Aligning Feature Distributions in VICReg Using Maximum Mean Discrepancy for Enhanced Manifold Awareness in Self-Supervised Representation Learning

M.Hadi Sepanj[1], Paul Fieguth[1]
[1]Vision and Image Processing Group, Systems Design Engineering, University of Waterloo
{mhsepanj, paul.fieguth}@uwaterloo.ca

Abstract

Self-supervised learning (SSL) methods like VICReg have shown considerable success in generating robust data representation by promoting invariance across augmented views. However, VICReg's focus on pairwise alignment between augmentations limits its capacity to ensure broader consistency across entire batches of diverse transformations. In this paper, we enhance VICReg by integrating a Maximum Mean Discrepancy (MMD) term, which aligns feature distributions across the entire batch in a Reproducing Kernel Hilbert Space (RKHS), thereby promoting batch-level invariance. By enforcing a unified feature distribution across a batch, MMD enables the model to capture higher-order dependencies and reduce variability among augmented views. We have evaluated our approach on MNIST, CIFAR-10, and STL-10, where the results demonstrate improved representation quality, as evidenced by clustering accuracy and linear classification performance. The results highlight the effectiveness of incorporating MMD term into VICReg in enhancing the representation quality.

1 Introduction

Self-supervised learning (SSL) [1, 2, 3] has emerged as a powerful framework for learning data representations without labeled supervision. VICReg [4, 5, 6] has advanced SSL by minimizing three key objectives: invariance, variance, and covariance regularization. Specifically, VICReg's invariance objective aims to produce consistent representations for different augmentations of the same input. This ensures that representations are stable to changes in data views [4]. However, while pairwise invariance promotes consistency between augmented pairs [7], capturing broader batch-level consistency remains a challenge, especially for complex datasets where robust invariance is essential. Batch-level consistency is crucial because it ensures the model generalizes effectively across a wide range of augmentations, producing a more robust feature space [8].

To address this limitation, we propose enhancing VICReg by incorporating a Maximum Mean Discrepancy (MMD) term [9]. MMD is a kernel-based distribution alignment technique that enforces consistency across the entire batch of augmented features by aligning distributions in a Reproducing Kernel Hilbert Space (RKHS) [9, 10, 11]. Unlike pairwise objectives, MMD captures both linear and non-linear dependencies across the batch, promoting smoother and more stable feature distributions. Our experiments on MNIST, CIFAR-10, and STL-10 demonstrate that integrating MMD into VICReg improves representation quality, as shown by clustering accuracy and linear classification performance.

2 Background and Related Work

2.1 Invariance in Self-Supervised Learning and VICReg

Self-supervised learning frameworks like SimCLR [12], BYOL [13], and VICReg [4] have achieved significant progress by training models to produce consistent representations across augmentations of the same input. VICReg minimizes three objectives:

$$\mathcal{L}_{\text{VICReg}} = \lambda\,\mathcal{L}_{\text{inv}} + \mu\,\mathcal{L}_{\text{var}} + \nu\,\mathcal{L}_{\text{cov}}, \tag{1}$$

where λ, μ, and ν are hyperparameters. The model uses a backbone-projector architecture [12, 14], a configuration frequently used in SSL frameworks [4, 14], to transform input samples into feature vectors through an encoder, followed by an expander network. The goal of invariance (to data augmentation) in VICReg is to

produce similar representations for different augmentations. However, this pairwise approach (comparison of representations between pairs of augmented views of the same sample) may not fully address variability across an entire batch of augmentations, limiting the model's robustness and generalization.

2.2 Maximum Mean Discrepancy (MMD) for Enhanced Invariance

Maximum Mean Discrepancy (MMD) [9] measures the distance between two probability distributions in an RKHS. It is defined as

$$\text{MMD}(P, Q) = \|\mathbb{E}_P[\phi(z_1)] - \mathbb{E}_Q[\phi(z_2)]\|^2_{\mathcal{H}}, \quad (2)$$

where P and Q are distributions over representations z_1 and z_2, and ϕ is a mapping to RKHS using a positive-definite kernel function. By aligning distributions across the batch, MMD captures both linear and non-linear dependencies [9], promoting a more robust and stable feature space.

2.3 Contributions of MMD to VICReg for Manifold-Aware Invariance

Integrating MMD into VICReg extends invariance from pairwise alignment to batch-level consistency. This enables VICReg to capture a broader range of dependencies and align features across diverse transformations within the batch. MMD's ability to enforce a unified feature distribution enhances robustness, making representations more stable and aligned with the data manifold. This approach addresses limitations of pairwise invariance, reinforcing the model's generalizability [15].

