WALKING ACROSS IMAGES: RANDOM WALK PAIRING FOR SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to capture both local invariances and global semantic structures is essential for effective self-supervised learning. In this paper, we address a fundamental challenge of ensuring robust generalization to unseen data while effectively capturing global semantic structures. Motivated by a spectral embedding perspective, we identify potential sources of suboptimal generalization in SSL methods that can be viewed through a spectral lens. We propose a simple yet effective training strategy that can be easily applied to a variety of SSL methods. We introduce SAG-VICReg (Stable and Generalizable VICReg) as a detailed instantiation, demonstrating improvements in generalization and global semantic understanding. Through comprehensive experiments across diverse SSL frameworks including SimCLR and TCL, we show broad applicability with consistent improvements. The same pairing also improves the fine-tuning of today's strongest pretrained backbones, with gains that persist under linear probing, grow as labels become scarce, and transfer to unseen datasets. Our enhanced methods exhibit superior performance on metrics designed to evaluate global semantic understanding while maintaining competitive results on local evaluation metrics. We also propose a label-free evaluation metric that accounts for global data structure without requiring labels–a key advantage when labeled data is scarce or unavailable.

1 INTRODUCTION

Self-supervised learning has emerged as a transformative paradigm in computer vision, achieving remarkable success through a variety of methods (e.g., Chen & He (2021); Tian et al. (2020b); Caron et al. (2021); Ermolov et al. (2021); Gidaris et al. (2021); He et al. (2022); Assran et al. (2022; 2023); Garrido et al. (2024)). A fundamental challenge in this paradigm is ensuring robust generalization to unseen data while effectively capturing global semantic structures. While numerous approaches exist to address these generalization challenges, in this work we adopt a spectral perspective to understand and tackle these issues.

Most SSL methods form positive pairs from augmentations of the same image. While effective, this may limit exposure to cross-image semantic relationships, such as viewpoint changes or intra-class variation that augmentations alone may not capture. To understand why same-image pairing alone may limit generalization, we adopt a spectral perspective. This proves particularly insightful because several prominent SSL methods can be viewed as spectral embedding techniques (e.g., Chen et al. (2020); Zbontar et al. (2021); Bardes et al. (2022)). By examining these methods through this framework, we identify a source of suboptimal generalization and develop a targeted solution.

While our analysis is grounded in spectral theory, the resulting approach is broadly applicable across various SSL frameworks. As a case study, we focus on VICReg Bardes et al. (2022), which serves as an ideal testbed due to its straightforward formulation and strong baseline performance, allowing us to clearly demonstrate the benefits of our enhancements. Through this spectral lens, we observe that same-image pairing may create a fragmented graph structure–each image forms an isolated cluster connected only through its augmentations. While Balestriero & LeCun (2022) initially observed this spectral connection, we extend this insight to explain potential generalization limitations.

Capturing semantic relationships that span across samples, while ensuring generalization to unseen data, remains a central challenge in SSL. Recent advances in joint-embedding predictive architectures (JEPAs) and world models LeCun et al. (2022); Assran et al. (2023); Garrido et al. (2024)

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.

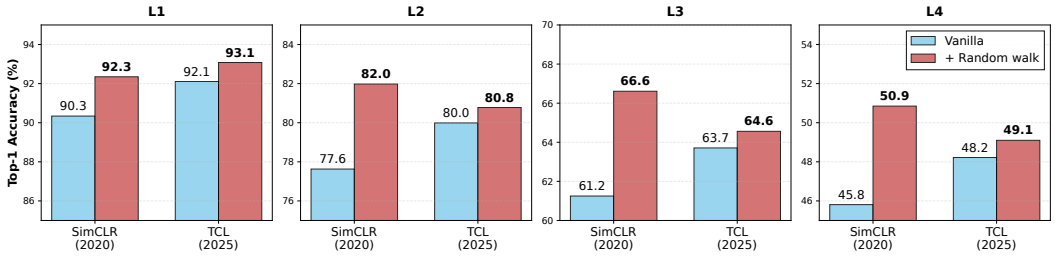


Figure 1: Our random walk-guided pairing strategy yields consistent improvements across different SSL methods and hierarchy levels. On CIFAR-100 with ResNet-34 trained for 500 epochs, k-NN evaluation shows improvements across all hierarchy levels (L1-L4), where L1 represents coarse-grained categories and L4 represents finest-grained classes. Notably, our approach improves SimCLR Chen et al. (2020), elevating it to performance on par with the state-of-the-art TCL Animesh & Chandraker (2025), demonstrating that our simple, architecture-free method can modernize existing approaches to achieve contemporary performance.

emphasize learning semantic similarities in latent space rather than relying solely on hand-crafted augmentations. These approaches share a common motivation with our work: capturing richer semantic structures across samples, beyond variations of single instances. The spectral perspective we adopt serves as a diagnostic tool to understand why same-image pairing may limit generalization–as illustrated in Section 4, but the underlying principle aligns with this broader trend in SSL: models benefit from learning broader semantic relationships.

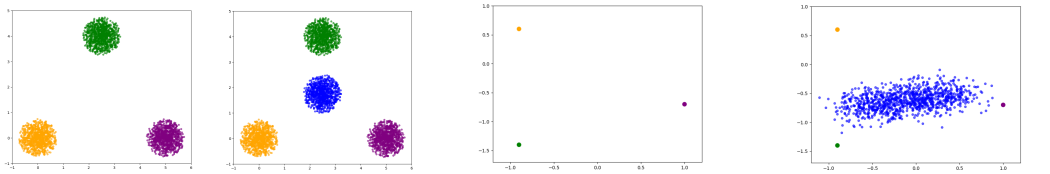
This motivates us to propose an enhanced training strategy that encourages models to recognize not only relationships between augmentations of the same image, but also meaningful semantic relationships between distinct images across the dataset. This broader perspective enables models to develop a richer understanding of global data structures and semantic hierarchies while improving generalization. Importantly, our approach is simple to implement, requires no architectural changes, and can be applied to various SSL frameworks–as we demonstrate through experiments, and shown in Figure 1, with consistent gains across all levels.

In proposing a method to improve model stability and generalization, we must determine whether it truly yields higher-quality embeddings. While existing evaluation methods (e.g., linear or semi-supervised classification) mainly gauge linear separability, they may fail to capture whether learned representations reflect the full scope of data structure, as discussed in Section 4.2. This issue is particularly important for datasets that exhibit hierarchical structures such as CIFAR Krizhevsky et al. (2009) and ImageNet Deng et al. (2009), where representations should capture both intra-class variation and inter-class semantics. We therefore propose a new standalone evaluation metric applicable to any embedding model and accounts for the global structure of the data. This metric directly evaluates the embeddings themselves without requiring any labels–a significant advantage in scenarios where labeled data is scarce or unavailable. Additionally, it enables us to assess the stability of any model when encountering new, unseen data. It also demonstrates that our techniques indeed effectively enhance this stability.

Our contributions are threefold: (1) We adopt a spectral perspective revealing that same-image pairing fragments graph connectivity, may limit generalization to unseen data. (2) We propose a simple yet effective strategy for learning richer semantic relationships beyond same-image augmentations, requiring no architectural changes and experimentally validated across SSL frameworks. (3) We introduce a label-free evaluation metric that captures global structure and hierarchical relationships, complementing traditional metrics.

2 Related Work

Generalization in SSL and spectral based methods. Modern SSL is dominated mostly by two main paradigms: contrastive methods that learn by distinguishing between similar and dissimilar samples Chen et al. (2020; 2021); Yang et al. (2025); Sharma et al. (2024); Animesh & Chandraker (2025), and self-distillation approaches that avoid explicit negative sampling through clustering or



(a) train points (b) test points (c) train points' embedding (d) test points' embedding

Figure 2: **Spectral embedding example demonstrating unstable generalization from fragmented connectivity.** (a,b) Input-space coordinates: training data contain three isolated clusters, while test data add an unseen blue cluster. (c,d) Embedding-space coordinates: training clusters map to a clean structure, but points from the unseen cluster are scattered and distorted relative to the original clusters. This illustrates how limited connectivity can lead to unstable embeddings for unseen data. While shown here in 2D, it illustrates issues that may occur in high-dimensional spaces as well.

teacher-student frameworks Caron et al. (2020; 2021); Oquab et al. (2024); Kang et al. (2023); Hızlı et al. (2025).

These methods primarily achieve generalization through two key strategies: extensive data augmentations Chen et al. (2020); Tian et al. (2020b); Caron et al. (2021), and large-scale pretraining Oquab et al. (2024); Singh et al. (2023); Dehghani et al. (2023). Recent work explores methods that go beyond single-image augmentations to improve generalization. Various approaches capture richer representations through cross-image relationships Tian et al. (2020a); Caron et al. (2020); Dwibedi et al. (2021); Estepa et al. (2023); Ge et al. (2023); Lebailly et al. (2024); Gao et al. (2024); Li et al. (2024); Long et al. (2025); Alshammari et al. (2025) while others focus on extracting richer within-image representations Zhou et al. (2022); Assran et al. (2022); Wang et al. (2023); Tian et al. (2023). Wen et al. (2025) is closest: it also uses a spectral view and k-NN positives, but targets dense/patch-level SSL, whereas we focus on image-level SSL with a simpler modification. Our work belongs to this cross-image trend, but is spectrally motivated, simpler than many prior approaches, and requires neither complex training procedures nor memory banks, reducing reliance on large-scale pretraining.

Traditional spectral embedding techniques Bengio et al. (2003); Fowlkes et al. (2004); Coifman & Lafon (2006) are primarily designed for fixed datasets and are not intended to generalize to new points. While approaches like Drineas & Mahoney (2005) partially address this through auxiliary mappings, they assume proximity between new and original data. SpectralNet Shaham et al. (2018) focuses on scaling spectral methods to large datasets, but also aims to tackle the problem of generalization, that the traditional methods do not address.

Evaluation of Image Embeddings. Standard SSL evaluation relies mainly on linear or k-NN probes, while hierarchy-aware protocols evaluate performance at multiple label granularities Silla Jr & Freitas (2011); Wehrmann et al. (2018); Giunchiglia & Lukasiewicz (2020); Valmadre (2022); however, these require predefined hierarchies or ground-truth labels. Recent unsupervised representation-quality measures Agrawal et al. (2022); Garrido et al. (2023); Thilak et al. (2024) instead estimate quality from statistics of a single embedding space, such as spectral decay, effective rank, or discriminative rank, and are primarily used as proxies for downstream probe performance. Dai et al. (2026) is closer in spirit, as it also considers structure in the representation space, but its objective is to predict dense downstream performance from class-separability and dimensionality cues rather than to assess whether hierarchical semantic organization is preserved. In contrast, our metric is explicitly structural and label-free: it compares the organization induced by two embedding spaces, allowing us to evaluate whether global and hierarchical relationships remain stable when generalizing to unseen data.

3 PRELIMINARIES

In order to explain the spectral connection, we use SpectralNet Shaham et al. (2018) and in Section 4 we establish the equivalence to VICReg, which is our case study.

3.1 SPECTRALNET

SpectralNet is a deep learning approach to spectral clustering, which is intended to tackle the scalability and generalization ability issues traditional spectral embedding methods lack. The loss of SpectralNet is defined as:

$$L_{\text{SpectralNet}}(\theta) = \frac{1}{n^2} \sum_{i,j=1}^{n} W_{i,j} \|y_i - y_j\|_2^2, \tag{1}$$

where n is the batch size, $y_i \in \mathbb{R}^k$, $y_i = F_\theta(x_i)$, k is the embedding dimension, y_i is the learned embedding of the data point x_i. The map F is a neural network with parameters θ and W is a $n \times n$ matrix such that $W_{i,j} = w(x_i, x_j)$, where w is a kernel function that quantifies the similarity between the points x_i, x_j. The loss is computed subject to an orthogonality constraint which is enforced using an orthogonalization layer in the architecture:

$$\frac{1}{n} Y^T Y = I_{k \times k}, \tag{2}$$

where Y is a $n \times k$ matrix of the outputs whose i-th row is y_i^T. SpectralNet is a spectral embedding method because the matrix Y, which minimizes the loss equation 1 while satisfying equation 2, corresponds to the leading eigenvectors of the Laplacian $L = D - W$, where D is the degree matrix, as explained in Appendix J.1. Minimizing equation 1 is equivalent to minimizing: trace $\big(Y^T(D-W)Y\big) = \text{trace}\,\big(Y^T L Y\big)$. Hence, SpectralNet functions as a spectral embedding technique. As it is a parametrized map (by a neural network), it has the ability to generalize to new points during inference.

3.2 VICREG

VICReg is an SSL method designed to produce robust embeddings by imposing specific regularization terms. The invariance term makes sure that the model is invariant to different augmentations, while the covariance and variance terms prevent a collapse.

Variance Regularization: This term ensures adequate spread in embeddings along each dimension. For a batch of embeddings $Y = [y_1, y_2, \ldots, y_n]$, define $y^j = (y_1^j, \ldots, y_n^j)^T$. The term is: $v(Y) = \frac{1}{k} \sum_{j=1}^{k} \max\big(0, \gamma - S\big(y^j, \epsilon\big)\big)$, where S is the regularized standard deviation defined by: $S(x, \epsilon) = \sqrt{\text{Var}(x) + \epsilon}$, k is the embedding dimension, γ is a target threshold and ϵ is a small constant.

