SOFT EQUIVARIANCE REGULARIZATION FOR INVARIANT SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A central principle in self-supervised learning (SSL) is to learn data representations that are invariant to semantic-preserving transformations *e.g.*, image representations should remain unchanged under augmentations like cropping or color jitter. While effective for classification, such invariance can suppress transformation-relevant information that is valuable for other tasks. To address this, recent works explore equivariant representation learning, which encourages representations to retain information about the applied transformations. However, how to effectively incorporate equivariance as an explicit regularizer on top of strong invariance-based SSL backbones at ImageNet scale remains underexplored. We conjecture that enforcing invariance and equivariance to the same layer is inherently difficult and, if handled naively, may even hinder learning. To overcome this, we propose soft equivariance regularization (SER), a simple yet scalable method that decouples the two objectives: learning invariant representations via standard SSL, while softly regularizing intermediate features with an equivariance loss. Our approach necessitates neither a transformation label nor its predictive objectives, but operates directly with group actions applied to the intermediate feature maps. We show that this soft equivariance regularization significantly improves the generalization performance of ImageNet-1k pre-training of vision transformers (ViT), leading to stronger downstream classification accuracy in ImageNet and in its variants, including both natural distributions and broad types of common corruptions and perturbations ImageNet-C and ImageNet-P. Our code is available at `https://anonymous.4open.science/r/erl-B5CE`.

1 INTRODUCTION

Self-supervised learning (SSL) has become a cornerstone in modern machine learning, especially within computer vision (Chen et al., 2020; Caron et al., 2021; Wang et al., 2023; Huang et al., 2023), enabling the extraction of rich and generalizable representations from large-scale unlabeled datasets. A prominent approach in SSL seeks representations invariant to predefined data augmentations, such as random cropping, color jittering, and rotations, under the assumption that these augmentations *should not alter the underlying semantic content*. While invariance encourages stable representation learning, relying solely on the invariance task may lead to the loss of valuable transformation-dependent information, potentially yielding suboptimal representations for downstream tasks. Incorporating equivariance explicitly modeling how representations should transform in response to input changes allows for the preservation and effective utilization of such information, thereby enriching the learned features and enhancing their relevance across diverse tasks (Dangovski et al., 2021; Marchetti et al., 2023).

This principle of equivariance ensures that representations transform predictably in response to changes in the input. Instead of discarding transformation-specific information, equivariant methods aim to encode it in a structured manner within the representation space. Existing approaches typically fall into two categories (Yu et al., 2024): implicit methods, which learn equivariance through auxiliary tasks such as predicting transformations applied to input pairs (Dangovski et al., 2021; Lee et al., 2021). Meanwhile, explicit methods directly model the transformation within the latent space, often requiring transformation labels to learn the corresponding representation transformation (Devillers & Lefort, 2023; Park et al., 2022; Garrido et al., 2023).

However, in practice, explicit methods often encounter significant challenges (Yu et al., 2024). These include reliance on transformation labels, which may not always be available; difficulty in capturing inter-dependencies between combined transformations (*e.g.*, simultaneous variations in cropping and color); and limitations in modeling complex, non-atomic augmentations (Yu et al., 2024). Additionally, most existing equivariant methods have been developed and evaluated predominantly on convolutional neural networks (CNNs), particularly ResNet variants (Devillers & Lefort, 2023; Yu et al., 2024). Their efficacy when applied to architectures with less inherent inductive bias, such as Vision Transformers (ViTs) (Dosovitskiy et al., 2021), remains largely unexplored. Equivariant self-supervised learning has been explored at ImageNet scale, including world-model–based approaches (Garrido et al., 2024). In contrast, we study explicit equivariance regularization on top of strong invariance-based SSL methods (MoCo-v3, DINO, BarlowTwins) and show that it can improve performance on ImageNet-scale datasets and robustness benchmarks. We hypothesize and empirically validate that expecting a single representation to exhibit complete invariance and nuanced transformation responsiveness simultaneously is both technically challenging and generally unnecessary.

To address these challenges, we propose a novel SSL framework that introduces transformation equivariance through a fundamentally different perspective. Unlike previous methods that impose equivariance constraints exclusively on spatially-collapsed representations via complex mechanisms, our framework employs soft regularization to minimize equivariance errors at intermediate, (spatial) structure-preserving layers. This strategy decouples invariance learning, achieved by standard contrastive objectives at the output layer, from equivariance learning, encouraged through regularization at earlier layers with preserved spatial structure.

It is worth noting that our method does not depend on inherently equivariant architectures such as CNNs for translation invariance. Instead, we utilize flexible models like ViTs suitable for large-scale training as up-to-date state-of-the-art backbones (Dosovitskiy et al., 2021), and known to even exceed architectures designed for certain symmetry, *e.g.*, CNNs for translation, at learning equivariance (Gruver et al., 2022) and introduce a soft inductive bias favoring equivariant representations. This principle incorporating subtle structural bias rather than enforcing rigid constraints has been demonstrated to enhance generalization both empirically and theoretically (Finzi et al., 2021; Kim et al., 2023; Wilson, 2025). Our equivariance regularizer, defined through a straightforward group-theoretic equivariance error, neither requires training transformation predictors nor access to explicit transformation labels.

We evaluate our method extensively across standard vision benchmarks and downstream tasks, including both natural distributions and broad types of common corruptions and perturbations. Our experiments show that the proposed method scales effectively to ViTs pre-training on ImageNet, consistently improving downstream classification performance across various base SSL methods used for invariance learning.