3 Methodology

Our proposed approach integrates Maximum Mean Discrepancy (MMD) into VICReg's to address the limitations of pairwise invariance by enforcing consistency across all samples in a batch. This section details the formulation of our objective function and the theoretical motivation behind using MMD to strengthen invariance in VICReg.

3.1 Objective: VICReg with MMD-Enhanced Invariance

VICReg aims to produce invariant features by minimizing the distance between representations of paired augmentations of the same input. However, this pairwise approach may be limited in capturing broader batch-level dependencies [16] and does not fully account for consistency across different transformations within a batch. MMD addresses this by aligning feature distributions across all augmented views in the batch.

To achieve this, we incorporate an MMD term into VICReg's loss, forming an objective function that balances both pairwise and batch-level alignment:

$$\mathcal{L} = \lambda\, \mathcal{L}_{\text{inv}} + \mu\, \mathcal{L}_{\text{var}} + \nu\, \mathcal{L}_{\text{cov}} + \alpha\, \mathcal{L}_{\text{MMD}}, \quad (3)$$

where $\mathcal{L}_{\text{inv}}$ denotes the original pairwise invariance in VICReg, and $\mathcal{L}_{\text{MMD}}$ is the MMD term that aligns feature distributions across the batch, defined in 2. The parameter α controls the influence of MMD relative to other terms, balancing pairwise invariance with batch-level alignment.

3.2 Theoretical Motivation for MMD-Enhanced Invariance in VICReg

The addition of MMD to VICReg's objective function offers several theoretical benefits:

Batch-Level Invariance Unlike pairwise objectives that enforce invariance only between specific views, MMD aligns all augmented samples in the batch. This caused a unified feature space where diverse transformations are consistently mapped. This batch-level consistency smooths variations among representations and contributes to a more robust, transformation-invariant representation [9].

Improved Generalization and Domain Adaptation By aligning feature distributions in a batch of different augmentations, MMD promotes a representation that is less susceptible to overfitting on individual transformations. This supports better generalization for unseen transformations [17]. Additionally, MMD has been shown to improve domain adaptability by aligning feature distributions in a manner that is stable across domains. It makes the learned representations more transferable across data settings [18].

Manifold Alignment and Higher-Order Dependencies MMD's alignment in RKHS allows the model to capture both linear and non-linear dependencies across the batch, supporting a manifold-aware alignment of features that respects the intrinsic geometry of the data space. This is especially beneficial for complex, non-linear datasets, where capturing higher-order dependencies aids in representing the underlying structure of the data [9].

3.3 Parameter Selection and Kernel Choice

The effectiveness of MMD is sensitive to the choice of kernel. In our experiments, we employ a Gaussian kernel:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \tag{4}$$

where σ controls the scale of the kernel, affecting the degree to which local versus global dependencies are captured. Choosing an appropriate σ is critical for the effectiveness of MMD. A smaller σ emphasizes local dependencies, making MMD sensitive to small-scale variations. A larger σ captures global dependencies and smooths out finer details. To balance these effects, we selected σ empirically testing a range of values. After experimentation, we selected $\sigma = 1$ as it provided the best results on the validation set, demonstrating a balanced contribution of MMD to batch-level alignment while maintaining the effectiveness of VICReg's original objectives.

Combined Impact The combined VICReg-MMD objective leverages both pairwise and batch-level alignment. By enhancing batch-level invariance through MMD, our method achieves representations that are robust to diverse transformations and resilient to domain variations. This dual approach provides a balanced regularization strategy, yielding high-quality self-supervised representations suited for downstream tasks.

4 Experiments

To assess the effectiveness of our proposed VICReg with MMD-enhanced invariance, we conduct experiments on MNIST, CIFAR-10, and STL-10 datasets. These datasets vary in complexity, providing a robust evaluation of our method's ability to capture invariance across simple and complex data distributions. We measure the quality of the learned representations by evaluating their performance on linear classification, and clustering accuracy. For evaluating the learned representations, we used a Multi-Layer Perceptron (MLP) classifier with two fully connected layers. The first layer maps the input dimension to 256 features, followed by a ReLU activation, and the second layer maps to 10 classes. For clustering, we used the k-Nearest Neighbors (k-NN) algorithm [19].Our experiments are designed to highlight how batch-level alignment with MMD improves representation consistency and robustness compared to baseline VICReg.

4.1 Experimental Setup

We implemented a lightweight version of the VICReg architecture, using a three-layer convolutional network as the backbone. Each image in these datasets is processed through standard data augmentations, including random cropping, flipping, and color jittering. These augmentations introduce variations that our model learns to be invariant to, with MMD contributing to the alignment of the feature distributions across all augmented samples in the batch.