Invariance Regularization: This term promotes consistency between different augmentations of the same image, effectively encourages semantically closer images to have similar embeddings. For two sets of embeddings Y and Y' of augmented views: $s(Y, Y') = \frac{1}{n} \sum_{i=1}^{n} \|y_i - y_i'\|_2^2$.

Covariance Regularization: This term minimizes redundancy by reducing correlations among embedding dimensions. For a batch of embeddings Y, the overall term is: $c(Y) = \frac{1}{k} \sum_{i \neq j} [C(Y)]_{i,j}^2$, where $C(Y)$ is the sample covariance matrix of Y. The overall loss function is a weighted combination of the three terms.

4 PROPOSED APPROACH

4.1 VICREG AS A SPECTRAL EMBEDDING METHOD

We first establish the fundamental connection between VICReg and spectral embedding, following the mathematical introduction provided in Appendix J and Section 3. Let Y, Y' denote the embeddings of two augmented batches. The VICReg invariance term can be written as $s(Y, Y') = \frac{1}{n} \sum_{i,j=1}^{n} W_{ij} \|y_i - y_j'\|_2^2$, where $W_{ij} = 1$ if $i = j$, i.e., y_i and y_j' come from two augmentations of the same image, and $W_{ij} = 0$ otherwise. This is precisely the same expression used in SpectralNet's loss equation 1, but in the specific case where $W_{ij} \in \{0, 1\}$, whereas in SpectralNet, $W_{ij} \in (0, 1]$. Additionally, when we consider the covariance term combined with the variance term in VICReg, we effectively obtain the identity covariance constraint[1] which is exactly the same orthonormalized

[1]Under the assumption of zero mean of the vectors $y_i \in Y$, but the Laplacian's eigenvectors are orthogonal to the constant vector (the first eigenvector), ensuring that the mean from the second and on is zero.

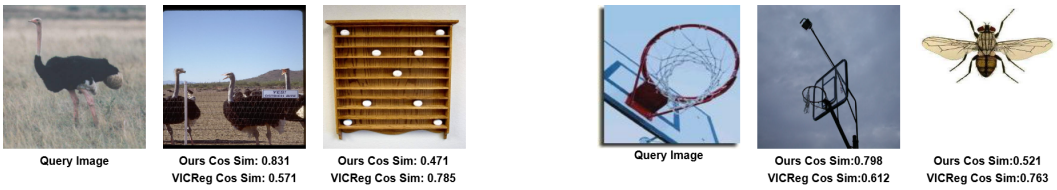


Figure 3: **A real-world demonstration of how VICReg behaves unstably when encountering images from unseen clusters during test time.** We compare cosine similarity scores between the query image (leftmost) and other images, showing results from both VICReg and our approach. Notice how VICReg's embeddings can place semantically similar images far apart and semantically unrelated images too close together–an empirical demonstration of the unexpected behavior.

output constraint equation 2 as in SpectralNet. Thus, VICReg can be viewed as a special spectral-embedding instance on a graph whose clusters are individual images: each cluster contains that image's augmentations, with edges only within the cluster and zero weight across different images. Therefore, VICReg is also a spectral embedding method, just like SpectralNet. This spectral view reveals a potential generalization issue: when connectivity is limited to within-cluster relationships, models may exhibit unstable behavior on data from unseen clusters and may fail to capture meaningful semantic structures that span across the dataset. As illustrated in Figure 2, such fragmented connectivity can produce clean embeddings for training clusters but unstable and distorted embeddings for samples from unseen clusters.

Figure 3 provides a real-world demonstration of the embedding inconsistency we identified, directly paralleling the 2D illustration in Figure 2. This empirical demonstration aligns with the findings in Figure 2, in that ostrich images are embedded in a manner lacking meaningful structure–so that a wooden shelf appears closer in the embedding space than other ostrich (just like the blue cluster in (d) that were not present during training, whose points are much dispersed). Alongside these illustrations, Figure 9 in Appendix E provides UMAP visualizations comparing our approach to VICReg when embedding images from unseen classes. These visualizations highlight our approach's ability to maintain coherent semantic structure when generalizing to new, unseen data, while VICReg's embeddings exhibit noticeable distortions and inconsistencies.

4.2 Rationale and Intuition

We now describe the rationale behind our proposed approach and the modifications we introduce to improve generalization and stability to new, unseen data in SSL methods while capturing richer semantic structures across the dataset. Alongside these modifications, we will give the motivation and illustrate why metrics like linear classification may be limited in capturing whether the learned representations reflect the full scope of the data, and how our method overcomes it. In addition, we provide a theoretical motivation to our approach in Appendix A.

Random walk-guided invariance. Motivated by the observation that same-image pairing creates fragmented graphs with isolated clusters, we enrich the graph connectivity to enhance generalization and semantic structure learning. Our approach modifies the invariance term by pairing not only augmentations of the same image, as in standard SSL, but also augmentations of different images that are "close" in some semantic sense. To enhance this effect, we weight this invariance term according to the similarity of each pair. This general adaptation encourages models to produce embeddings that are more stable and meaningful for new, unseen data while learning broader relationships between data points. When applied to VICReg, we name the modified algorithm SAG-VICReg.

Embedding evaluation. We propose a new metric that goes beyond linear separability to evaluate the structural integrity of embeddings, *without requiring any labels*. The rationale is that linear-separability scores tell only part of the story–an embedding can preserve the data's global structure even when classes are not cleanly split by a linear probe, as the top row of Figure 8 in Appendix E shows. To capture this broader quality notion, we introduce a *label-free* metric that evaluates embeddings directly. Rather than focusing solely on immediate neighbors like k-NN, it analyses relationships across the entire dataset, revealing coarse-to-fine semantic structure. As it is computed

on the representations themselves without any downstream task or ground-truth labels, it offers a direct, *label-free* measure of structural integrity, as can be seen in the bottom row of Figure 8.

4.3 SAG-VICReg: Random walk-guided invariance

Our approach initially follows the VICReg procedure up to the point where two batches of embeddings are produced. Let $S = \{s_1, s_2, \ldots, s_n\} \subset \mathcal{D}$ be a mini-batch of images sampled from dataset $\mathcal{D}$. For each image $s_i \in S$, we apply two random transformations $t, t' \sim \mathcal{T}$, where $\mathcal{T}$ is a set of augmentations. The transformed views x and x' are then passed through a SAG-VICReg network. First, they are encoded by the feature extractor f_θ (which is the backbone, a ResNet He et al. (2016) or a ViT Dosovitskiy et al. (2020)) into representations $y = f_\theta(x)$ and $y' = f_\theta(x')$. These representations are then mapped by an expander h_ϕ (a Neural network consisting of fully-connected layers) to the embeddings $z = h_\phi(y)$, $z' = h_\phi(y')$, which will be used to compute the loss. The embeddings $Z = \{z_1, z_2, \ldots, z_n\}$,$Z' = \{z'_1, z'_2, \ldots, z'_n\}$ form two batches, for further processing.

At this stage, we introduce our primary contribution by incorporating relationships among the embeddings. In order to capture relationships between the embeddings, we construct an affinity matrix $W \in \mathbb{R}^{n \times n}$ based on cosine similarity. For each sample $z_i \in Z$, we identify its k-nearest neighbors in Z' and compute the cosine distance $d_{ij} = 1 - \cos(z_i, z'_j)$ between z_i and each of its neighbors z'_j. We define each entry of W using a Gaussian kernel: $W_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma_i^2}\right)$, where σ_i is a local scaling factor whose value varies per point. Implementation details including local scaling factor calculations are provided in Appendix L, with hyperparameters considerations in Appendix C and computational overhead in Appendix D. SAG-VICReg remains stable across different hyperparameter configurations. Next, we construct a random walk matrix P and sample pairs as in Appendix J.2. Some examples of non-trivial pairs generated through this sampling are shown in Figure 4. This sampling process creates a new batch of pairs (z_i, z''_i), where z''_i is selected from Z' using the random walk P starting from z_i. Finally, we compute the invariance term in the VICReg loss, weighted by the similarity values W_{ij}. Specifically, the invariance constraint is adjusted by weighting each pair (z_i, z''_i) according to W_{ij}. The remaining terms in the VICReg loss function remain unchanged. To limit false-positive pairs, pairing starts after a warm-up of plain VICReg (50/90/150 epochs on CIFAR-100/Caltech-256/ImageNet). Our random walk pairing can be viewed as a form of *nearest-neighbour debiasing*, similar to the neighbour-propagation step used in I-Con Alshammari et al. (2025). The key difference is motivational: I-Con debiases to correct false-negative noise from random negative sampling, whereas our goal is to bolster out-of-cluster generalization. The full pseudo code is described in Algorithm 1 in Appendix I.



Figure 4: **Non-trivial pairs.** Each column shows the random walk start image above and the selected endpoint below.

Applicability to other SSL methods. Although instantiated with VICReg, random walk pairing applies more broadly to SSL objectives that consume positive pairs: the sampled partner replaces the host method's original positive target, while its architecture and remaining losses remain unchanged (Appendix B.10 describes this for each method). We validate this beyond VICReg on pretrained models, including DINOv2, and on additional SSL methods trained from scratch; full results appear in Section 5 and Appendix B.

4.4 Global Structure-Preserving Evaluation

We introduce a novel *label-free* evaluation metric designed to capture both the hierarchical and global structure preservation of embeddings, as well as the generalization capabilities of embedding models. This is a stand-alone contribution, useful for assessing embedding models in general. Let $A = \{a_1, a_2, \ldots, a_n\} \subset \mathbb{R}^d$ and $B = \{b_1, b_2, \ldots, b_n\} \subset \mathbb{R}^m$ be two sets of embeddings, each containing n vectors, where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}^m$ are embeddings of the same entity for $i = 1, \ldots, n$. The dimensions d and m may or may not be equal. These sets might come from two

different models or represent original data alongside its embeddings. We use these two sets to construct hierarchical trees (dendrograms), and evaluate their structural similarity, capturing how well the global organization of one embedding space corresponds to the other. Using this metric, we can assess the generalization capabilities of models. For example, if A represents embeddings generated by a model on data it was trained on and B represents embeddings of data the model has not seen before, high similarity between their dendrograms indicates that the model has effectively captured the underlying data relationships, even in unseen data.

Hierarchical Tree Construction. To construct hierarchical trees for A, B, we use an agglomerative clustering algorithm McQuitty (1957) with cosine distance and Ward linkage Anderberg (2014), which minimizes within-cluster variance, creating compact clusters and a clear hierarchy.

Dendrogram-based Similarity. Once the dendrograms are built, we compare their structure using two complementary quantities. First, for every pair of embeddings, we compute the distance to their *lowest common ancestor* (LCA) in each tree; smaller LCA distances indicate closer relationships within the dendrogram. We then correlate the LCA-distance vectors from the two trees using Pearson, Spearman, and Kendall correlations, which we call **LCA similarity**. Second, we compute **cophenetic similarity** (see Appendix J.3), which correlates dendrogram-based cophenetic distances, i.e., the merge heights at which pairs first join the same cluster, with the pairwise distances in the other embedding space. Together, these quantities assess whether hierarchical and local-to-global relationships are preserved across embedding spaces.

Interpretation and Generalization. By correlating dendrogram-based distances, our metric quantifies how well an embedding preserves the *full, multi-scale hierarchy* of the data. In Figure 8 in Appendix E, embedding 2 yields dendrograms that closely mirror those of the original points, resulting in substantially higher LCA and cophenetic correlations than embedding 1. As the computation is entirely *label-free* and involves no downstream task, the metric remains applicable even when annotations are scarce or absent. It complements local tests such as *k*-NN or linear probes by capturing broader semantic relationships and hierarchical structures across the dataset, rather than only pairwise similarity or class separability. Our metric also naturally extends to measuring generalization, as it assesses whether a model preserves similar structural organization on unseen data. The full pseudocode is provided in Algorithm 2 in Appendix I, with hyperparameter analysis in Appendix C. Section 5 compares it with other label-free measures.

5 Experiments and Results

We demonstrate the broad applicability of our approach by integrating it with multiple SSL methods including VICReg, SimCLR, and TCL Animesh & Chandraker (2025). We conduct comprehensive comparisons against baseline spectral embedding methods (VICReg, Barlow Twins, SimCLR).

Datasets and Architecture. We use 3 well-known datasets, ImageNet, CIFAR-100 and Caltech-256 Griffin et al. (2007). These datasets are selected not only because they are well-studied, but also because they have clear hierarchy Yan et al. (2015); Redmon & Farhadi (2017); Kim et al. (2023). This makes them suitable for evaluating whether our approach captures global semantic structure across different SSL methods. The architecture is detailed in Appendix F.

Setup. A model with strong generalization capabilities should embed an image similarly whether or not it encountered that image (or its class) during training. To test this, we examine how the embeddings for the same images differ depending on whether the model has seen those images during training or not. The information on the experiments is provided in Appendices G and H.

We evaluate our approach using multiple complementary metrics. First, we evaluate hierarchical linear and *k*-NN classification at different hierarchy levels (where level 1 corresponds to high-level category labels and level 4 to the finest-grained class label) using ground-truth labels. This reveals whether our approach enhances both coarse-grained category understanding and fine-grained class discrimination across SSL methods.