To summarize, our contribution is threefold:

- We empirically demonstrate that imposing equivariance and invariance on the *same final layer* is sub-optimal: it significantly degrades downstream accuracy while increasing transformation sensitivity (Figure 3, Table 4). This validates our core conjecture that these two objectives fundamentally conflict when applied jointly on the final representation.

- Motivated by this observation, we propose SER, a framework that decouples invariance and equivariance learning by applying a soft equivariance regularizer at an intermediate, spatially structured layer while keeping the final representation trained purely with a standard invariance-based SSL objective. The method is mathematically simple, relying on direct group actions on intermediate feature maps as the regularization mechanism.

- SER leverages known geometric group actions and avoids supervision from transformation labels or additional modules to model transformation information. When plugged into strong invariance-based SSL methods (MoCo-v3, DINO, Barlow Twins), it consistently improves performance on ImageNet-scale classification and robustness benchmarks (e.g., ImageNet-C/P) as well as downstream tasks such as COCO detection and 3DIEBench.

2 BACKGROUNDS

2.1 SELF-SUPERVISED LEARNING

Self-supervised learning (SSL) leverages intrinsic supervisory signals derived directly from the data, circumventing the need for costly human-annotated labels. SSL methods typically construct proxy tasks such as predicting rotations (Gidaris et al., 2018), solving jigsaw puzzles (Noroozi & Favaro, 2016), or performing instance discrimination via contrastive learning (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Zbontar et al., 2021) to guide neural networks in learning meaningful representations. Central to many SSL approaches is the enforcement of invariance to semantically irrelevant data augmentations, ensuring the representations capture intrinsic content rather than superficial variations. Recent advances demonstrate that enforcing invariance through contrastive losses or similarity constraints yields representations competitive with or superior to supervised learning in various vision tasks (Chen et al., 2020; Caron et al., 2021; Bardes et al., 2022).

In practice, SSL frameworks often employ multiple (usually 2) augmented views generated by independently sampling transformations from a predefined augmentation distribution. Increasing the number of these views (crops) can easily improve representation quality but incurs extra computational and memory costs (Caron et al., 2020). Contemporary SSL algorithms utilize diverse invariance objectives: SimCLR and MoCo-v3 use noise-contrastive estimation losses; SimSiam and BYOL rely on cosine similarity; and Barlow Twins combines covariance-based redundancy reduction with invariance constraints (Chen et al., 2020; He et al., 2020; Chen & He, 2021; Grill et al., 2020; Zbontar et al., 2021). Our proposed method complements these approaches by introducing a joint optimization of an equivariance regularization term alongside standard invariance-based objectives (see Section 3.3).

2.2 EQUIVARIANT REPRESENTATION LEARNING

The goal of equivariant representation learning in SSL is to complement invariant representation learning by encouraging representations to be responsive to transformations. Most existing approaches implement this by introducing additional loss functions to impose equivariance, typically applied to the same layer from which invariant representations are derived. These losses capture equivariance either implicitly or explicitly. For example, methods such as E-SSL (Dangovski et al., 2021) and AugSelf (Lee et al., 2021) indirectly promote equivariance by training models to predict transformation labels applied to the inputs. However, such approaches often struggle to capture structured or complex transformations precisely.

In contrast, explicit methods directly model transformations in the representation space. For example, EquiMod (Devillers & Lefort, 2023) constrains latent spaces to predict embedding displacements, but its heavy reliance on transformation labels limits its effectiveness with interdependent or complex augmentations such as AugMix (Hendrycks et al., 2019). Self-supervised Transformation Learning (STL) (Yu et al., 2024), on the other hand, mitigates label dependency by modelling transformation representations from image pairs, making it more flexible with complex augmentations. Nevertheless, STL can suffer from spatial collapse, reducing its sensitivity to subtle transformations. Common limitations across existing methods include dependency on transformation labels, difficulty handling multiple augmentations simultaneously, and restricted applicability beyond CNN-based architectures. Our approach overcomes these issues by softly enforcing equivariance at intermediate layers of ViTs, without relying on explicit labels or auxiliary modules to extract spatial information once collapsed (*e.g.*, through global average pooling). By directly applying group actions as regularization, our method preserves domain structure, avoids spatial collapse, and enhances scalability and downstream task performance in ViTs.

2.3 SYMMETRY, GROUPS, AND EQUIVARIANCE

Symmetry refers to a transformation that leaves an object unchanged (Bronstein et al., 2021). For example, rotating a perfect circle around its center does not alter its appearance. The set of all such transformations that preserve an object's structure forms a *symmetry group*. Formally, a group is a mathematical structure consisting of a set of elements and a binary operation (here, composition of transformations) that satisfies four properties: closure (the composition of two symmetries is also a symmetry), associativity, existence of an identity element, and existence of inverses. Symmetry

