance Analysis:* Compare randomly picked images and visualize the semantic level by applying dimension reduction algorithms [25, 30].
4) *Discussion:* Interpretation of the empirical results to determine if predictive latent objectives facilitate more effective generalization in real-world, unconstrained environments than the other SSL methods.

IV. Outline

1) *Introduction*
   - *Motivation and Problem Statement:* The gap between benchmark success and real-world distribution shifts.
   - *The Role of SSL:* Using self-supervision to extract invariant features.
   - *Contributions:* Evaluating the robustness of predictive latent objectives (I-JEPA) against generative and invariance-based methods.
2) *Related Work*
   - *SSL methods:* Comparison of MAE, VICReg/Barlow Twins, and I-JEPA.
   - *ViT:* Architecture and scaling properties.
   - *Robustness:* Overview of covariate shifts and shortcut learning.
3) *Methodology*
   - *Architecture:* Formalizing the three objective functions.
   - *Experimental Setup:* Pre-training on unconstrained datasets (e.g., iWildCam).
   - *Evaluation:* Linear probing and label efficiency testing.
4) *Experimental Results*
   - *Quantitative Analysis:* In-distribution vs. Out-of-distribution accuracy.
   - *Qualitative Analysis:* Latent space visualization via t-SNE/UMAP.
5) *Discussion*
   - *Predictive vs. Reconstructive:* Interpreting the decoupling of semantic features from noise.
   - *Practical Implications:* Recommendations for robust model selection.

6) *Conclusion:* Summary of findings and future research directions.

REFERENCES

[1] Alain, Guillaume and Bengio, Yoshua. *Understanding intermediate layers using linear classifier probes*. arXiv:1610.01644 [stat]. Nov. 2018. DOI: 10.48550/arXiv.1610.01644. URL: http://arxiv.org/abs/1610.01644 (visited on 02/23/2026).

[2] Arjovsky, Martin et al. *Invariant Risk Minimization*. arXiv:1907.02893 [stat]. Mar. 2020. DOI: 10.48550/arXiv.1907.02893. URL: http://arxiv.org/abs/1907.02893 (visited on 02/27/2026).

[3] Assran, Mahmoud et al. *Masked Siamese Networks for Label-Efficient Learning*. arXiv:2204.07141 [cs]. Apr. 2022. DOI: 10.48550/arXiv.2204.07141. URL: http://arxiv.org/abs/2204.07141 (visited on 03/03/2026).

[4] Assran, Mahmoud et al. "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2023, pp. 15619–15629. DOI: 10.1109/CVPR52729.2023.01499. URL: https://ieeexplore.ieee.org/document/10205476/ (visited on 02/13/2026).

[5] Bardes, Adrien, Ponce, Jean, and LeCun, Yann. *VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning*. arXiv:2105.04906 [cs]. Jan. 2022. DOI: 10.48550/arXiv.2105.04906. URL: http://arxiv.org/abs/2105.04906 (visited on 02/24/2026).

[6] Caron, Mathilde et al. "Emerging Properties in Self-Supervised Vision Transformers". en. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9630–9640. ISBN: 978-1-6654-2812-5. DOI: 10.1109/ICCV48922.2021.00951. URL: https://ieeexplore.ieee.org/document/9709990/ (visited on 02/24/2026).

[7] Caron, Mathilde et al. *Unsupervised Learning of Visual Features by Contrasting Cluster Assignments*. arXiv:2006.09882 [cs]. Jan. 2021. DOI: 10.48550/arXiv.2006.09882. URL: http://arxiv.org/abs/2006.09882 (visited on 02/24/2026).

[8] Chang, Huiwen et al. "MaskGIT: Masked Generative Image Transformer". en. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 11305–11315. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.01103. URL: https://ieeexplore.ieee.org/document/9878676/ (visited on 02/24/2026).

[9] Chen, Ting et al. "A Simple Framework for Contrastive Learning of Visual Representations". en. In: *Proceedings of the 37th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, Nov. 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html (visited on 02/23/2026).

[10] Chen, Xinlei and He, Kaiming. "Exploring Simple Siamese Representation Learning". en. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, June 2021, pp. 15745–15753. ISBN: 978-1-6654-4509-2. DOI: 10.1109/CVPR46437.2021.01549. URL: https://ieeexplore.ieee.org/document/9578004/ (visited on 02/24/2026).

[11] Chen, Xinlei, Xie, Saining, and He, Kaiming. "An Empirical Study of Training Self-Supervised Vision Transformers". en. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9620–9629. ISBN: 978-1-6654-2812-5. DOI: 10.1109/ICCV48922.2021.00950. URL: https://ieeexplore.ieee.org/document/9711302/ (visited on 02/24/2026).

[12] Dehghani, Mostafa et al. "Scaling Vision Transformers to 22 Billion Parameters". en. In: *Proceedings of the 40th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2023, pp. 7480–7512. URL: https://proceedings.mlr.press/v202/dehghani23a.html (visited on 02/24/2026).

[13] Dosovitskiy, Alexey et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv:2010.11929 [cs]. June 2021. DOI: 10.48550/arXiv.2010.11929. URL: http://arxiv.org/abs/2010.11929 (visited on 02/13/2026).

[14] Geirhos, Robert et al. "Shortcut Learning in Deep Neural Networks". In: *Nature Machine Intelligence* 2.11 (Nov. 2020). arXiv:2004.07780 [cs], pp. 665–673. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00257-z. URL: http://arxiv.org/abs/2004.07780 (visited on 02/27/2026).