Second, to evaluate generalization capabilities, we compare embedding sets generated by models trained on different class subsets using LCA and Cophenetic similarity (Section 4.4). Higher corre-

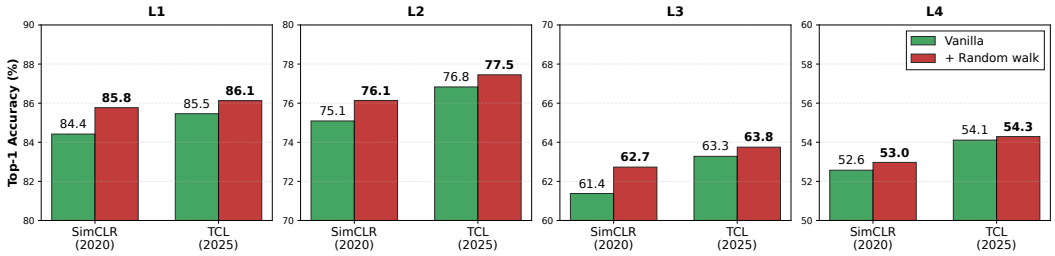


Figure 5: Cross-dataset generalization: hierarchical linear classification on CIFAR-100 using models pretrained on Caltech-256. Our approach yields consistent improvements across all hierarchy levels, demonstrating robust transfer learning and cross-dataset applicability.

lation indicates better generalization ability. Third, we also evaluate the embeddings with the Rand index at each level of the CIFAR-100 hierarchy, where SAG-VICReg achieves the highest scores at every level (Appendix B.6).

Results. We begin by demonstrating the broad effectiveness of our approach across different SSL methods and training scenarios. Figure 1 shows results when pretraining and evaluation are on the same dataset (CIFAR-100), while Figure 5 demonstrates cross-dataset transfer (Caltech-256 to CIFAR-100). Both scenarios reveal consistent improvements across all hierarchy levels, confirming that our method enhances generalization whether within-dataset or across different visual domains. Additional results using these methods are provided in Appendix B.

Table 1: LCA similarity results. Relative gains over the second-best are shown; best scores are bolded.

	ImageNet 1k			CIFAR-100			Caltech-256		
	Pearson	**Spearman**	**Kendall**	**Pearson**	**Spearman**	**Kendall**	**Pearson**	**Spearman**	**Kendall**
SimCLR	0.248 ± 0.04	0.348 ± 0.1	0.296 ± 0.06	0.235 ± 0.02	0.226 ± 0.01	0.187 ± 0.05	0.289 ± 0.07	0.262 ± 0.03	0.211 ± 0.04
Barlow Twins	0.247 ± 0.02	0.321 ± 0.02	0.257 ± 0.02	0.204 ± 0.01	0.219 ± 0.06	0.181 ± 0.05	0.306 ± 0.02	0.317 ± 0.11	0.254 ± 0.09
VICReg	0.287 ± 0.03	0.403 ± 0.06	0.329 ± 0.05	0.202 ± 0.01	0.223 ± 0.02	0.193 ± 0.02	0.266 ± 0.01	0.239 ± 0.04	0.201 ± 0.04
SAG-VICReg (Ours)	$\mathbf{0.318 \pm 0.04}$	$\mathbf{0.433 \pm 0.07}$	$\mathbf{0.351 \pm 0.05}$	$\mathbf{0.268 \pm 0.02}$	$\mathbf{0.296 \pm 0.12}$	$\mathbf{0.25 \pm 0.09}$	$\mathbf{0.321 \pm 0.04}$	$\mathbf{0.413 \pm 0.08}$	$\mathbf{0.35 \pm 0.07}$
Gain	10.8%	7.4%	6.7%	14.04%	31.0%	29.53%	4.9%	30.28%	37.79%

Table 1 shows that SAG-VICReg achieves the highest LCA similarity across all datasets, with particularly strong gains in Spearman and Kendall correlations, indicating improved preservation of global semantic structure on unseen data. Rank correlations are especially informative for hierarchy because they measure preserved ordering rather than absolute distances. Cophenetic similarity results in Appendix B show the same trend.

Validating the metric. To validate our metric as a meaningful label-free signal, we show that models with higher agreement under our metric also achieve higher hierarchical k-NN accuracy. Across 18 diverse CIFAR-100 models, the correlation reaches $\rho = 0.73$ (Levels 1–2) and $\rho = 0.76$ (Level 4), above RankMe Garrido et al. (2023) (0.47, 0.49), α-ReQ Agrawal et al. (2022) (0.42, 0.36), and linear CKA Kornblith et al. (2019) (0.20, 0.13). Using raw cosine distances instead drops it to 0.10 and -0.02 (Appendix B.1).

Table 2: Hierarchical k-NN classification accuracy (%) on CIFAR-100.

(a) **Pretraining on CIFAR-100**

Model	**Level 1**	**Level 2**	**Level 3**	**Level 4**
SimCLR	90.34 ± 0.81	77.63 ± 0.84	61.25 ± 0.64	45.81 ± 0.52
BarlowTwins	92.87 ± 0.66	82.14 ± 0.57	71.34 ± 0.63	58.17 ± 0.37
VICReg	92.62 ± 0.77	83.08 ± 0.59	71.21 ± 0.43	57.99 ± 0.48
SAG-VICReg (Ours)	$\mathbf{93.89 \pm 0.52}$	$\mathbf{84.89 \pm 0.70}$	$\mathbf{73.18 \pm 0.65}$	$\mathbf{60.01 \pm 0.57}$

(b) **Pretraining on ImageNet-1k**

Level 1	**Level 2**	**Level 3**	**Level 4**
88.17 ± 0.53	79.85 ± 0.38	65.46 ± 0.59	47.21 ± 0.44
90.99 ± 0.41	82.54 ± 0.24	70.68 ± 0.41	55.56 ± 0.19
91.32 ± 0.53	82.61 ± 0.38	70.65 ± 0.21	$\mathbf{55.95 \pm 0.33}$
$\mathbf{91.68 \pm 0.37}$	$\mathbf{83.08 \pm 0.46}$	$\mathbf{70.79 \pm 0.25}$	55.82 ± 0.23

Tables 2 and 3 report complementary evaluations. Table 2 shows that SAG-VICReg improves hierarchical k-NN accuracy across the hierarchy. When pretrained on CIFAR-100, it achieves the best result at every level, including a clear gain at the finest level over VICReg. When pretrained on ImageNet-1k, it remains strongest at the coarser levels and performs on par with the best methods at the finer levels, indicating robust transfer of the learned structure. Table 3 further shows that these

Table 3: ImageNet-1k: linear, semi-supervised (1%, 10%), and k-NN accuracies (%).

Method	Linear		Semi-supervised (1%)		Semi-supervised (10%)		k-NN
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	
SimCLR	68.13 ± 0.54	88.74 ± 0.41	44.17 ± 0.59	72.28 ± 0.31	61.81 ± 0.66	84.79 ± 0.37	54.22 ± 0.39
Barlow Twins	71.58 ± 0.41	$\underline{90.43 \pm 0.22}$	$\underline{50.64 \pm 0.27}$	74.99 ± 0.48	$\mathbf{65.81 \pm 0.52}$	86.66 ± 0.26	59.28 ± 0.31
VICReg	$\underline{71.74 \pm 0.29}$	90.39 ± 0.33	50.49 ± 0.21	$\underline{75.59 \pm 0.29}$	65.18 ± 0.39	$\underline{86.79 \pm 0.41}$	$\underline{59.41 \pm 0.46}$
SAG-VICReg (Ours)	$\mathbf{72.46 \pm 0.36}$	$\mathbf{90.97 \pm 0.45}$	$\mathbf{50.86 \pm 0.29}$	$\mathbf{76.18 \pm 0.32}$	$\underline{65.62 \pm 0.34}$	$\mathbf{87.09 \pm 0.45}$	$\mathbf{60.15 \pm 0.53}$

gains do not come at the expense of standard ImageNet-1k evaluation: SAG-VICReg achieves the best or competitive performance across linear, semi-supervised, and k-NN protocols. Due to computational constraints, all ImageNet models in Table 3 are trained for 250 epochs rather than the 1000-epoch setup commonly used. Together, these results indicate that improving global structure preservation also maintains strong local discriminative performance. Additional results on Caltech-256 and hierarchical linear classification are provided in Appendix B.

For out-of-distribution transfer, Table 4 evaluates VOC07 Everingham et al. (2010) linear classification and VOC07+12 Everingham et al. (2010) object detection. SAG-VICReg consistently improves over VICReg for both source datasets, with particularly clear gains on VOC07+12 object detection. This suggests that the richer structure learned by our method transfers beyond classification and benefits localization-heavy downstream tasks. Additional cross-dataset transfer results across five diverse classification datasets are provided in Appendix B.3.

Table 4: OOD transfer: VOC07 linear classification and VOC07+12 object detection.

Pretrained on	Method	VOC07	VOC07+12
Caltech-256	VICReg	66.67 ± 0.41	49.73 ± 0.77
	SAG-VICReg (Ours)	$\mathbf{67.31 \pm 0.61}$	$\mathbf{51.07 \pm 0.56}$
ImageNet-1k	VICReg	82.85 ± 0.36	73.89 ± 0.24
	SAG-VICReg (Ours)	$\mathbf{83.49 \pm 0.28}$	$\mathbf{74.22 \pm 0.43}$

We next test whether random walk pairing can improve already-pretrained representations while preserving the host method's native objective. We fine-tune DINOv2, TIPSv2 Cao et al. (2026), and CAPI Darcet et al. (2025), each with its own objective, comparing the vanilla fine-tune with an otherwise identical version using random walk pairing. We evaluate different model sizes across several CIFAR-100 data fractions and on ImageNet-100.

Table 5: Percentages denote the fraction of training data used; values report k-NN top-1 gains, in percentage points, over the corresponding vanilla fine-tune.

Backbone	CIFAR-100 (10%)	CIFAR-100 (20%)	CIFAR-100 (100%)	ImageNet-100 (20%)
DINOv2 ViT-S/14	$+0.88$	$+1.16$	$+0.87$	$+0.47$
DINOv2 ViT-B/14	$+1.43$	$+1.18$	$+1.65$	$+0.63$
TIPSv2 ViT-B/14	$+0.43$	$+0.78$	$+0.76$	$+0.45$
TIPSv2 ViT-L/14	$+0.59$	$+0.86$	$+0.57$	$+0.39$
CAPI ViT-L/14	$+0.26$	$+0.57$	$+0.73$	$+0.62$

Table 5 shows that random walk pairing improves all 20 evaluated settings, with gains of up to $+1.65$ points across model sizes, data fractions, datasets, and objectives. These results show that the benefit extends beyond VICReg and training from scratch to the adaptation of pretrained SSL models (more details in Appendix B.8).

These results show that our pairing is not tied to a particular SSL objective or training regime, as the consistent gains across methods, model scales, data regimes, and datasets show.

6 Conclusion

In this work, we addressed the challenge of generalization and global semantic understanding in SSL. Motivated by spectral insights revealing how same-image pairing creates fragmented graph structures, we proposed a random walk-guided training strategy applicable across various SSL methods. Our approach enriches connectivity to capture hierarchical semantic structures while maintaining strong local performance, as demonstrated through consistent improvements in diverse SSL methods. Additionally, we introduced a *label-free* evaluation metric that directly assesses global structure preservation and generalization capabilities, complementing standard local metrics.

AI USE STATEMENT

We used generative AI tools to help draft parts of the paper and to improve the wording, grammar, and clarity of the text. We did not use generative AI for literature search, research ideation, running experiments, generating synthetic data, or proving mathematical claims.

All AI-assisted text was reviewed and edited by the authors. The AI tools were used only for writing assistance; the research ideas, experiments, analyses, mathematical work, and references were developed and carried out by the authors. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our method and experiments throughout the paper and appendices. Algorithms 1 and 2 give pseudocode for random walk pairing and the proposed label-free evaluation. Appendix C reports the random walk, optimization, and sensitivity hyperparameters, including settings for all additional SSL methods. Appendices F, G, and H describe the architectures, dataset splits, and training protocols used in the main experiments. Appendix B.9 specifies the fine-tuning protocol for DINOv2, TIPSv2, and CAPI, while Appendix B.10 details how random walk pairing is incorporated into each host objective. Appendix D reports the hardware and computational overhead. We use publicly available datasets and pretrained checkpoints, and matched baseline and random walk runs use the same experimental settings except for the proposed pairing mechanism.

REFERENCES

Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. \alpha-req : Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17626–17638. Curran Associates, Inc., 2022. doi: 10.52202/068431-1281. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/70596d70542c51c8d9b4e423f4bf2736-Paper-Conference.pdf.

Shaden Alshammari, John Hershey, Axel Feldmann, William Freeman, and Mark Hamilton. A unifying framework for representation learning. In *International Conference on Learning Representations*, volume 2025, pp. 59257–59287, 2025.

Michael R Anderberg. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, volume 19. Academic press, 2014.

Chaitanya Animesh and Manmohan Chandraker. Tuned contrastive learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7225–7234. IEEE, 2025.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Yoshua Bengio, Jean-françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, 16, 2003.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. In *International conference on machine learning*, pp. 3252–3298. PMLR, 2023.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, et al. Tipsv2: Advancing vision-language pretraining with enhanced patch-text alignment. *arXiv preprint arXiv:2604.12012*, 2026.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021. doi: 10.1109/CVPR46437.2021.01549.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9620–9629. IEEE, 2021.

Ronald R. Coifman and Stéphane Lafon. Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions. *Applied and Computational Harmonic Analysis*, 21(1):31–52, 2006. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2005.07.005. URL https://www.sciencedirect.com/science/article/pii/S1063520306000522. Special Issue: Diffusion Maps and Wavelets.