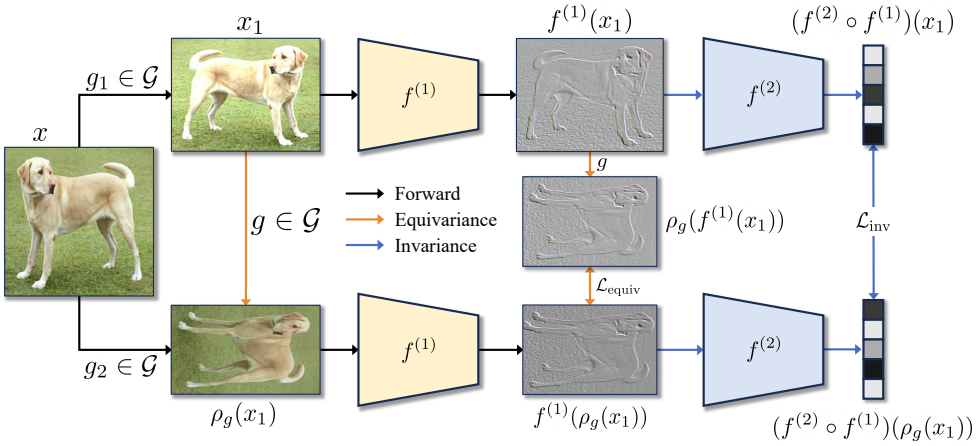

Figure 1: An exemplary overview of our soft equivariance regularization for self-supervised learning. The image pair is created via the group actions g_1 and g_2. For simplicity, we omit the intensity transformation applied to the original image (see Section 3.1).

arises in many domains, such as images defined on a 2D grid or molecules in 3D space, and encodes a form of structure or redundancy in data. Leveraging such symmetries allows machine learning models to generalize better from limited data, as they can capture invariances or equivariances induced by the underlying group actions.

To formalize how functions respond to symmetries, we consider *group representations*. A *representation* of g in a group $\mathcal{G}$ on a Euclidean space $\mathbb{R}^n$ is a homomorphism $\rho_g : \mathcal{G} \to \mathrm{GL}(n)$, where $\mathrm{GL}(n)$ is the group of invertible $n \times n$ matrices. This mapping preserves the group structure, meaning $\rho_{gh}(\cdot) = \rho_g \rho_h(\cdot)$ for all $g, h \in \mathcal{G}$. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be *$\mathcal{G}$-equivariant* if for all $g \in \mathcal{G}$ and $x \in \mathcal{X}$,

$$f\big(\rho_g(x)\big) = \rho_g\big(f(x)\big).$$

Note that we unify the notation of the representation ρ_g for both $\mathcal{X}$ and $\mathcal{Y}$ for simplicity. In practice, they use different representations due to different dimension size. Intuitively, section 2.3 means that applying a transformation g to the input and then computing f is equivalent to computing f first and then transforming the output by g. Equivariance implies that the function respects the structure imposed by the group action, rather than discarding it.

CNNs exemplify this principle: their convolution layers are equivariant to translations, assuming an idealized setting over $\mathbb{R}^2$. This built-in translation symmetry has been crucial to their success in image analysis. Motivated by this, a wide range of architectures, such as group-equivariant CNNs (Cohen & Welling, 2016; 2017) and equivariant graph networks (Keriven & Peyré, 2019), have been developed to encode other symmetry types, leading to improved generalization, data efficiency, and interpretability.

3 Soft Equivariance Regularization for Invariant Self-Supervised Learning

Previous methods for introducing equivariance into invariant SSL typically impose both invariance and equivariance objectives on the output layer representations. However, these representations are often spatially collapsed, which may be suitable for enforcing invariance but are generally inadequate for capturing transformation-sensitive equivariant structures. Therefore, we explicitly encourage equivariance at the *intermediate representations* computed at earlier layers, which retain spatial structure and are better aligned with group actions.

3.1 Soft Equivariance at Intermediate Features

A straightforward way to introduce equivariance would be to impose it directly on the final representation (e.g., globally pooled feature in ResNets or the `[CLS]` token in ViTs), as explored in prior work. However, these final representations are spatially collapsed and no longer admit a natural

spatial group action. We therefore leverage non [CLS] patch tokens where spatial structure is still explicit (e.g., the gray feature maps in Figure 1).

For ViTs, which are our primary focus, spatial structure is disrupted after the introduction of the [CLS] token. To preserve a spatial lattice for equivariance, we decompose the encoder f into two components:

$$f = f^{(2)} \circ f^{(1)},$$

where $f^{(1)}$ is a structure-preserving, equivariant feature extractor, and $f^{(2)}$ is an invariance-oriented head. The [CLS] token is introduced only at the input of $f^{(2)}$, so it does not affect the feature maps produced by $f^{(1)}$. As a result, the outputs of $f^{(1)}$ remain defined on a regular spatial grid and are amenable to group actions. The overall architecture is illustrated in Figure 2.

We adopt the standard two-view SSL protocol, following the principle of Gupta et al. (2023) that equivariance should be learned from pairs of augmented samples, analogous to invariant contrastive learning. Given an image x and two sampled transformations $g_1, g_2 \sim \mathcal{G}$, we form two views

$$x_1 = \rho_{g_1}(x), \qquad x_2 = \rho_{g_2}(x) = \rho_g(x_1),$$

where the relative group element $g = g_2 g_1^{-1}$ maps x_1 to x_2. We exploit this relative transform both in input space and feature space. On the feature side, we apply g to the intermediate representation via the group action ρ_g to obtain $\rho_g\big(f^{(1)}(x)\big)$, and we compare it to the representation obtained by transforming the input first and then encoding:

$$\rho_g\big(f^{(1)}(x)\big) \quad \text{vs.} \quad f^{(1)}\big(\rho_g(x)\big).$$

The equivariance constraint is formalized as

$$\mathcal{L}_{\text{equiv}} = \mathbb{E}_{x,\,(g_1,g_2)\sim\mathcal{G}}\Big[d\big(\rho_g(f^{(1)}(x)), f^{(1)}(\rho_g(x))\big)\Big], \quad g = g_2 g_1^{-1}, \tag{1}$$

where $d(\cdot,\cdot)$ is a distance measure between feature maps; in this work we instantiate d with a contrastive loss as described in Section 3.3. This form of equivariance objective has appeared before (e.g., Eq. 4 in Yu et al. (2024)); our key difference is that we apply it to an intermediate, spatially structured representation and use the known group action ρ_g directly, without training an additional transformation-prediction module or latent action network. Minimizing $\mathcal{L}_{\text{equiv}}$ does not enforce exact equivariance, but rather encourages *soft equivariance* at that layer.