Dataset	VICReg		VICReg + MMD	
	Classifier	Clustering	Classifier	Clustering
MNIST	88.67%	81.65%	91.68%	88.26%
CIFAR-10	59.73%	48.96%	61.96%	50.36%
STL-10	53.78%	40.66%	60.22%	42.26%

Table 1: Performance comparison of VICReg and VICReg + MMD on MNIST, CIFAR-10, and STL-10, evaluated by linear classifier and clustering accuracy.

4.2 Results and Analysis

The results in Table 1 indicate that incorporating MMD into VICReg consistently improves representation quality across all datasets. The most significant improvements are observed on the STL-10 dataset, which contains high-dimensional and complex images, where MMD-enhanced VICReg achieves a classifier accuracy increase from 53.78% to 60.22% and a clustering accuracy increase from 40.66% to 42.26%. This improvement suggests that MMD's batch-level alignment is especially effective in scenarios where invariance across complex transformations is critical.

Batch-Level Consistency Integrating MMD into VICReg enhances clustering performance, as demonstrated by improved metrics on complex datasets like CIFAR-10 and STL-10. This aligns with prior work highlighting the effectiveness of batch-level distribution alignment in capturing richer feature representations [9]. By enforcing consistency across the entire batch in a Reproducing Kernel Hilbert Space (RKHS), MMD reduces variability among augmentations, which is theoretically supported to promote more cohesive clustering [10]. While the observed increase in clustering accuracy provides empirical evidence of this benefit, further exploration is needed to fully characterize the relationship between batch-level alignment and feature space stability in diverse self-supervised settings.

Improved Robustness and Generalization The benefits of MMD's batch-level alignment are evident in

the linear classification results. By ensuring that augmented samples across the batch map to a cohesive feature distribution, MMD reduces variability, which enhances feature robustness and generalizability across transformations. This improvement is consistent across datasets, highlighting the advantage of enforcing batch-level invariance in a variety of data contexts.

4.3 t-SNE Visualization

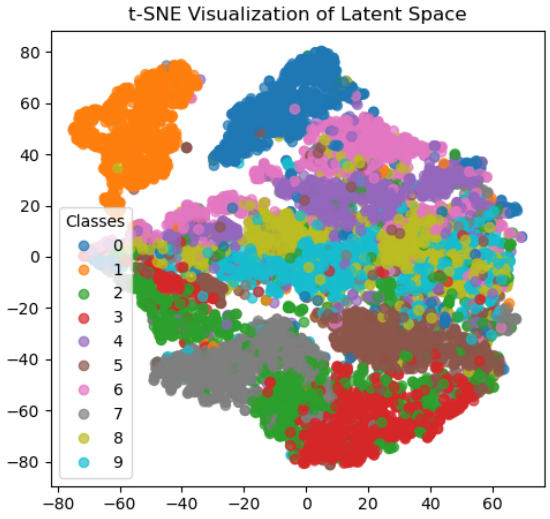
To illustrate the impact of MMD on representation structure, we visualize the learned features using t-SNE on the MNIST dataset. As shown in Figure 1, representations learned by VICReg + MMD exhibit more distinct clusters with less overlap between classes compared to the baseline VICReg. This improvement in cluster separability suggests that MMD's batch-level alignment supports a more organized feature space, making the learned representations more interpretable and easier to classify.

4.4 Ablation Study on MMD Influence

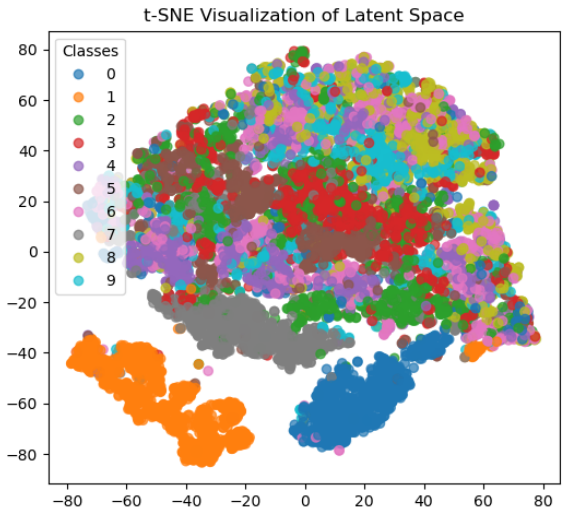
We further conducted an ablation study to analyze the effect of the MMD term's weight (α) on representation quality. Figure 2 shows that as α increased, clustering accuracy improved up to a certain threshold, beyond which performance plateaued. This suggests that MMD's influence is most effective when balanced with other VICReg objectives, reinforcing that batch-level alignment enhances invariance without overwhelming pairwise alignment.