Siran Dai, Qianqian Xu, Peisong Wen, Yang Liu, and Qingming Huang. Exploring structural degradation in dense representations for self-supervised learning. *Advances in Neural Information Processing Systems*, 38:16715–16764, 2026.

Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latent patches for improved masked image modeling. *arXiv preprint arXiv:2502.08769*, 2025.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Petros Drineas and Michael W. Mahoney. On the nystrom method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(72):2153–2175, 2005. URL http://jmlr.org/papers/v6/drineas05a.html.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9568–9577. IEEE, 2021.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html.

Imanol G Estepa, Ignacio Saras'ua, Bhalaji Nagarajan, and P Radeva. All4one: Symbiotic neighbour contrastive learning via self-attention and redundancy reduction. 2023 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pp. 16197–16207, 2023.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004. doi: 10.1109/TPAMI.2004.1262185.

Zheng Gao, Chen Feng, and Ioannis Patras. Self-supervised representation learning with cross-context learning between global and hypercolumn features. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1762–1772. IEEE, 2024.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning, 2024. URL https://arxiv.org/abs/2403.00504.

Chongjian Ge, Jiangliu Wang, Zhan Tong, Shoufa Chen, Yibing Song, and Ping Luo. Soft neighbors are positive supporters in contrastive visual representation learning. *arXiv preprint arXiv:2303.17142*, 2023.

Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Obow: Online bag-of-visual-words generation for self-supervised learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6826–6836, 2021. doi: 10.1109/CVPR46437.2021.00676.

Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. *Advances in neural information processing systems*, 33:9662–9673, 2020.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. *CalTech Report*, 03 2007.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

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.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Çağlar Hızlı, Çağatay Yıldız, and Pekka Marttinen. Object-level self-distillation for vision pretraining. *arXiv preprint arXiv:2506.05409*, 2025.

Dahyun Kang, Piotr Koniusz, Minsu Cho, and Naila Murray. Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19627–19638. IEEE, 2023.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Hyeongji Kim, Pekka Parviainen, Terje Berge, and Ketil Malde. Inspecting class hierarchies in classification-based metric learning models. *arXiv preprint arXiv:2301.11065*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Tim Lebailly, Thomas Stegmüller, Behzad Bozorgtabar, Jean-Philippe Thiran, and Tinne Tuytelaars. Adaptive similarity bootstrapping for self-distillation based representation learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16459–16468. IEEE, 2023.

Tim Lebailly, Thomas Stegmüller, Behzad Bozorgtabar, Jean-Philippe Thiran, and Tinne Tuytelaars. Cribo: Self-supervised learning via cross-image object-level bootstrapping. In *International Conference on Learning Representations*, volume 2024, pp. 46794–46814, 2024.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Shaojie Li, Sheng Ouyang, and Yong Liu. Understanding the generalization performance of spectral clustering algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8614–8621, 2023.

Xiaojie Li, Yibo Yang, Xiangtai Li, Jianlong Wu, Yue Yu, Bernard Ghanem, and Min Zhang. Genview: Enhancing view quality with pretrained generative model for self-supervised learning. In *European Conference on Computer Vision*, pp. 306–325. Springer, 2024.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Xianzhong Long, Chen Peng, and Yun Li. Mnn: Mixed nearest-neighbors for self-supervised learning. *Pattern Recognition*, 158:110998, 2025.

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. Tips: Text-image pretraining with spatial awareness. In *International Conference on Learning Representations*, volume 2025, pp. 64256–64279, 2025.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Louis L McQuitty. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *Educational and psychological measurement*, 17(2):207–229, 1957.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=a68SUt6zFt`.

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.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, 2017. doi: 10.1109/CVPR.2017.690.

Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Rohan Sharma, Kaiyi Ji, Zhiqiang Xu, and Changyou Chen. Auc-cl: A batchsize-robust framework for self-supervised contrastive representation learning. In *International Conference on Learning Representations*, volume 2024, pp. 7010–7030, 2024.

Carlos N Silla Jr and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22(1):31–72, 2011.

Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5461–5471. IEEE, 2023.

Robert R Sokal and F James Rohlf. The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40, 1962.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua Susskind, and Etai Littwin. Lidar: Sensing linear probing performance in joint embedding ssl architectures. In *International Conference on Learning Representations*, volume 2024, pp. 56726–56765, 2024.

Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c2e5eaae9152079b9e95845750bb9ab-Paper.pdf.

Jack Valmadre. Hierarchical classification at multiple operating points. *Advances in neural information processing systems*, 35:18034–18045, 2022.

Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2122–2131. IEEE, 2023.

Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5075–5084. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/wehrmann18a.html.

Peisong Wen, Qianqian Xu, Siran Dai, Runmin Cong, and Qingming Huang. Semantic concentration for self-supervised dense representations learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2740–2748, 2015. doi: 10.1109/ICCV.2015.314.

Xiaoyu Yang, Lijian Xu, Hongsheng Li, and Shaoting Zhang. One leaf reveals the season: Occlusion-based contrastive learning with semantic-aware views for efficient visual representation, 2025. URL https://arxiv.org/abs/2411.09858.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. *arXiv preprint arXiv:2203.14415*, 2022.

A Theoretical Motivation for Our Approach

Our random walk-guided pairing strategy builds on recent theoretical results that highlight the centrality of spectral analysis for self-supervised embedding and generalization.

First, HaoChen et al. (2021) show that in a contrastive learning framework–modeled by a so-called *augmentation graph*, whose nodes represent augmented data points, richer or denser edges yield smoother eigenfunctions and more robust embeddings. Concretely, adding edges for augmentations that capture genuine relationships in data improves stability for points outside of the original training clusters. This insight directly underpins our random walk sampling strategy, where we deliberately introduce additional "positive" edges to foster cross-image relationships beyond strict instance-level augmentations. The effect is to densify the augmentation graph, thereby improving spectral regularization and leading to more stable embeddings in unseen regions.

Second, Li et al. (2023) provide *excess risk bounds* for spectral clustering and analyze how one can extend the learned embedding map to newly arrived samples. Their analysis confirms that spectral methods can reliably embed unseen points from the same manifold, provided that the embeddings are anchored in sufficiently connected local structures. In our setting, ensuring that edge creation (i.e., sampling cross-image pairs) captures correct semantic neighborhoods helps reduce distortion when new data, possibly from new clusters, arrives at test time. The random walk procedure thus strengthens these local-to-global relationships and mitigates the generalization gap.

Finally, Cabannes et al. (2023) emphasize that creating nontrivial positive pairs (i.e., carefully chosen augmentations rather than trivial identity mappings) reduces the variance of the learned embedding and, more importantly, boosts generalization. Building on this perspective, our approach employs a *weighted random walk* over the batch's similarity graph to select pairs that share broader semantic similarity (e.g., partial overlap or common neighborhood). Essentially, by sampling from a denser subgraph under an appropriate random walk transition, we encourage models to recognize higher-level or cross-image relationships, consistent with the theoretical prediction that such "denser edges" yield better accuracy on downstream tasks. Hence, our design can be viewed as a principled approach toward bridging the gap between purely local augmentation invariance and capturing global manifold structure in learned representations, applicable across diverse SSL frameworks that rely on positive pair construction.

B Additional results

B.1 Validation of the Label-Free Metric

At a high level, we ask whether agreement in the hierarchical organization learned by different models can serve as a label-free signal of representation quality. If so, models that recover similar meaningful structure should receive higher consensus scores, and these scores should track downstream hierarchical accuracy without using labels.

To test this, we evaluate the proposed metric across a diverse collection of pretrained representations and compare it with both label-aware hierarchy measures and existing label-free representation-quality metrics.

Setup. We evaluate 18 checkpoints spanning supervised, self-supervised, and image–text pretraining: supervised ResNet-18 and ResNet-50 He et al. (2016), ViT-B/16 Dosovitskiy et al. (2020) and ConvNeXt-T Liu et al. (2022); DINO ViT-S/16 Caron et al. (2021), DINOv2 ViT-S/14, ViT-B/14 and ViT-L/14 Oquab et al. (2024), ResNet-50 models trained with Barlow Twins Zbontar et al. (2021), SwAV Caron et al. (2020) and VICReg Bardes et al. (2022), and MAE ViT-B/16 He et al. (2022); CLIP ViT-B/32 Radford et al. (2021), SigLIP ViT-B/16 Zhai et al. (2023) and EVA-02 ViT-B Fang et al. (2024); a randomly initialized ResNet-50; and our two split-half CIFAR-100 models from Section 5. All models embed the same 5,000 CIFAR-100 test images, and all scores below are computed from ℓ_2-normalized embeddings.

As a label-based reference, we use hierarchical k-NN accuracy (k=20, cosine), averaged over the two coarse levels (L1–L2) and separately at the finest level (L4). We define the CIFAR-100 hierarchical distance between two images as 0 if they share the same finest-grained class, 1 if they share

Table 6: Spearman correlation between representation-quality scores and hierarchical k-NN accuracy across 18 checkpoints. Label-free methods are shown below the separator; best label-free results are bolded.

Score	Labels	ρ (L1–L2)	ρ (L4)
Tree vs. hierarchy (cophenetic)	yes	$+0.84$	$+0.85$
Tree vs. hierarchy (LCA distance)	yes	$+0.81$	$+0.81$
Raw distances vs. hierarchy (no tree)	yes	$+0.34$	$+0.29$
Consensus, our pair metric	**no**	$\mathbf{+0.73}$	$\mathbf{+0.76}$
Consensus, raw distances (no tree)	no	$+0.10$	-0.02
Consensus, linear CKA	no	$+0.20$	$+0.13$
RankMe	no	$+0.47$	$+0.49$
α-ReQ	no	$+0.42$	$+0.36$

Level 3 but not Level 4, 2 if they share Level 2 but not Level 3, 3 if they share Level 1 but not Level 2, and 4 if they do not share the same Level 1 class.

For each model, we construct a Ward dendrogram from the pairwise cosine distances of its embeddings. This associates every image pair with both a cophenetic distance and an LCA distance. We measure alignment with the CIFAR-100 hierarchy using the Spearman correlation between these tree-based distances and the corresponding hierarchical distances. As a no-tree reference, we perform the same comparison directly using cosine distances.

For the fully label-free evaluation, we compare our metric with RankMe Garrido et al. (2023), α-ReQ Agrawal et al. (2022), and linear CKA Kornblith et al. (2019). For every pair of models (m, m'), our pairwise metric from Section 4.4 is the Spearman correlation between their cophenetic-distance vectors, measuring how similarly the two models organize the same images hierarchically. Each model's *consensus* score is its average agreement with the remaining 17 models. We construct analogous consensus scores from raw cosine distances and linear CKA, while RankMe and α-ReQ are computed independently for each model.

Results. Table 6 shows that our structural metric is a strong label-free indicator of hierarchy-aware representation quality. Its consensus score reaches $\rho = 0.73$ at the coarse levels and $\rho = 0.76$ at the finest level, substantially exceeding all other label-free measures. The next-best baseline, RankMe, reaches only 0.47 and 0.49, while CKA consensus reaches 0.20 and 0.13.

The hierarchical tree is central to this result. Replacing our tree-based comparison with raw cosine distances reduces the label-free correlation from 0.73 to 0.10 at L1–L2 and from 0.76 to -0.02 at L4. The same effect appears in the label-aware comparison with the CIFAR-100 hierarchy: tree-based alignment reaches 0.84–0.85, compared with only 0.29–0.34 for raw distances.

Notably, despite using no labels, our consensus score approaches the correlation obtained when the learned tree is compared directly with the ground-truth CIFAR-100 hierarchy. This suggests that agreement between independently learned representation hierarchies contains substantial information about downstream hierarchical quality. Encouragingly, the ranking is qualitatively meaningful: stronger pretrained models such as DINOv2 and supervised ViT-B receive the highest scores, while MAE and the randomly initialized network appear near the bottom. This suggests that the metric assigns high consensus for a reason, it tracks meaningful representation quality rather than arbitrary structural agreement.

Across all 153 model pairs, our pairwise metric also correlates with linear CKA ($\rho = 0.73$), showing agreement with a standard representation-similarity measure while capturing additional hierarchical structure. We therefore use the metric as a label-free model-level measure of representation structure, rather than as an epoch-wise training diagnostic.

B.2 Additional classification results

In addition to the main-paper results, Table 7 reports hierarchical linear classification on CIFAR-100. This complements the hierarchical *k*-NN results in the main text: when pretrained on CIFAR-

Table 7: Hierarchical linear classification accuracy (%) on CIFAR-100.

(a) **Pretraining on CIFAR-100**

Model	Level 1	Level 2	Level 3	Level 4
SimCLR	91.96 ± 0.74	82.88 ± 0.85	67.70 ± 0.69	56.27 ± 0.92
BarlowTwins	93.26 ± 0.82	85.38 ± 0.91	74.49 ± 0.79	65.03 ± 0.88
VICReg	93.36 ± 0.58	85.61 ± 0.69	74.77 ± 0.64	64.95 ± 0.80
SAG-VICReg (Ours)	$\mathbf{94.25 \pm 0.66}$	$\mathbf{86.94 \pm 0.71}$	$\mathbf{76.56 \pm 0.83}$	$\mathbf{66.45 \pm 0.67}$

(b) **Pretraining on ImageNet-1k**

Level 1	Level 2	Level 3	Level 4
90.56 ± 0.51	81.83 ± 0.62	71.43 ± 0.48	61.95 ± 0.51
93.42 ± 0.41	86.99 ± 0.43	78.44 ± 0.22	71.32 ± 0.28
93.72 ± 0.25	87.26 ± 0.48	78.59 ± 0.36	$\mathbf{71.83 \pm 0.66}$
$\mathbf{94.11 \pm 0.31}$	$\mathbf{87.66 \pm 0.53}$	$\mathbf{79.12 \pm 0.28}$	71.58 ± 0.42

Table 8: Caltech-256: linear, semi-supervised (1%, 10%), and k-NN accuracies (%).