Because $\mathcal{L}_{\text{equiv}}$ alone does not provide an instance-discrimination signal for the final representation, we jointly train the invariance-oriented head $f^{(2)}$ with a standard SSL loss on the [CLS] token, as in MoCo-v3, DINO, and Barlow Twins (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Zbontar et al., 2021). The full procedure is summarized in Figure 1. Importantly, we do not augment the network with any extra module to model transformations (unlike, e.g., EquiMod or STL (Devillers & Lefort, 2023; Yu et al., 2024)); instead, we reuse the known image-level group actions to define ρ_g on the intermediate feature maps and regularize the encoder accordingly.

3.2 Augmentation Policy and Batch Partitioning

Typical augmentation policies used in invariance-based SSL include RandomResizedCrop, RandomHorizontalFlip, and photometric modifications such as color jittering and grayscale. However, RandomResizedCrop does not form a group; after cropping, the discarded region cannot be recovered by applying another crop, so no inverse exists (see Section 2.3). Most importantly, it changes the spatial support of the image; as a result, the relative transform $g = g_2 g_1^{-1}$ and the corresponding ρ_g for the spatially-structured feature map cannot be well-defined with RandomResizedCrop. Therefore, we split each mini-batch into two sub-batches, i.e., b_1 and b_2 (see Figure 2).

b_1 employs the existing invariant SSL framework (including its augmentation policy denoted as $\mathcal{T}$), which SER aims to enhance. On the other hand, b_2 leverages $\mathcal{G}$, which is the modified version of $\mathcal{T}$; it excludes RandomResizedCrop and adds Rotation90°:

$$b_1 : \mathcal{T}, \qquad b_2 : \mathcal{G} = \mathcal{T} \setminus \{\text{Random Crop}\} \cup \{\text{Rotation } 90^\circ\}.$$

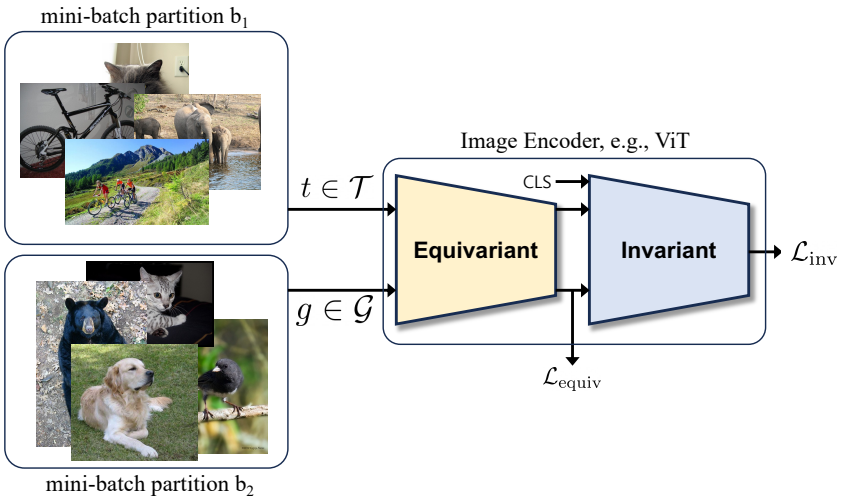

Figure 2: An overview of the training pipeline. Mini-batch is randomly divided into two partitions; the standard augmentation set for self-supervised learning applies to partition 1, whereas a slightly modified policy applies to subset 2. Differences are as follows: 1) random crop is removed from $\mathcal{T}$ because symmetry cannot hold for crop, and 2) rotation 90° is added to $\mathcal{G}$.

Moreover, within $\mathcal{G}$, we define the group action ρ_g only on the (invertible) geometric subset (anisotropic scaling inherited from `RandomResizedCrop` without cropping, `RandomHorizontalFlip`, and `Rotation90°`). Photometric augmentations (color jitter, grayscale, blur, solarization) are retained in $\mathcal{G}$ but do not contribute to equivariance, for no group action is associated with them. In summary, we randomly split the mini-batch into b_1 and b_2. b_1 adopts the exact baseline invariant SSL algorithm including $\mathcal{T}$ and its loss function that yields L_{inv1}, whereas b_2 employs $\mathcal{G}$, and we define the group action ρ_g only on the geometric subset of $\mathcal{G}$. Note that b_2 outputs both L_{inv2} (using the same invariant loss function for b_1) and L_{equiv}, which we describe in Section 3.3.