[15] Grill, Jean-Bastien et al. *Bootstrap your own latent: A new approach to self-supervised Learning*. arXiv:2006.07733 [cs]. Sept. 2020. DOI: 10.48550/arXiv.2006.07733. URL: http://arxiv.org/abs/2006.07733 (visited on 02/24/2026).

[16] He, Kaiming et al. "Masked Autoencoders Are Scalable Vision Learners". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2022, pp. 15979–15988. DOI: 10.1109/CVPR52688.2022.01553. URL: https://ieeexplore.ieee.org/document/9879206/ (visited on 02/13/2026).

[17] He, Kaiming et al. "Momentum Contrast for Unsupervised Visual Representation Learning". en. In: ().

[18] Hendrycks, Dan and Dietterich, Thomas. *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*. arXiv:1903.12261 [cs]. Mar. 2019. DOI: 10.48550/arXiv.1903.12261. URL: http://arxiv.org/abs/1903.12261 (visited on 02/23/2026).

[19] Jiang, Zihao. "Masked Autoencoders for Vision Representation Learning on CIFAR-10". In: *2025 IEEE 5th International Conference on Data Science and Computer Application (ICDSCA)*. Oct. 2025, pp. 395–400. DOI: 10.1109/ICDSCA67083.2025.11350287. URL: https://ieeexplore.ieee.org/document/11350287/ (visited on 02/24/2026).

[20] Koh, Pang Wei et al. "WILDS: A Benchmark of in-the-Wild Distribution Shifts". en. In: *Proceedings of the 38th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2021, pp. 5637–5664. URL: https://proceedings.mlr.press/v139/koh21a.html (visited on 02/27/2026).

[21] Kraus, Oren et al. "Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2024, pp. 11757–11768. DOI: 10.1109/CVPR52733.2024.01117. URL: https://ieeexplore.ieee.org/document/10657479/ (visited on 02/24/2026).

[22] Kumar, Ananya et al. *Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution*. arXiv:2202.10054 [cs]. Feb. 2022. DOI: 10.48550/arXiv.2202.10054. URL: http://arxiv.org/abs/2202.10054 (visited on 02/23/2026).

[23] Kyathanahally, S. P. et al. "Ensembles of data-efficient vision transformers as a new paradigm for automated classification in ecology". en. In: *Scientific Reports* 12.1 (Nov. 2022), p. 18590. ISSN: 2045-2322. DOI: 10.1038/s41598-022-21910-0. URL: https://www.nature.com/articles/s41598-022-21910-0 (visited on 02/24/2026).

[24] Marques, Fernando G. et al. "Applying ViT Masked Autoencoders to Seismic Data for Feature Extraction and Few-Shot Learning". In: *IEEE Geoscience and Remote Sensing Letters* 23 (2026), pp. 1–5. ISSN: 1558-0571. DOI: 10.1109/LGRS.2025.3639172. URL: https://ieeexplore.ieee.org/document/11271686/ (visited on 02/23/2026).

[25] McInnes, Leland, Healy, John, and Melville, James. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426 [stat]. Sept. 2020. DOI: 10.48550/arXiv.1802.03426. URL: http://arxiv.org/abs/1802.03426 (visited on 02/23/2026).

[26] Recht, Benjamin et al. *Do ImageNet Classifiers Generalize to ImageNet?* arXiv:1902.10811 [cs]. June 2019. DOI: 10.48550/arXiv.1902.10811. URL: http://arxiv.org/abs/1902.10811 (visited on 02/23/2026).

[27] RoßTeutscher, Immanuel, Drese, Klaus S., and Uphues, Thorsten. "Masked Autoencoders for Ultrasound Signals: Robust Representation Learning for Downstream Ap-

plications". In: *IEEE Access* 13 (2025), pp. 212620–212634. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3644232. URL: https://ieeexplore.ieee.org/document/11299642/ (visited on 02/24/2026).

[28] Sagawa, Shiori et al. *Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization*. arXiv:1911.08731 [cs]. Apr. 2020. DOI: 10.48550/arXiv.1911.08731. URL: http://arxiv.org/abs/1911.08731 (visited on 02/27/2026).

[29] Shimodaira, Hidetoshi. "Improving predictive inference under covariate shift by weighting the log-likelihood function". en. In: *Journal of Statistical Planning and Inference* 90.2 (Oct. 2000), pp. 227–244. ISSN: 03783758. DOI: 10.1016/S0378-3758(00)00115-4. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378375800001154 (visited on 02/23/2026).

[30] Van der Maaten, Laurens and Hinton, Geoffrey. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[31] Vaswani, Ashish et al. "Attention is All you Need". en. In: ().

[32] Wang, Lifeng et al. "DeLoCo: Decoupled location context-guided framework for wildlife species classification using camera trap images". en. In: *Ecological Informatics* 85 (Mar. 2025), p. 102949. ISSN: 15749541. DOI: 10.1016/j.ecoinf.2024.102949. URL: https://linkinghub.elsevier.com/retrieve/pii/S1574954124004916 (visited on 11/30/2025).

[33] Zbontar, Jure et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". en. In: *Proceedings of the 38th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2021, pp. 12310–12320. URL: https://proceedings.mlr.press/v139/zbontar21a.html (visited on 02/24/2026).