Method	Linear		Semi-supervised (1%)		Semi-supervised (10%)		k-NN
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	
SimCLR	73.48 ± 0.52	89.24 ± 0.54	26.49 ± 0.35	41.98 ± 0.48	67.51 ± 0.32	81.85 ± 0.54	62.18 ± 0.46
Barlow Twins	76.25 ± 0.43	91.14 ± 0.49	29.49 ± 0.29	43.01 ± 0.34	67.35 ± 0.37	83.44 ± 0.39	65.41 ± 0.47
VICReg	76.40 ± 0.39	91.31 ± 0.51	$\mathbf{30.37 \pm 0.22}$	43.51 ± 0.31	67.17 ± 0.42	83.40 ± 0.54	65.34 ± 0.38
SAG-VICReg (Ours)	$\mathbf{77.33 \pm 0.34}$	$\mathbf{91.83 \pm 0.45}$	30.19 ± 0.26	$\mathbf{43.66 \pm 0.41}$	$\mathbf{67.86 \pm 0.36}$	$\mathbf{84.18 \pm 0.43}$	$\mathbf{66.17 \pm 0.34}$

100, SAG-VICReg improves over all baselines across all hierarchy levels, showing gains in both coarse and fine-grained semantic discrimination. When pretrained on ImageNet-1k, SAG-VICReg achieves the best results at the coarser levels and remains competitive at the finest levels, indicating that the learned structure transfers across datasets. Table 8 further reports standard Caltech-256 evaluation on finest-grained labels, where SAG-VICReg is again superior or competitive across linear, semi-supervised, and k-NN protocols. Overall, these results show that the improvements in global structure preservation do not sacrifice local classification accuracy, and mostly improve it.

B.3 Additional Cross-Dataset Transfer Evaluation

We further evaluate cross-dataset transfer by comparing VICReg and SAG-VICReg pretrained on either Caltech-256 or ImageNet-1k. We freeze the learned representations and train linear classifiers on five downstream datasets: Food-101 Bossard et al. (2014), SUN397 Xiao et al. (2010), Stanford Cars Krause et al. (2013), Stanford Dogs Khosla et al. (2011), and EuroSAT Helber et al. (2019). These datasets cover diverse domains, including scene recognition, fine-grained classification, and remote-sensing imagery.

Table 9: Linear classification accuracy (%) on five downstream datasets using representations pretrained on Caltech-256 or ImageNet-1k. Best results within each pretraining setting are bolded.

Pretrained on	Method	Food-101	SUN397	Cars	Dogs	EuroSAT
Caltech-256	VICReg	52.39 ± 0.28	46.36 ± 0.61	$\mathbf{42.30 \pm 0.41}$	38.79 ± 0.19	94.23 ± 0.55
	SAG-VICReg	$\mathbf{52.86 \pm 0.46}$	$\mathbf{46.91 \pm 0.69}$	42.25 ± 0.18	$\mathbf{39.37 \pm 0.34}$	$\mathbf{94.66 \pm 0.64}$
ImageNet-1k	VICReg	71.11 ± 0.81	61.28 ± 0.33	46.46 ± 0.93	$\mathbf{71.90 \pm 0.17}$	95.81 ± 0.72
	SAG-VICReg	$\mathbf{71.51 \pm 0.38}$	$\mathbf{61.75 \pm 0.55}$	$\mathbf{47.61 \pm 0.22}$	71.82 ± 0.49	$\mathbf{95.99 \pm 0.29}$

Table 9 shows consistent transfer gains across both pretraining settings. SAG-VICReg improves four of the five downstream datasets when pretrained on Caltech-256 and four of five when pretrained on ImageNet-1k, yielding gains in eight of the ten transfer settings overall. The two remaining differences are small: 0.05 points on Stanford Cars after Caltech-256 pretraining and 0.08 points on Stanford Dogs after ImageNet-1k pretraining.

The improvements span different downstream domains, with the largest gain on Stanford Cars after ImageNet-1k pretraining. Together with the VOC results in the main paper, these results indicate that the richer structure encouraged by random walk pairing transfers beyond the source dataset without sacrificing general downstream performance.

B.4 Cophenetic similarity results

Table 10 presents cophenetic similarity results, which evaluate how well pairwise distances correlate with dendrogram-based cophenetic distances in both directions: D1 to P2 (correlating the second model's pairwise distances with the first model's cophenetic distances) and D2 to P1 (vice versa). SAG-VICReg achieves the highest scores across all datasets, with particularly strong gains on CIFAR-100 and Caltech-256. The bidirectional evaluation confirms that our method preserves

Table 10: Cophenetic similarity results. **D1 to P2** correlates second model's pairwise distances with first model's cophenetic distances, and vice versa for **D2 to P1**. **Average** shows the average of both directions. Relative gains over the second-best are shown.

	ImageNet 1k			CIFAR-100			Caltech-256		
	D1 to P2	**D2 to P1**	**Average**	**D1 to P2**	**D2 to P1**	**Average**	**D1 to P2**	**D2 to P1**	**Average**
SimCLR	$\underline{0.289 \pm 0.03}$	$\mathbf{0.259 \pm 0.06}$	$\underline{0.274}$	$\underline{0.349 \pm 0.08}$	$\underline{0.341 \pm 0.09}$	$\underline{0.345}$	0.305 ± 0.06	0.322 ± 0.09	0.314
Barlow Twins	0.199 ± 0.01	0.156 ± 0.01	0.178	0.309 ± 0.004	0.286 ± 0.01	0.298	$\underline{0.341 \pm 0.01}$	$\underline{0.364 \pm 0.03}$	$\underline{0.353}$
VICReg	0.263 ± 0.03	0.213 ± 0.01	0.238	0.287 ± 0.006	0.277 ± 0.01	0.282	0.342 ± 0.02	0.335 ± 0.01	0.339
SAG-VICReg (Ours)	$\mathbf{0.308 \pm 0.03}$	$\underline{0.246 \pm 0.02}$	$\mathbf{0.277}$	$\mathbf{0.405 \pm 0.02}$	$\mathbf{0.384 \pm 0.03}$	$\mathbf{0.395}$	$\mathbf{0.382 \pm 0.01}$	$\mathbf{0.375 \pm 0.03}$	$\mathbf{0.379}$
Gain	6.57%	-5.0%	1.09%	16.0%	12.61%	14.49%	11.7%	3.02%	7.37%

both local structure (captured by pairwise distances) and global hierarchical organization (captured by cophenetic distances) during generalization. These results complement the LCA similarity findings in the main paper, providing converging evidence that our random walk-guided approach effectively maintains multi-scale semantic relationships when encountering unseen data, a fundamental requirement for robust visual representations.

B.5 Comparison against other cross-image methods

Table 11: Hierarchical k-NN classification accuracy (%) on CIFAR-100. Best results are bolded and second-best results are underlined.

Method	Level 1	Level 2	Level 3	Level 4
NNCLR	$\underline{91.87 \pm 0.48}$	$\underline{82.38 \pm 0.52}$	$\underline{68.82 \pm 0.35}$	$\underline{53.54 \pm 0.41}$
I-Con	89.12 ± 0.27	76.83 ± 0.33	63.81 ± 0.66	48.11 ± 0.51
SAG-VICReg (Ours)	$\mathbf{93.89 \pm 0.52}$	$\mathbf{84.89 \pm 0.70}$	$\mathbf{73.18 \pm 0.65}$	$\mathbf{60.01 \pm 0.57}$

Table 12: Hierarchical linear classification accuracy (%) on CIFAR-100. Best results are bolded and second-best results are underlined.

Method	Level 1	Level 2	Level 3	Level 4
NNCLR	92.30 ± 0.56	84.13 ± 0.31	70.72 ± 0.68	59.74 ± 0.26
I-Con	$\underline{93.60 \pm 0.48}$	$\underline{85.99 \pm 0.27}$	$\underline{73.27 \pm 0.51}$	$\underline{62.23 \pm 0.19}$
SAG-VICReg (Ours)	$\mathbf{94.25 \pm 0.66}$	$\mathbf{86.94 \pm 0.71}$	$\mathbf{76.56 \pm 0.83}$	$\mathbf{66.45 \pm 0.67}$

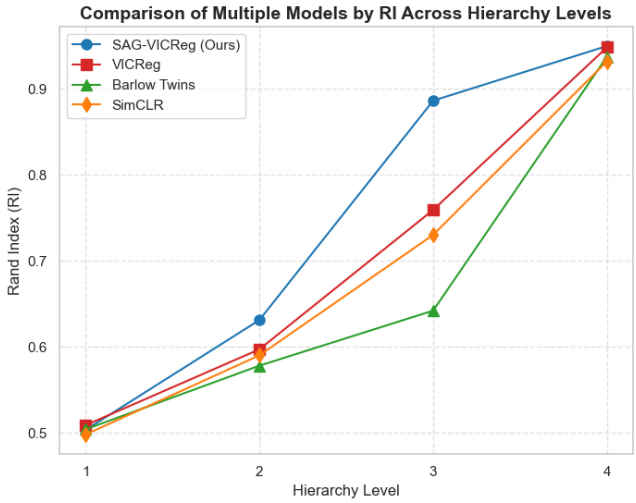
Tables 11 and 12 compare SAG-VICReg directly with representative cross-image SSL methods. SAG-VICReg outperforms both NNCLR Dwibedi et al. (2021) and I-Con Alshammari et al. (2025) across all hierarchy levels and under both k-NN and linear evaluation, with particularly large gains at the finer levels. This suggests that our cross-image pairing strategy improves not only coarse semantic grouping, but also fine-grained hierarchical discrimination, while avoiding NNCLR's memory-bank requirement.

These methods add cross-image relations in different ways. NNCLR Dwibedi et al. (2021) uses the nearest neighbor of an image in a queue of past embeddings as its positive, and soft-neighbor methods Ge et al. (2023) weight several neighbors from a similar support set. AdaSim Lebailly et al. (2023) builds self-distillation targets from neighbors in a queue, and adapts how much it relies on them to the estimated quality of the embedding. MNN Long et al. (2025) mixes the features of neighbors. Mugs Zhou et al. (2022), All4One Estepa et al. (2023) and CrIBo Lebailly et al. (2024) add objectives, heads, attention modules or object-level bootstrapping. I-Con Alshammari et al. (2025) matches a learned neighbor distribution to a supervisory one with a KL divergence, and widens the supervisory distribution by debiasing or neighbor propagation. Our pairing needs none of these parts. The graph is built online from the current batch and model, a partner is sampled from it, and the pair is passed to the host method's own loss, with no memory bank, extra loss, extra head or change to the architecture.

B.6 Rand Index Results

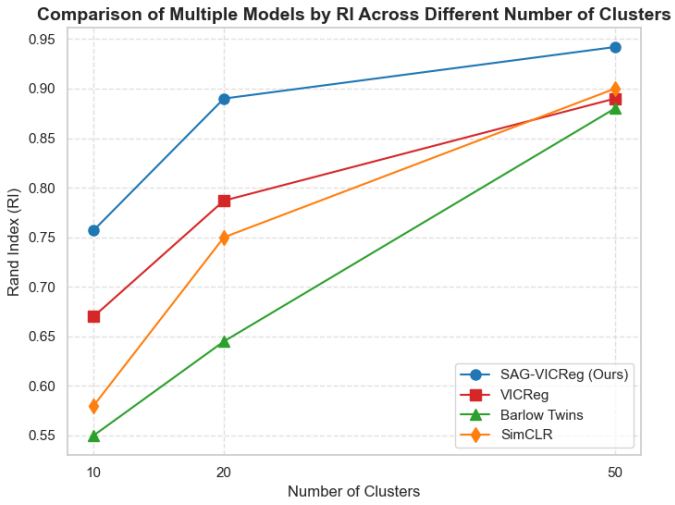
The setup is described in Appendices J.4 and K. The Rand-Index evaluation in Figure 6 shows that SAG-VICReg consistently achieves the highest scores across all CIFAR-100 hierarchy levels, with especially pronounced advantages at the more abstract levels (2, 3). This indicates that our method

Figure 6: Rand-Index scores across CIFAR-100 hierarchy levels.



not only distinguishes between fine-grained categories but also correctly organizes these categories into their broader semantic relationships, reflecting the natural hierarchical organization present in the data.

Figure 7: Comparison of Rand-Index scores across number of clusters on finest-grained CIFAR-100.



Building on these results, Figure 7 further demonstrates SAG-VICReg's robustness by evaluating performance across varying cluster counts (10, 20, and 50) on CIFAR-100's finest-grained labels. Our method consistently achieves the highest Rand Index scores across all tested cluster numbers, maintaining a clear performance advantage over other spectral embedding-based methods.

Notably, while other methods exhibit significant performance fluctuations with changing cluster granularity, SAG-VICReg maintains stable superior performance regardless of clustering resolution. This robustness confirms that our approach captures meaningful semantic structure that generalizes across different levels of clustering granularity–a crucial property for real-world applications where the optimal number of semantic groups may be unknown or vary across different domains and tasks.