3.3 Training Objective for Soft Equivariance Regularization

To encourage predictable responses to input transformations, we introduced the equivariance regularizer in Equation (1) as a patch-wise NT-Xent (noise-contrastive) loss applied on the sub-batch b_2 (Chen et al., 2020). Let H_f, W_f, and D_f denote the height, width, and channel dimension of the intermediate feature maps. We write

$$\boldsymbol{z} = \rho_g\big(f^{(1)}(x)\big) \quad \text{and} \quad \boldsymbol{z}' = f^{(1)}\big(\rho_g(x)\big),$$

with $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^{H_f \times W_f \times D_f}$ the two transformed feature maps for a given relative group element $g = g_2 g_1^{-1}$. We index images in b_2 by i and spatial locations by $j \in \{0, \ldots, H_f W_f - 1\}$, and denote by z_{ij} and z'_{ij} the corresponding feature vectors from $\boldsymbol{z}$ and $\boldsymbol{z}'$, respectively. Each vector is first projected by a 2-layer MLP with GELU into a 512-dimensional space (Caron et al., 2021). The equivariance contrastive loss for anchor (i, j) is then

$$\mathcal{L}^{i,j}_{\text{equiv}} = -\log \frac{\exp\big(s(z_{ij}, z'_{ij})\big)}{\exp\big(s(z_{ij}, z'_{ij})\big) + \sum_{m \neq i} \sum_n \big[\exp\big(s(z_{ij}, z_{mn})\big) + \exp\big(s(z_{ij}, z'_{mn})\big)\big]},$$

where $s(x, y)$ denotes temperature-scaled cosine similarity, $s(x, y) = \frac{1}{\tau} x^\top y / (\|x\| \|y\|)$. We set $\tau = 0.3$ for MoCo-v3 and Barlow Twins, and $\tau = 0.5$ for DINO. Following O Pinheiro et al. (2020), negatives are sampled only from *other* images in the batch; i.e., we omit all tokens from the same image as the anchor. The overall equivariance loss averages this quantity over all images and spatial locations in b_2,

$$\mathcal{L}_{\text{equiv}} = \frac{1}{|b_2| H_f W_f} \sum_i \sum_j \mathcal{L}^{i,j}_{\text{equiv}}.$$

The full training objective combines the standard invariance loss with our equivariance regularizer:

$$\mathcal{L} = \mathcal{L}_{\text{inv1}} + \mathcal{L}_{\text{inv2}} + \lambda \mathcal{L}_{\text{equiv}},$$

where the hyperparameter $\lambda > 0$ controls the strength of equivariance regularization. Both $\mathcal{L}_{\text{inv1}}$ and $\mathcal{L}_{\text{inv2}}$ employ exactly the baseline invariance loss function (e.g., MoCo-v3, DINO, Barlow Twins) but applied to sub-batches b_1 and b_2, respectively. Thus, our objective is agnostic to the choice of base SSL algorithm and can be seamlessly integrated with different invariance-based methods, consistently improving downstream classification performance (see Section 4.3). This loss encourages *soft* equivariance at a spatially-structured representation, thereby preserving flexibility of the representation space.

4 EXPERIMENTS

In this section, we empirically demonstrate the effectiveness of the proposed SSL algorithm with soft equivariance regularization against state-of-the-art equivariant representation learning baselines through comprehensive experiments. We first detail our experimental setup in Section 4.1, and subsequently address the following key research questions in our experimental results:

- Does the proposed soft equivariant regularization improve the generalization performance of ViTs compared to purely invariant and other equivariant SSL baselines?
- Does our approach scale to large-scale pre-training scenarios, and what is its impact on downstream classification tasks that rely on transformation-specific information?
- How robust is our approach when facing complex and combined augmentations or shifts that challenge existing equivariant representation methods?

4.1 EXPERIMENTAL SETUP

Baselines. We evaluate our method by comparing it with other approaches designed to encourage equivariance within self-supervised learning frameworks. Our baselines include both implicit equivariance methods such as E-SSL (Dangovski et al., 2021) and AugSelf (Lee et al., 2021) and explicit equivariance methods, including EquiMod (Devillers & Lefort, 2023) and STL (Yu et al., 2024). Notably, EquiMod utilizes three global crops, whereas E-SSL employs two global and four local crops, making direct comparison challenging due to these differing cropping strategies. It is well-known that increasing the number of crops generally improves performance, albeit at the expense of greater memory usage and computational cost (Caron et al., 2020; 2021). To address this discrepancy and ensure a fair evaluation, we reimplement our method using a consistent 2+4 cropping scheme and report these adjusted results as well.

Dataset. To assess the efficacy and scalability of our equivariance regularization approach, we conduct pre-training and evaluation experiments using the ImageNet-1k dataset (Deng et al., 2009), adhering to standard evaluation protocols established in the self-supervised learning literature (Chen et al., 2020; Caron et al., 2021). Additionally, we evaluate our method on ImageNet variants specifically designed to measure robustness and generalizability to a broad spectrum of natural distribution shifts, ImageNet-Sketch (Wang et al., 2019), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021) and commonly-induced corruptions and perturbations ImageNet-C, and ImageNet-P (Hendrycks & Dietterich, 2019). Though these sets are all designed to evaluate whether the model is robust to corruption and perturbation, it has to be noted that ImageNet-P is more corrupted with geometric distortion, *e.g.*, translation, rotation, and scaling, whereas the distortion to ImageNet-C is primarily focused on appearance-based corruption *e.g.*, blurring, pixel noise, brightness changes, fog), which affect the texture or color of the image rather than its geometric structure. In addition, we employ the 3DIEBench dataset (Garrido et al., 2023) as an out-of-domain dataset for transfer learning on semantic classification, especially suited for evaluating the model's ability towards invariance and equivariance equipped with realistic 3D transformation. As a whole, this comprehensive evaluation aims to demonstrate the improved generalization capabilities enabled by our soft equivariance regularization.