4.5 Summary of Findings

Our experimental results validate the hypothesis that incorporating MMD into VICReg enhances representation quality by strengthening batch-level invariance. MMD's ability to align feature distributions across the batch yields representations that are not only more invariant but also better suited for downstream tasks, particularly in complex data contexts. These results highlight the importance of batch-level consistency in self-supervised learning, demonstrating MMD's role as a valuable addition to VICReg for manifold-aware, robust representation learning.



(a) VICReg + MMD



(b) VICReg

Figure 1: t-SNE visualization of representations on MNIST: (a) VICReg + MMD, (b) VICReg. The addition of MMD leads to more distinct, well-separated clusters.

5 Conclusion

We introduced an enhancement to VICReg by incorporating a Maximum Mean Discrepancy (MMD) term to address the limitations of pairwise invariance in self-supervised representation learning. By aligning feature distributions across the batch in a Reproducing Kernel Hilbert Space (RKHS), MMD promotes batch-level consistency, resulting in a unified and stable feature space.

Our experiments on MNIST, CIFAR-10, and STL-10 show that MMD improves representation quality, as evidenced by higher linear classification accuracy and enhanced clustering performance. These results underscore the benefits of capturing higher-order dependencies and suggest that batch-level alignment contributes to more robust and generalizable features.

Future work could explore MMD's integration into other self-supervised frameworks, potentially enhancing the robustness of methods like SimCLR or BYOL. Additionally, testing this approach on larger and more complex datasets would further reveal the role of batch-

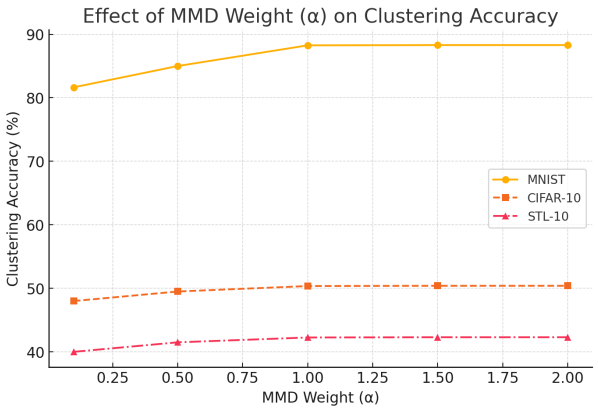


Figure 2: Enter Caption

level invariance. Our VICReg-MMD approach lays the groundwork for more effective manifold-aware representations, supporting a wide range of downstream applications.

References

[1] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *International Journal of Computer Vision*, vol. 132, no. 1, pp. 208–223, 2024.

[2] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] H. Sepanj and P. Fieguth, "Context-aware augmentation for contrastive self-supervised representation learning," *Journal of Computational Vision and Imaging Systems*, vol. 9, no. 1, pp. 4–7, 2023.

[4] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022, pp. 1–12.

[5] Y. Chen, A. Bardes, Z. Li, and Y. LeCun, "Intra-instance vicreg: Bag of self-supervised image patch embedding," *arXiv preprint arXiv:2206.08954*, vol. 2, 2022.

[6] S. T. P. Raghu, D. T. MacIsaac, and E. J. Scheme, "Self-supervised learning via vicreg enables training of emg pattern recognition using continuous data with unclear labels," *arXiv preprint arXiv:2409.11632*, 2024.

[7] H. Cheng, H. Wen, X. Zhang, H. Qiu, L. Wang, and H. Li, "Contrastive continuity on augmentation stability rehearsal for continual self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5707–5717.

[8] W. Sun, Y. Zhang, Z. Qin, Z. Liu, L. Cheng, F. Wang, Y. Zhong, and N. Barnes, "All-pairs consistency learning forweakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 826–837.

[9] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[10] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[11] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nystr\" om method, and use of kernels in machine learning: Tutorial and survey," *arXiv preprint arXiv:2106.08443*, 2021.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[13] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[14] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*. PMLR, 2021, pp. 12 310–12 320.

[15] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, "Domain adaptation with conditional transferable components," in *International conference on machine learning*. PMLR, 2016, pp. 2839–2848.

[16] P. Koromilas, G. Bouritsas, T. Giannakopoulos, M. Nicolaou, and Y. Panagakis, "Bridging minibatch and asymptotic analysis in contrastive learning: From infonce to kernel-based losses," *arXiv preprint arXiv:2405.18045*, 2024.

[17] W. Wang, H. Li, Z. Ding, F. Nie, J. Chen, X. Dong, and Z. Wang, "Rethinking maximum mean discrepancy for visual domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 264–277, 2021.

[18] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[19] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. Springer, 2003, pp. 986–996.