Table 13: Hierarchical linear classification accuracy (%) on CIFAR-100 with random walk enhancement (ResNet-34, 500 epochs, pretrained on CIFAR-100).

(a) SimCLR

Method	Level 1	Level 2	Level 3	Level 4
Vanilla	91.96 ± 0.74	82.88 ± 0.85	67.70 ± 0.69	56.27 ± 0.92
+ Random walk	$\mathbf{93.06 \pm 0.92}$	$\mathbf{84.93 \pm 0.81}$	$\mathbf{70.77 \pm 0.46}$	$\mathbf{58.92 \pm 0.73}$

(b) TCL

Method	Level 1	Level 2	Level 3	Level 4
Vanilla	92.87 ± 0.65	84.21 ± 0.71	71.91 ± 0.48	62.11 ± 0.66
+ Random walk	$\mathbf{93.31 \pm 0.66}$	$\mathbf{85.07 \pm 0.86}$	$\mathbf{72.84 \pm 0.91}$	$\mathbf{62.77 \pm 0.74}$

B.7 Additional results using other methods

Table 13 validates the broad applicability of our random walk enhancement across different SSL methods. SimCLR shows substantial gains with random walk pairing, particularly at middle hierarchy levels. These improvements enable the 2020-era SimCLR to match contemporary state-of-the-art performance, as demonstrated in Figures 1, 5. TCL, already state-of-the-art, also benefits from consistent improvements across all hierarchy levels. While more modest than SimCLR's gains–expected given TCL's stronger baseline–the consistent enhancements confirm that our approach benefits even leading methods.

B.8 Additional Fine-Tuning Results

DINOv2 beyond k-NN. We evaluate the DINOv2 ViT-S/14 model fine-tuned on 20% of CIFAR-100 with several other protocols.

Table 14: Additional evaluation of DINOv2 ViT-S/14 fine-tuned on 20% of CIFAR-100 (accuracy, %). Δ: gain of random walk pairing over the vanilla fine-tune.

Evaluation	Vanilla	Random walk	Δ
Hierarchical k-NN, Level 1	97.57	97.83	$+0.27$
Hierarchical k-NN, Level 2	93.83	94.27	$+0.44$
Hierarchical k-NN, Level 3	88.97	89.75	$+0.77$
Hierarchical k-NN, Level 4	80.12	81.28	$+1.16$
Linear probe (100% labels)	85.34	85.60	$+0.26$
Linear probe (10% labels)	78.79	79.17	$+0.38$
Linear probe (1% labels)	60.80	62.52	$+1.72$
Transfer: CIFAR-10 k-NN	94.26	94.58	$+0.32$

Random walk pairing improves every protocol in Table 14: hierarchical k-NN at all four levels, linear probes with 100%, 10% and 1% of the labels, and k-NN transfer to CIFAR-10. The gain is largest with 1% of the labels, so the improved representation helps most when labels are scarce.

TIPSv2 and CAPI. Table 5 also includes two recent models trained with different objectives: the TIPSv2 image encoder Cao et al. (2026), evaluated at two model sizes, and CAPI Darcet et al. (2025), each fine-tuned with its own native objective (Appendix B.9).

Random walk pairing improves all 12 evaluated TIPSv2 and CAPI settings. For TIPSv2 ViT-B/14, the gains are $+0.43$, $+0.78$, and $+0.76$ points with 10%, 20%, and 100% of CIFAR-100, respectively, and $+0.45$ on ImageNet-100. The corresponding gains for TIPSv2 ViT-L/14 are $+0.59$, $+0.86$, $+0.57$, and $+0.39$. CAPI ViT-L/14 similarly improves by $+0.26$, $+0.57$, and $+0.73$ points across the three CIFAR-100 fractions, and by $+0.62$ on ImageNet-100. Thus, the improvement persists across objectives, model scales, data regimes, and datasets, with no observed regression.

On the 20% CIFAR-100 setting, the gains also carry over beyond k-NN: hierarchical k-NN improves significantly at all four levels for TIPSv2 ViT-B/14 and at the two finest levels for CAPI, while linear probes with 1% of the labels improve by 2.19 and 1.77 points, respectively.

B.9 Fine-tuning protocol

We fine-tune DINOv2 ViT-S/14 and ViT-B/14 Oquab et al. (2024), the TIPSv2 ViT-B/14 and ViT-L/14 image encoders Cao et al. (2026), and CAPI ViT-L/14 Darcet et al. (2025), each from the authors' released checkpoint. Each model keeps its own self-supervised objective: DINO+iBOT for DINOv2; the self-distillation and masked image modeling losses of the TIPS image encoder Maninis et al. (2025), without the text losses; and, for CAPI, the prediction of the teacher's cluster assignments for masked patches. The objective heads are re-initialized and all inputs are 112×112. The vanilla and random walk runs of a setting share all other settings and seeds, and differ only in the random walk pairing (Appendix B.10). For TIPS and CAPI, the number of epochs of each setting is chosen on the vanilla runs alone and then used for both arms. Hyperparameters are given in Appendix C.

B.10 Random walk pairing in other SSL methods

In every method, the walk is computed on the current batch as in Section 4: an affinity matrix W between the embeddings of the two views, the transition matrix $P = D^{-1}W$, and one partner $j \sim P_{i,\cdot}$ sampled for every image i. The partner can be the image itself. The partner is used for the positive of image i, as follows.

VICReg. The second view of image i is replaced by the second view of its partner: the invariance term is computed between z_i and $z''_i = z'_j$, weighted by W_{ij} (Algorithm 1). The variance and covariance terms are unchanged.

SimCLR. We apply the SimCLR loss, in both directions, to the batch of pairs (Z, Z'') instead of (Z, Z'), where $Z'' = \{z''_1, \ldots, z''_n\}$ holds the partners' views, $z''_i = z'_j$. For an anchor z_i, the positive is z''_i and the candidates are $(Z \backslash \{z_i\}) \cup Z''$; the reverse direction is symmetric. We keep the standard denominator: the partner's other view z_j and repeated partners stay among the $2n - 1$ candidates, and when every partner is the image itself this is the standard SimCLR loss.

TCL. In the same way, we apply the TCL loss, with its tuned denominator, to the pairs (Z, Z''): the positive of z_i is z''_i, and all other $2n - 1$ embeddings of $Z \cup Z''$ act as negatives.

In both SimCLR and TCL, the partner positive is used after an initial warm-up period.

DINOv2 and TIPS. Each student crop of image i is trained with a cross-entropy to the teacher's centered and sharpened output on the other global view of image i (global crops are paired crosswise; local crops use both global views). We replace this target by a weighted pair of targets: the teacher output on the corresponding global view of image i, with weight 1, and of its partner j, with weight w. The partner is sampled from the teacher's CLS features of the two global views, and the masked image modeling (iBOT Zhou et al. (2021)) term is unchanged.

CAPI. CAPI trains the student to predict, for every masked patch, the cluster that the teacher assigns to it by Sinkhorn–Knopp, and has no image-level target. The walk partner provides one: the student's predictions at the masked patches of image i are averaged and passed through the student head, and this image-level prediction is trained, with weight w, to match the partner's cluster histogram, i.e., the teacher's cluster assignments on the partner's second view averaged over its patches. The partner is sampled from the teacher's global representations of the two views, and the patch-level objective is unchanged.

For the fine-tuned models, w is switched on after one sixth of training and increased linearly to its final value over the next third (Appendix C).

C Analysis and Hyperparameters

C.1 Hyperparameters Overview

Beyond the standard VICReg hyperparameters ($\lambda = 25$, $\mu = 25$, $\nu = 1$), which we maintain unchanged, our approach introduces two additional parameters: k (number of neighbors for constructing the affinity matrix) and the local scaling factor for the Gaussian kernel (computed as a percentile of the distance distribution). Through systematic evaluation, we identified $k = 5$ neighbors and the 20th percentile for local scaling as effective default values.

Our ablation studies reveal that while performance varies across different hyperparameter configurations, the method demonstrates reasonable robustness within practical ranges. SAG-VICReg consistently outperforms the baseline across most tested configurations, though optimal performance requires appropriate hyperparameter selection. The following subsections present detailed ablations examining core design choices, sensitivity to key hyperparameters, and batch size effects, providing guidance for practitioners on configuration selection and expected performance trade-offs.

Experimental setup. Unless otherwise specified, all ablation experiments in this section use *k*-NN classification on finest-grained labels for evaluation.

C.2 Core Design Ablations

Table 15 examines two fundamental design choices: how to weight the invariance term and how many steps to take in the random walk process.

Table 15: Ablation of invariance term weighting and random walk length (fine-grained *k*-NN accuracy, %).

Configuration	CIFAR-100	ImageNet-1k	Caltech-256
Invariance Term Weighting			
Uniform weighting	59.21 ± 0.68	59.74 ± 0.53	65.79 ± 0.67
W_{ij}-weighted	$\mathbf{60.01 \pm 0.57}$	$\mathbf{60.15 \pm 0.53}$	$\mathbf{66.17 \pm 0.34}$
Walk Length			
1-step	$\mathbf{60.01 \pm 0.57}$	$\mathbf{60.15 \pm 0.53}$	$\mathbf{66.17 \pm 0.34}$
2-step	59.61 ± 0.83	59.69 ± 0.85	65.24 ± 0.54
3-step	59.36 ± 0.92	59.34 ± 0.69	66.05 ± 0.59

Invariance term weighting. When constructing cross-image pairs, we can either treat all selected pairs equally (uniform weighting, $W_{ij} = 1$ for all pairs) or weight them according to their semantic similarity (W_{ij}-weighted, as described in Section 4). The results demonstrate consistent improvements from similarity-based weighting across all datasets: CIFAR-100 improves by 0.80 percentage points (59.21% $\rightarrow$ 60.01%), ImageNet-1k by 0.41 points (59.74% $\rightarrow$ 60.15%), and Caltech-256 by 0.38 points (65.79% $\rightarrow$ 66.17%). This confirms that emphasizing stronger semantic connections while still allowing weaker relationships to contribute yields better representations than treating all cross-image pairs uniformly.

Random walk length. The walk length determines how we sample cross-image pairs: 1-step means directly selecting from the k nearest neighbors, while multi-step walks traverse multiple transitions through the affinity graph (e.g., 2-step samples from neighbors of neighbors). Results show that 1-step sampling performs best across all datasets, with slight to moderate performance decreases as walk length increases. On CIFAR-100, performance drops from 60.01% (1-step) to 59.61% (2-step) and 59.36% (3-step). Similar patterns emerge on ImageNet-1k (60.15% $\rightarrow$ 59.69% $\rightarrow$ 59.34%) and Caltech-256 (66.17% $\rightarrow$ 65.24% $\rightarrow$ 66.05%). This indicates that direct semantic neighbors provide the most reliable cross-image relationships, while longer walks may introduce pairs that are too semantically distant, potentially diluting the learning signal. Based on these findings, we adopt 1-step sampling with W_{ij}-weighting as our default configuration.

Table 16 isolates the contribution of the random walk pairing strategy. Across both uniform and similarity-weighted invariance losses, the random walk variant outperforms deterministic nearest-neighbor pairing. The best result is obtained when random walk sampling is combined with W_{ij}-

Table 16: Effect of pairing strategy and invariance weighting (fine-grained k-NN accuracy on CIFAR-100, %).

Invariance Weighting	Nearest Neighbor	Non-trivial Nearest Neighbor	Our Random Walk
Uniform	59.13 ± 0.68	57.94 ± 0.88	$\mathbf{59.21 \pm 0.68}$
W_{ij}-weighted	59.42 ± 0.71	59.27 ± 0.63	$\mathbf{60.01 \pm 0.57}$

Non-trivial nearest neighbor selects the closest neighbor in Z' that is not the augmentation counterpart of z_i, i.e., $j \neq i$.

weighted invariance, showing that the gain comes not only from using cross-image positives, but also from propagating through the affinity graph and weighting pairs by their semantic similarity.

Table 17: Sensitivity to hyperparameters. Bold indicates our default configuration.

Configuration	Fine-grained top-1 (%)
Number of Neighbors (k)	
$k = 1$	59.42 ± 0.71
$k = 3$	59.19 ± 0.78
$\mathbf{k = 5}$	$\mathbf{60.01 \pm 0.57}$
$k = 7$	59.28 ± 0.49
$k = 10$	59.09 ± 0.55
Local Scale Percentile	
10th	59.82 ± 0.44
15th	59.76 ± 0.64
20th	$\mathbf{60.01 \pm 0.57}$
25th	59.48 ± 0.73
30th	58.99 ± 0.78

C.2.1 Hyperparameter Sensitivity

Table 17 examines sensitivity to the number of neighbors k and the local scaling percentile used for the Gaussian kernel.

Number of neighbors (k). The results reveal an optimal value at $k = 5$ (60.01%), with performance declining for both smaller and larger neighborhood sizes. Too few neighbors ($k = 1$: 59.42%, $k = 3$: 59.19%) limit the diversity of semantic relationships captured, while too many neighbors ($k = 7$: 59.28%, $k = 10$: 59.09%) likely introduce increasingly distant, less semantically related pairs that dilute the learning signal. The relatively modest performance variation across the range ($k = 3$ to $k = 10$ spans only 0.92 percentage points) suggests reasonable robustness, though the clear peak at $k = 5$ confirms our default choice.