Implementation Details. Unless otherwise noted, we pretrained ViT-small using the ImageNet-1k dataset. We follow standard augmentation practices with a scaling ranging between 0.7 and 1.3. As

Table 1: **Top-1 and top-5 accuracy (%) under linear evaluation.** Note that all equivariant representation learning methods use MoCo (He et al., 2020) as their baseline, which outperformed DINO and BarlowTwins in our setting (see Table 2). Concatenated `[CLS]` tokens from the last 4 layers were used as an input to the linear classifier, following the feature-based evaluations in (Devlin et al., 2019; Caron et al., 2021). 'View' refers to the number of crops sampled per image (see Section B for more detail). ImageNet-1k scores are averaged over 3 runs.

View	Algorithm	Param (M)	ImageNet-1k		ImageNet-Sketch		ImageNet-V2		ImageNet-R		3DIEBench	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
2 view	MoCo-v3	42.9	68.44 ±0.07	88.02 ±0.04	17.65	31.87	56.54	78.68	18.59	30.08	68.43	91.96
	+ AugSelf	43.7	67.55 ±0.05	87.62 ±0.05	13.30	25.35	53.74	76.68	17.62	28.66	64.97	90.73
	+ STL	62.2	65.49 ±0.12	85.91 ±0.08	15.40	28.96	55.43	78.02	17.22	28.49	-	-
	+ Ours	43.4	**69.28 ±0.01**	**88.79 ±0.02**	**17.68**	**32.54**	**56.95**	**79.29**	**18.95**	**30.72**	**70.17**	**92.78**
3 view	+ EquiMod	43.3	68.95 ±0.02	88.87 ±0.01	14.81	28.11	56.31	79.93	16.54	27.32	67.97	91.97
2+4 view	+ E-SSL	43.3	70.6 ±0.04	89.85 ±0.02	19.23	34.77	58.33	**80.93**	19.86	32.36	-	-
	+ Ours	43.4	**71.56 ±0.03**	**90.04 ±0.01**	**19.76**	**34.81**	**59.50**	80.72	**20.27**	**32.54**	**70.91**	**93.15**

Table 2: **Top-1 and top-5 accuracy (%) under linear evaluation with different baseline invariant self-supervised learning (SSL) methods.** All methods use 2-view augmentation policy, and ImageNet-1k scores are averaged over 3 runs.

Algorithm	ImageNet-1k		ImageNet-Sketch		ImageNet-V2		ImageNet-R	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
MoCo-v3	68.44 ±0.07	88.02 ±0.04	17.65	31.87	56.54	78.68	18.59	30.08
+ Ours	**69.28 ±0.01**	**88.79 ±0.02**	**17.68**	**32.54**	**56.95**	**79.29**	**18.95**	**30.72**
DINO	67.37 ±0.02	87.55 ±0.01	17.13	32.09	55.00	77.38	18.28	30.38
+ Ours	**67.63 ±0.01**	**87.56 ±0.01**	**18.07**	**34.03**	**55.19**	**77.84**	**18.96**	**31.55**
Barlow Twins	63.34 ±0.03	84.3 ±0.04	10.90	21.17	47.69	70.73	12.30	20.94
+ Ours	**64.02 ±0.03**	**84.73 ±0.01**	**12.39**	**24.39**	**50.89**	**74.20**	**13.90**	**23.99**

our approach integrates seamlessly into existing SSL frameworks (MoCo-v3 (Chen et al., 2021), DINO (Caron et al., 2021), and Barlow Twins (Zbontar et al., 2021)), we preserve their original architectures and hyperparameters. Our modifications are limited to: (i) partitioning the mini-batch into subsets with one subset subjected to group transformations; (ii) adjusting the position of the `[CLS]` token to accommodate our equivariance objective whereas the other applies to the conventional augmentation including crop; and (iii) introducing the soft equivariance regularization constraint and its corresponding projection MLP layers. For all studies in this paper, we pre-trained ViT-S/16 with ImageNet using the AdamW optimizer. Similar to SimCLR (Chen et al., 2020), we pre-trained the network at batch size 2048 for 100 epochs with linear warmup for the first 10 epochs and decayed the learning rate using the cosine decay scheduler (without restart). For the linear evaluation protocol for ViT, we concatenated `[CLS]` tokens from the last 4 layers as an input to the linear classifier following (Devlin et al., 2019; Caron et al., 2021). A single linear layer is trained for 50 epochs with a cosine decaying learning schedule without a warmup, similar to Chen et al. (2020).

4.2 Main Results

Linear Evaluation. To assess the quality of the representation from our regularization constraint, we apply the linear evaluation method on the ImageNet-1k dataset. As shown in Table 1, we compare

Table 3: **Nonlinear evaluations using ImageNet-1k with different equivariant representation learning methods.** Note that all equivariant representation learning methods use MoCo as their baseline. View refers to the number of crops sampled per image.

View	Algorithm	MLP		20-NN	Fine-tune	
		Top-1	Top-5	Top-1	Top-1	Top-5
2 view	MoCo-v3	67.84	**88.37**	61.56	73.83	**91.91**
	+ AugSelf	63.24	85.86	60.63	73.50	91.67
	+ STL	65.74	86.32	57.34	73.90	91.49
	+ Ours	**68.04**	**88.37**	**61.64**	**74.33**	91.79
3 view	+ EquiMod	68.33	88.50	58.28	74.08	91.75
2+4 view	+ E-SSL	68.45	88.31	64.56	75.00	**92.36**
	+ Ours	**70.99**	**89.72**	**65.32**	**75.02**	92.12

the performance of our method with that of both implicit (E-SSL, AugSelf) and explicit methods (STL, EquiMod) as addressed in Section 4.1. Note that Equimod and E-SSL utilize three global views and two global with four additional local views, which can be denoted as a 2+4 view setting, respectively (Caron et al., 2020; 2021). Due to the discrepancies in the number of views used in differing equivariance algorithms, direct comparison in performance is undesirable as addressed in Section 2.1, and we therefore adopt our method to 2+4 setting, *i.e.*, 2 global and 4 local views, and report the performance in Table 1 to compare the performance with E-SSL. We omit reproducing our method for the three global views, as our method with 2 views already outperforms EquiMod in most scenarios. Note that in the conventional 2-view self-supervised learning setup, only our method scores higher than the baseline MoCo-v3, which may indicate that other methods to impose equivariance may have increased the equivariance at the last layer but sacrificed the downstream task performance. Other than Top-5 accuracy on ImageNet-v2, our method scores the highest on every side. Note that the parameter increment by adding our method on the conventional SSL method, i.e., MoCo, is marginal, because we do not have to train an additional module.

Generalizability to diverse SSL methods Though we have mainly used MoCo-v3 as a baseline SSL method due to its performance superiority over others, we evaluate the benefit of adding our method to diverse invariant SSL algorithms, i.e., DINO (Caron et al., 2021) and BarlowTwins (Zbontar et al., 2021), and show the results in Table 2. Note that our method always brings a performance increment when combined with diverse invariant SSL methods, as shown in Table 2.

Nonlinear Evaluation. Following the evaluation protocol from (Garrido et al., 2023), we also evaluate our learned representation in a nonlinear evaluation setting, i.e., 3-layer MLP as well as 20-nearest neighbour following (Caron et al., 2021). We also evaluate via finetuning the whole ViT encoder, and show the result at Table 3

4.3 Ablation and Analysis

One of the key contributions of our study is to encourage equivariance to the intermediate representation, while previously suggested approaches impose both invariance and equivariance at the last layer (Lee et al., 2021; Dangovski et al., 2021; Devillers & Lefort, 2023; Yu et al., 2024). Note that, in addition, we append the `[CLS]` not at the beginning but at the following layer of equivariance loss imposition as illustrated in Section 3.1. Figure 3 shows that there exists a sweet spot for both equivariance loss and `[CLS]`. When ablating `[CLS]` locations, we fixed the location to impose equivariance loss (layer 3).

Furthermore, we examine the relationship between the level of equivariance and the discrimination quality of the final representation when moving the equivariance loss layer closer to the final layer. As we shift the equivariance loss location towards the final layer, the equivariance score of various transformations at the final layer increased, albeit at the expense of representation quality as described in Table 4. More details can be found in Section B.

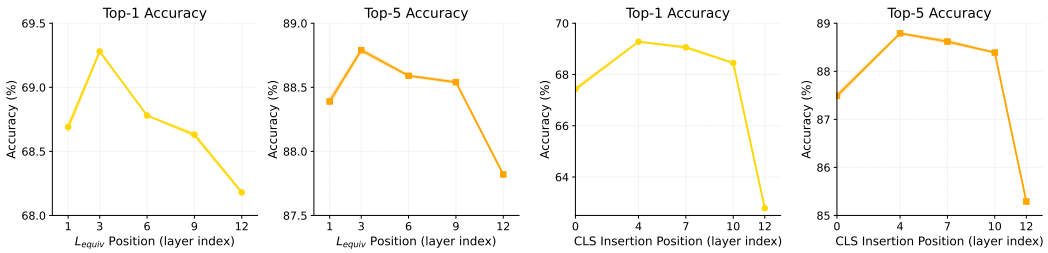

Figure 3: Ablation study on the location to regularize towards equivariance (left) and to insert the `[CLS]` token in the ViT encoder with fixed equivariance regularization layer at the 3rd layer (right). Both Top-1 (left) and Top-5 (right) accuracies peak when the equivariance loss and `[CLS]` is introduced near the middle of the network.

Table 4: **Layer-wise equivariance score and representation quality from learned representation.** Higher ↑ indicate greater equivariance, measured at the final representation layer. Regularizing equivariance at progressively later layers increases the equivariance score of the final representation, but at the cost of lower Top-1 accuracy, illustrating a trade-off between transformation sensitivity and discriminative power (see Section B for more detail)

Metric	**MoCo + Ours**			**MoCo + STL**	**MoCo + AugSelf**
	$\mathcal{L}_{\text{equiv}}$@layer3	$\mathcal{L}_{\text{equiv}}$@layer9	$\mathcal{L}_{\text{equiv}}$@layer12		
Top-1	69.21	68.72	68.18	67.58	64.98
Rotation ↑	0.840	0.873	0.875	0.731	0.997
H-Flip ↑	0.963	0.970	0.974	0.944	0.999
Scale ↑	0.937	0.946	0.946	0.915	0.999

5 Conclusion

In this paper, we have introduced a novel soft equivariance regularization framework that seamlessly integrates existing invariant self-supervised learning algorithms. Recognizing that purely invariant SSL methods may suppress valuable transformation-related information, our approach decouples invariance and equivariance by using standard SSL for invariant final representations and softly enforcing equivariance at intermediate layers. Our method avoids complexities like explicit transformation labels, additional modules, or auxiliary prediction tasks. Instead, we directly apply group actions as a soft regularization, preserving domain structure, preventing spatial collapse, and enhancing robustness against minor distortions. Empirical evaluations show our approach significantly improves downstream classification performance for ViTs pre-trained on ImageNet, effectively scaling to large datasets and consistently outperforming invariant baselines. We believe this strategy provides a simple yet effective means of incorporating equivariance into SSL, enhancing generalization and applicability for ViT architectures.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 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 (ICML)*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2990–2999. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/cohenc16.html.