Local scale percentile. The local scaling factor, computed as a percentile of the distance distribution, shows optimal performance at the 20th percentile (60.01%). Lower percentiles (10th: 59.82%, 15th: 59.76%) produce tighter kernels that may be overly restrictive, while higher percentiles (25th: 59.48%, 30th: 58.99%) create broader kernels that may assign excessive weight to distant pairs. The performance remains relatively stable between the 10th and 20th percentiles (0.19-0.25 point differences), but degrades more substantially at the 30th percentile (1.02 point drop), indicating that overly broad kernels are more problematic than overly tight ones. These findings confirm that the 20th percentile provides an effective balance for capturing meaningful semantic neighborhoods while filtering out noise.

C.2.2 Batch Size Sensitivity

Table 18 examines how performance scales with batch size across three datasets with varying numbers of classes. Since our neighbor mining is intra-batch, batch size directly affects the probability of sampling semantically related images.

Table 18: Sensitivity to batch size (fine-grained k-NN accuracy, %). Bold indicates our default configuration.

Configuration	CIFAR-100	ImageNet-1k	Caltech-256
Batch Size			
128	59.88 ± 0.55	58.47 ± 0.75	65.11 ± 0.77
256	$\mathbf{60.01 \pm 0.57}$	59.23 ± 0.77	65.56 ± 0.72
512	58.43 ± 0.72	59.71 ± 0.65	$\mathbf{66.17 \pm 0.34}$
1024	54.18 ± 0.88	$\mathbf{60.15 \pm 0.53}$	65.91 ± 0.66

The results reveal a clear relationship between optimal batch size and the number of dataset classes, driven by **batch semantic density**, the probability of finding true semantic neighbors within a batch. CIFAR-100 (100 classes) achieves peak performance at batch size 256, with significant degradation at larger batches (54.18% at 1024). In contrast, ImageNet-1k (1000 classes) shows monotonic improvement with increasing batch size, reaching optimal performance at 1024 (60.15%). Caltech-256 (256 classes) falls between these extremes, peaking at batch size 512 (66.17%).

This pattern reflects a fundamental trade-off: with fewer classes, smaller batches are sufficient to capture semantic relationships, as the probability of sampling multiple instances from the same or related classes remains high. For example, in CIFAR-100 with batch size 256, we expect approximately 2.56 samples per class on average, providing adequate semantic neighborhood density. However, with 1000 classes in ImageNet-1k, batch size 256 yields only 0.26 samples per class on average, making it difficult to find meaningful cross-image pairs. Larger batches (1024) improve this to roughly 1 sample per class, enabling more effective semantic neighbor discovery.

The degradation at excessively large batches on CIFAR-100 suggests that beyond a certain point, additional batch diversity may introduce training dynamics challenges or reduce the quality of intra-batch relationships. These findings indicate that datasets with more classes benefit from larger batches to maintain sufficient batch semantic density for effective cross-image pairing.

C.2.3 Hyperparameters for Other SSL Methods

Table 19 lists the random walk settings for the other methods; the rest of their training setup is in Appendix H (SimCLR, TCL) and Appendix B.9 (fine-tuned models). As with VICReg, 1-step walks work best for every method. SimCLR and TCL have no separately weighted invariance term, so we do not use the W_{ij} weighting for them.

Table 19: Random walk and batch settings for the other methods.

Method	Neighbors k	Scale percentile	Batch size
SimCLR	5	20th	512
TCL	5	15th	256 (CIFAR-100), 512 (Caltech-256)
DINOv2	4	20th	256 (ViT-S), 128 (ViT-B)
TIPSv2	4	20th	128
CAPI	4	20th	128

The fine-tuned models use AdamW with learning rate 2×10^{-5} ($5\times$ higher for the heads) and weight decay 0.04, and the teacher is an exponential moving average of the student, with momentum going from 0.996 to 1. The partner weight is $w = 0.5$ for all models; it is switched on after one sixth of training and ramped up linearly over the next third. DINOv2 and TIPS use two 112-pixel global and six 56-pixel local crops. CAPI keeps its own settings: 16,384 clusters, a masking ratio of 0.65, a separate AdamW optimizer for its clustering head (learning rate 5×10^{-4}), and bfloat16 mixed precision.

DINOv2 and TIPS use new DINO and iBOT heads, each a three-layer MLP with hidden size 2048 and 8192 prototypes. The student temperature is 0.1. The teacher temperature is 0.04 during the learning-rate warm-up and then follows a cosine schedule between 0.04 and 0.07. iBOT masks 40% of the patches of each global crop. Global crops cover 40–100% of the image and local crops 5–40%. CAPI keeps its own temperatures (0.12 for the student, 0.06 for the targets and 0.12 for the predictor) and crops covering 60–100% of the image. Compared with the DINOv2 pretraining

recipe, we change only what the smaller images and datasets need: 112 and 56-pixel crops instead of 224 and 98, six local crops instead of eight, 8192 prototypes instead of 65,536 because the heads are trained from scratch, and a weight of 0.1 on the iBOT loss. All models share these settings, the vanilla and random walk runs share them too, and the number of epochs is fixed before any random walk run (Appendix B.9). With these settings, the vanilla fine-tune beats the frozen checkpoint in every CIFAR-100 setting, by 3 to 15 points of k-NN accuracy, on the exact same setting.

C.3 Analysis and Considerations for the Label-Free Evaluation Metric

Our label-free evaluation metric, based on LCA and cophenetic similarity, relies on hierarchical clustering to assess global structure preservation. While our experiments demonstrate its effectiveness, several considerations warrant discussion. First, we evaluate robustness to clustering hyperparameters by testing different linkage methods (Ward, average, complete) and distance metrics (Euclidean, cosine), finding consistent rankings across methods, with Ward linkage and cosine distance providing the most stable results. Ward linkage minimizes within-cluster variance during agglomeration, producing hierarchical structures with clear semantic boundaries that align well with the multi-scale organization we aim to measure. Cosine distance is preferable to Euclidean distance in high-dimensional embedding spaces because it captures angular similarity rather than absolute distance, avoiding the distance concentration phenomenon that can reduce discriminative power when dimensionality increases. Second, we analyze sensitivity to augmentation strength, showing that the metric maintains discriminative power across varying perturbation levels without suffering from distance concentration issues in high-dimensional spaces. Third, to assess scalability across different representation sizes, we evaluate the metric on embedding dimensions of 768, 1024, 2048, 4096, and 8192. Across all tested dimensions, the metric yields meaningful and consistent results, indicating robustness to embedding dimensionality and confirming its applicability to diverse architectural choices, from standard backbone outputs to expanded projector representations. Finally, we investigate the correlation between our metric and downstream task performance, demonstrating that models scoring higher on LCA/cophenetic similarity generally achieve better performance on hierarchical classification tasks, though the relationship with fine-grained classification is more nuanced.

D Hardware and Computational overhead

We use NVIDIA L40 for our evaluations, and NVIDIA GeForce GTX 1080 Ti as well as the L40 for our training.

We benchmarked SAG-VICReg versus VICReg on an NVIDIA L40 GPU, using the complete ImageNet-1k dataset (1,281,167 samples per epoch).

- **Runtime per epoch:** SAG-VICReg: 2520s, VICReg: 2300s.
- **GPU memory:** SAG-VICReg: 25,034 MB, VICReg: 24,208 MB.

Overall, SAG-VICReg incurs a $\sim 9.6\%$ increase in computational time and a $\sim 3.4\%$ increase in memory usage compared to VICReg.

E Visualizations

Figure 8 illustrates why our label-free evaluation metric provides more comprehensive assessment than traditional linear separability measures. While the first embedding achieves perfect linear separability, it fails to preserve global structure by placing semantically distant points closer than related ones. Our second embedding preserves structural relationships despite reduced linear separability, demonstrating superior global semantic understanding and highlighting the limitations of relying solely on linear classification for embedding evaluation.

Figure 9 confirms our method's enhanced ability to preserve meaningful semantic relationships when generalizing to new data, with SAG-VICReg maintaining better semantic organization than VICReg on unseen classes. The visualization shows tighter grouping of related classes and more

Figure 8: **Comparing embedding quality assessment methods**. **Top:** Original data (left) and two embeddings, where the first (middle) is linearly separable but distorts global structure, while the second (right) preserves structural relationships but isn't linearly separable. **Bottom:** Dendrograms reveal that the second embedding's hierarchical structure closely matches the original data, maintaining key relationships (e.g., sample 0's proximity to 1, 8, 9) that are lost in the first embedding despite its linear separability. This demonstrates why our metric provides a more comprehensive evaluation than traditional approaches.

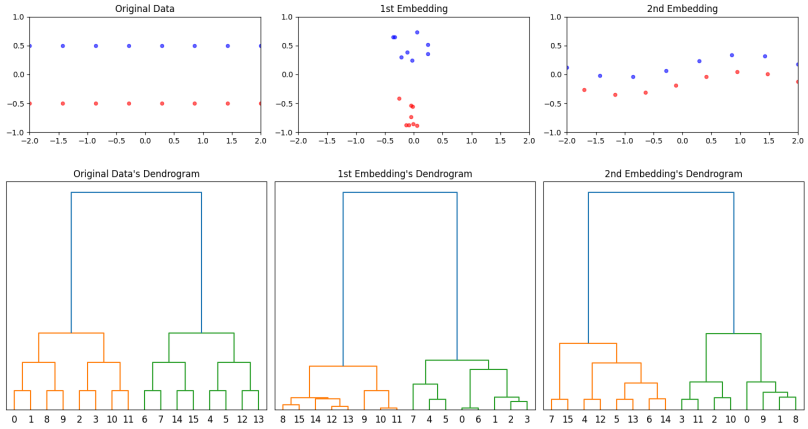
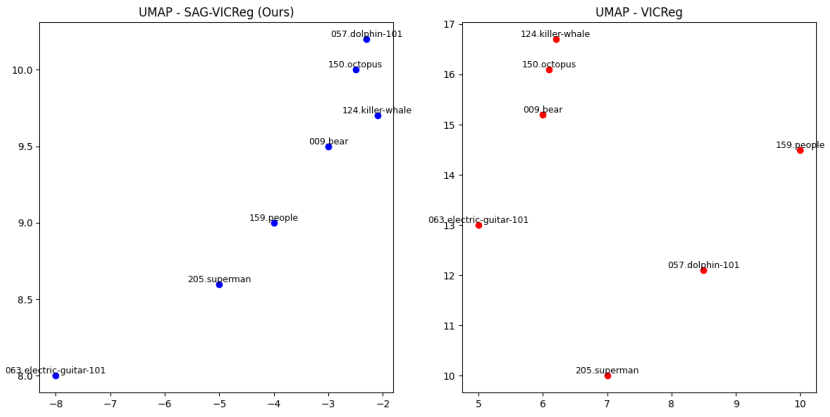


Figure 9: **UMAP visualizations of the class centroids (mean embeddings) for SAG-VICReg and VICReg.** Both models trained on the same set of classes, which differ from those used to generate the embedded images. SAG-VICReg more effectively preserves global structure even on unseen data, drawing semantically related centroids closer and pushing distant ones further apart. For example, aquatic classes cluster more tightly in SAG-VICReg. Additionally, "superman" is closer to the "people" class centroid, while "electric guitar" is placed farther from unrelated classes, showcasing SAG-VICReg's stronger semantic separation compared to VICReg.



coherent semantic organization, validating our approach's effectiveness in capturing global structure while generalizing to previously unseen data.

F ARCHITECTURE

We maintained architectural consistency across our experimental framework to ensure fair comparisons. Our implementation follows the original VICReg architecture as detailed in Figure 1 of Bardes et al. (2022).

Spectral embedding-based methods. For experiments involving spectral embedding-based SSL methods (Tables 1, 10), we employed dataset-specific backbone configurations. ImageNet and

Caltech-256 experiments utilized a ResNet-50 backbone that generated 2048-dimensional representation vectors, subsequently projected to 8192-dimensional embeddings through a series of three fully connected layers. For CIFAR-100, we adapted the architecture to use a ResNet-34 backbone, producing 512-dimensional representations expanded to 1024-dimensional embeddings via three fully connected layers.

Cross-method applicability experiments. For demonstrating broad applicability across different SSL methods, we used architecture-matched configurations. Figure 1 as well as Table 13 employed ResNet-34 for SimCLR and ResNet-34 for TCL on CIFAR-100. Figure 5 specifically used ResNet-50 for both methods in the cross-dataset transfer setting (Caltech-256 to CIFAR-100).

Hierarchical and standard evaluation experiments. For the hierarchical evaluation experiments (Figures 6, 7 and Tables 2, 7) and standard classification evaluation experiments (Tables 3, 8), we maintained the same ResNet architectures as used in Tables 1 and 10, ensuring that all comparative models utilized identical backbone networks to facilitate direct performance comparison under controlled conditions.

G DATA

Our experiments employed a carefully structured approach to data partitioning across multiple datasets.

Generalization evaluation with split datasets. For the comparisons in Tables 1 and 10, we used a consistent split methodology across datasets. We selected the first 100 classes of ImageNet (detailed class names provided in Appendix M), dividing them into two equal parts: training the first model on classes 1-50 and the second model on classes 51-100. We applied an identical splitting strategy to CIFAR-100. For Caltech-256, we implemented a 1-128 and 129-256 split. Testing was consistently performed on the first portion of each dataset (classes 1-50 for ImageNet and CIFAR-100, classes 1-128 for Caltech-256).

Cross-method applicability experiments. For demonstrating broad applicability across different SSL methods, all experiments used complete datasets. Figures 1 and Table 13 used the full CIFAR-100 dataset, while Figure 5 employed the complete Caltech-256 for pretraining with evaluation on CIFAR-100.

Hierarchical and standard evaluation experiments. For the Rand Index comparisons (Figures 6, 7), hierarchical classification experiments (Tables 2, 7), and the standard classification experiments (Tables 3, 8), we utilized the complete CIFAR-100 dataset when pretraining on CIFAR-100, the complete Caltech-256 when pretraining on Caltech-256, and the complete ImageNet-1k dataset (1,281,167 images) when pretraining on ImageNet-1k, to provide comprehensive evaluation across the entire data distributions.

H TRAINING DETAILS

We implemented a comprehensive training protocol tailored to each dataset and model architecture to ensure optimal performance while maintaining experimental validity. All models used 10 warmup epochs with cosine learning rate decay schedules. Each method is trained with at least seven seeds, and we report the mean and standard deviation; the vanilla and random walk versions of a method use the same seeds.

Spectral embedding-based SSL methods. For the spectral embedding-based SSL models comparisons (Tables 1, 10), we trained ImageNet models for 300 epochs with a batch size of 512 and a learning rate of 1.2. CIFAR-100 and Caltech-256 models underwent more extensive training for 500 epochs, with batch size of 256 and learning rate of 1.4. All invariance-based models utilized the LARS optimizer You et al. (2017); Goyal et al. (2017) with a weight decay of 10^{-6}, consistent with the original VICReg implementation.

CIFAR-100 settings. CIFAR-100 models follow the ImageNet recipe, with the changes that 32×32 images need: a ResNet-34 whose first layer is a 3×3 convolution with stride 1 and no max-pooling, 32×32 crops, an expander of three fully connected layers with 1024 units instead of 8192, and batch size 256 with learning rate 1.4 for 500 epochs. The LARS optimizer, weight decay, cosine

schedule, 10 learning-rate warm-up epochs and VICReg loss weights (25, 25, 1) are the same as for ImageNet.

Cross-method applicability experiments. For demonstrating broad applicability across different SSL methods (Figures 1, 5, Table 13), all models were trained for 500 epochs. SimCLR used LARS optimizer with batch size 512 and learning rate 0.5 on Caltech-256, and CIFAR-100. TCL used SGD optimizer with batch size 512 and learning rate 0.2 on Caltech-256, and batch size 256 with learning rate 0.1 on CIFAR-100.

Hierarchical and standard evaluation experiments. Models used in the Rand Index comparisons (Figures 6, 7) were trained for 500 epochs. For the hierarchical and standard classification experiments (Tables 2, 7, 3, 8), we trained models for 500 epochs on CIFAR-100 and Caltech-256, and 250 epochs on ImageNet-1k. Batch sizes were set to 256 for CIFAR-100 (learning rate 1.4), 512 for Caltech-256 (learning rate 0.7), and 1024 for ImageNet-1k (learning rate 0.3), using LARS optimizer with weight decay of 10^{-6}.

I Algorithms

I.1 SAG-VICReg

Algorithm 1 SAG-VICReg

Input: Mini-batch of images $S = \{s_1, s_2, \ldots, s_n\} \subset \mathcal{D}$, transformations $t, t' \sim \mathcal{T}$
Output: None

1: Apply transformations t and t' to S, generating two transformed views $x = t(i)$ and $x' = t'(i)$
2: Forward the views x and x' through the network by $z = h_\phi(f_\theta(x))$, $z' = h_\phi(f_\theta(x'))$ and form 2 batches $Z = \{z_1, \ldots, z_n\}$ and $Z' = \{z'_1, \ldots, z'_n\}$
3: Compute affinity matrix $W \in \mathbb{R}^{n \times n}$ between samples in Z and Z'
4: Compute a random walk matrix P and sample n pairs to form the new batch $Z'' = \{z''_1, \ldots, z''_n\}$, where every z''_i is selected from Z'
5: Compute the weighted VICReg's invariance loss using W_{ij} between Z, Z'', and compute the regular VICReg's variance and covariance on each Z, Z'' separately

I.2 Global Structure-Preserving Evaluation

Algorithm 2 Evaluation of Structural Similarity between Embedding Sets

Input: Two embedding sets $A = \{a_1, a_2, \ldots, a_n\} \subset \mathbb{R}^d$ and $B = \{b_1, b_2, \ldots, b_n\} \subset \mathbb{R}^m$
Output: Similarity scores based on hierarchical and global structure preservation

1: Construct hierarchical trees T_1 and T_2 for A and B using agglomerative clustering, and calculate the distances to LCA and the cophenetic distances D_1, D_2 for all pairs
2: Compute Pearson, Spearman, and Kendall correlation coefficients between the LCA distances of A and B
3: Compute cophenetic correlation coefficient between D_1 and the pairwise distances in B, D_2 and the pairwise distances in A
4: Return all the calculated correlation coefficients

J Additional Preliminaries

J.1 Spectral Embedding

Spectral embedding transforms high-dimensional data into a lower-dimensional representation by exploiting the eigenstructure of a graph's Laplacian matrix. This technique preserves the underlying relationships between data points through their graph representation. Consider a weighted, undirected graph $G = (V, E)$ with n vertices. The graph structure is characterized by two fundamental matrices: The weight matrix $W \in \mathbb{R}^{n \times n}$, where: $W_{ij} =$ similarity between vertices i and j. The degree matrix $D \in \mathbb{R}^{n \times n}$, a diagonal matrix where: $D_{ii} = \sum_{j=1}^{n} W_{ij}$ The graph Laplacian matrix L is then defined as: $L = D - W$.

To compute the k-dimensional spectral embedding, we first need to find the eigendecomposition of L. Next, from this decomposition, we take the k eigenvectors $v_1, \ldots, v_k$ that correspond to the k smallest non-zero eigenvalues. Finally, we construct the embedding matrix $X \in \mathbb{R}^{n \times k}$, where each row i contains the embedded coordinates for vertex i in the lower-dimensional space. This process transforms each vertex from our original graph into a point in a k-dimensional space, where vertices that were strongly connected in the original graph remain close to each other in this new representation.

A major limitation of spectral embedding is that adding new points to the graph requires recalculating W, D, L. When a new point is introduced, not only do we need to calculate its similarities to all existing vertices, but we also need to update the graph's structure to incorporate these new relationships. Therefore, finding the spectral embeddings requires a recalculation of the eigendecomposition to obtain the new k-dimensional embeddings. This limitation shows why spectral embeddings are intended primarily for fixed graphs, as in Bengio et al. (2003), Fowlkes et al. (2004) and Coifman & Lafon (2006). SpectralNet Shaham et al. (2018), aims to solve this generalization issue using a deep neural network that learns the spectral embedding.

J.2 Random walk on a Graph

A random walk on a Graph is a stochastic process where each step involves moving from one node to another according to edge weights, capturing the connectivity structure of the graph. Given the weight matrix W and degree matrix D, the random walk matrix P is defined as: $P = D^{-1}W$, where P_{ij} represents the probability of transitioning from node i to node j in a single step: $P_{ij} = \frac{W_{ij}}{D_{ii}}$. This matrix P is row-stochastic, meaning each row sums to 1, representing the transition probabilities of a Markov chain. To sample an index j from a random walk matrix P in the i-th row, we choose one index according to the probability distribution defined by Pij.

J.3 Cophenetic Correlation

Cophenetic correlation Sokal & Rohlf (1962) is a metric used to evaluate how well a hierarchical clustering structure preserves the pairwise distances between observations in a dataset. This measure is particularly useful in embedding evaluation, as it assesses whether embeddings reflect the hierarchical relationships inherent in the data. Given a hierarchical clustering, the cophenetic distance between any two points i and j is defined as the height of the dendrogram at which these two points are first joined into a single cluster. Let $T(i, j)$ denote this cophenetic distance for points i and j based on the hierarchical clustering. The cophenetic correlation coefficient c quantifies the correlation between the original pairwise distances $D(i, j)$ in the data and the cophenetic distances $T(i, j)$. Mathematically, it is calculated as:

$$c = \frac{\sum_{i<j} \left(D(i,j) - \bar{D}\right)\left(T(i,j) - \overline{T}\right)}{\sqrt{\sum_{i<j} \left(D(i,j) - \bar{D}\right)^2 \sum_{i<j} \left(T(i,j) - \overline{T}\right)^2}}$$

where: $\bar{D}$ is the mean of all pairwise distances $D(i, j)$, and $\overline{T}$ is the mean of all cophenetic distances. The distance metric of D can be Euclidean or cosine distance, as well as other metrics. The cophenetic distance T can be calculated by tracing the dendrogram tree and identifying the height at which clusters merge.

J.4 Rand index

The Rand Index Rand (1971) is a measure for quantifying the similarity between two clusterings of the same dataset. Let there be a dataset $\{x_1, x_2, \ldots, x_n\}$ of n elements, and suppose we have two partitions (clusterings) of this dataset, denoted by X and Y. Each partition assigns every element x_i to a cluster label. The Rand Index measures how consistently pairs of elements are assigned to the same cluster or to different clusters in both partitions.

Formally, consider all possible $\binom{n}{2}$ pairs of distinct elements. A pair (x_i, x_j) is counted as an *agreement* if they are either:

- Placed in the same cluster under both partitions X and Y, or

- Placed in different clusters under both partitions X and Y.

The Rand Index is given by the fraction of such pairwise agreements out of all possible pairs:

$$\text{RandIndex}(X, Y) = \frac{\text{number of pairwise agreements}}{\binom{n}{2}}.$$

A Rand Index close to 1 indicates a high level of agreement between the two clusterings, while a value closer to 0 indicates lower agreement.

K RAND INDEX SETUP

In this configuration, all models are trained on the full CIFAR-100 dataset for 500 epochs, and unlike the LCA and Cophenetic similarity approaches, we do not train two separate models per method. We begin by constructing a hierarchical tree of the CIFAR-100 label structure, following the hierarchy described by Kim et al. (2023) (Table S1). For each level in this hierarchy, we compute the Rand index by comparing two partitions: (1) the clustering assignments obtained via spectral clustering on the embeddings, and (2) the corresponding labels at that level. The number of clusters is set to match the number of classes at each level.

Beyond this hierarchical approach, we also calculate the Rand index at the finest level of granularity (i.e., all 100 CIFAR-100 classes). In this case, the number of clusters produced by spectral clustering is allowed to vary, rather than being fixed to the total number of classes.

The corresponding results for these approaches are presented in Figures 6 and 7, respectively.

L IMPLEMENTATION DETAILS

The affinity matrix W is computed to represent the relationships between points in X and Y. The steps are as follows:

L.1 NEAREST NEIGHBORS AND DISTANCE ADJUSTMENT

The distances to the k-nearest neighbors (D) are computed using cosine distance, where $k = 5$. To emphasize relative relationships, the distance to the nearest neighbor is subtracted from all distances, inspired by McInnes et al. (2018):

$$\mathrm{D}_{ij} = \mathrm{D}_{ij} - \mathrm{D}_{i0}, \quad \forall j$$

L.2 SCALE CALCULATION

The scale parameter determines the Gaussian kernel width and is derived based on the adjusted distances (after decreasing the distance to the nearest neighbor from the distance to every sample point, which results in nearest distance $= 0$). For local scaling, the 20th percentile of the adjusted distances is used:

$$\text{scale}_i = \max(\text{Percentile}(\mathrm{D}_i, 20), 1e^{-7})$$

The $1e^{-7}$ ensures numerical stability by clamping the scale to a minimum value.

L.3 GAUSSIAN KERNEL

Using the computed scale, the Gaussian kernel is applied to the pairwise distances D:

$$W_{ij} = \exp\left(-\frac{D_{ij}^2}{\text{scale}_i^2}\right)$$

This creates the affinity matrix, emphasizing points within the local neighborhood.

L.4 MASKING NON-NEIGHBORS

A binary mask is applied to sparsify W, keeping only the entries corresponding to the $k = 5$ nearest neighbors:

$$W_{ij} = \begin{cases} W_{ij}, & j \in \text{Ids}_i \\ 0, & \text{otherwise} \end{cases}$$

M IMAGENET CLASSES

The following are the first 100 classes of ImageNet used in our experiments:

n01440764, n01443537, n01484850, n01491361,
n01494475, n01496331, n01498041, n01514668,
n01514859, n01518878, n01530575, n01531178,
n01532829, n01534433, n01537544, n01558993,
n01560419, n01580077, n01582220, n01592084,
n01601694, n01608432, n01614925, n01616318,
n01622779, n01629819, n01630670, n01631663,
n01632458, n01632777, n01641577, n01644373,
n01644900, n01664065, n01665541, n01667114,
n01667778, n01669191, n01675722, n01677366,
n01682714, n01685808, n01687978, n01688243,
n01689811, n01692333, n01693334, n01694178,
n01695060, n01697457, n01698640, n01704323,
n01728572, n01728920, n01729322, n01729977,
n01734418, n01735189, n01737021, n01739381,
n01740131, n01742172, n01744401, n01748264,
n01749939, n01751748, n01753488, n01755581,
n01756291, n01768244, n01770081, n01770393,
n01773157, n01773549, n01773797, n01774384,
n01774750, n01775062, n01776313, n01784675,
n01795545, n01796340, n01797886, n01798484,
n01806143, n01806567, n01807496, n01817953,
n01818515, n01819313, n01820546, n01824575,
n01828970, n01829413, n01833805, n01843065,
n01843383, n01847000, n01855032, n01855672.