Taco S. Cohen and Max Welling. Steerable cnns. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rJQKYt5ll.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv preprint arXiv:2111.00899*, 2021.

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.

Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve visual instance discrimination. In *International Conference on Learning Representations (ICLR)*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

M. Finzi, G. Benton, and A. G. Wilson. Residual pathway priors for soft equivariance constraints. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Q. Garrido, L. Najman, and Y. LeCun. Self-supervised learning of split invariant equivariant representations. *arXiv preprint arXiv:2302.10283*, 2023.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. *arXiv preprint arXiv:2210.02984*, 2022.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. *arXiv preprint arXiv:2306.13924*, 2023.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Shih-Cheng Huang, Anuj Pareek, Malte E. K. Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digit. Medicine*, 6, 2023. doi: 10.1038/S41746-023-00811-0. URL https://doi.org/10.1038/s41746-023-00811-0.

Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 7090–7099, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/ea9268cb43f55d1d12380fb6ea5bf572-Abstract.html.

H. Kim, H. Lee, H. Yang, and J. Lee. Regularizing towards soft equivariance under mixed symmetries. In *International Conference on Machine Learning (ICML)*, 2023.

Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Giovanni Luca Marchetti, Gustaf Tegnér, Anastasiia Varava, and Danica Kragic. Equivariant representation learning via class-pose decomposition. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent (eds.), *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4745–4756. PMLR, 2023. URL https://proceedings.mlr.press/v206/marchetti23b.html.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pp. 69–84. Springer, 2016. doi: 10.1007/978-3-319-46466-4_5. URL https://doi.org/10.1007/978-3-319-46466-4_5.

Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in neural information processing systems*, 33:4489–4500, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

J. Y. Park, O. Biza, L. Zhao, J. W. van de Meent, and R. Walters. Learning symmetric embeddings for equivariant world models. *arXiv preprint arXiv:2204.11371*, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae V2: scaling video masked autoencoders with dual masking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 14549–14560. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01398. URL https://doi.org/10.1109/CVPR52729.2023.01398.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

A. G. Wilson. Deep learning is not so mysterious or different. *arXiv preprint arXiv:2503.02113*, 2025.

Jaemyung Yu, Jaehyun Choi, DongJae Lee, HyeongGwon Hong, and Junmo Kim. Self-supervised transformation learning for equivariant representations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL http://proceedings.mlr.press/v139/zbontar21a.html.

Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pp. 7324–7334. PMLR, 2019.

A ALGORITHM

Algorithm 1 outlines our proposed soft equivariance regularization for invariant self-supervised learning, employing consistent notation with Figure 1.

Algorithm 1 Soft Equivariance Regularization for Invariant Self-Supervised Learning

Input: Batch B, partition ratio r, SSL augmentation $\mathcal{T}$, equivariant group $\mathcal{G}$, encoder $f = f^{(2)} \circ f^{(1)}$, invariance distance $d(\cdot, \cdot)$, weight λ
Partition B into B_1 and B_2 where $|B_2| = r|B|$ and $|B_1| = (1-r)|B|$
// Invariance-only path on B_1
Initialize $\mathcal{L}_{\text{inv}_1} \leftarrow 0$
for each $x \in B_1$ **do**
    Sample two views $t_1, t_2 \sim \mathcal{T}$
    Compute invariance loss:

$$\mathcal{L}_{\text{inv}_1} \leftarrow \mathcal{L}_{\text{inv}_1} + d\big(f(t_1(x)), f(t_2(x))\big)$$

end for
// Joint invariance and equivariance on B_2
Initialize $\mathcal{L}_{\text{inv}_2} \leftarrow 0$, $\mathcal{L}_{\text{equiv}} \leftarrow 0$
for each $x \in B_2$ **do**
    Sample two views $g_1, g_2 \sim \mathcal{G}$
    Extract intermediate features $z_1 = f^{(1)}(g_1(x))$ and $z_2 = f^{(1)}(g_2(x))$
    Compute invariance loss:

$$\mathcal{L}_{\text{inv}_2} \leftarrow \mathcal{L}_{\text{inv}_2} + d\big(f^{(2)}(z_1), f^{(2)}(z_2)\big)$$

    Apply group action to intermediate features: $\hat{z_1} = \rho_g(z_1), \quad g = g_2 g_1^{-1}$
    Update equivariance loss:

$$\mathcal{L}_{\text{equiv}} \leftarrow \mathcal{L}_{\text{equiv}} + d\big(\hat{z_1}, z_2\big)$$

end for
// Combine losses
Total loss:

$$\mathcal{L} \leftarrow \mathcal{L}_{\text{inv}_1} + \mathcal{L}_{\text{inv}_2} + \lambda\, \mathcal{L}_{\text{equiv}}$$

Update encoder parameters by minimizing $\mathcal{L}$
Output: Pre-trained model parameters

B FURTHER DISCUSSIONS AND EXPERIMENTS

B.1 DIVERSE NUMBER OF AUGMENTATION

It is well established that increasing the number of global or local views (augmentations) improves representational quality, albeit with additional computational cost (Caron et al., 2020; 2021). Hence, comparing algorithms that use different numbers of augmentations can lead to unfair evaluations. In particular, direct comparisons between E-SSL, Equimod, and other equivariance-based methods are misleading, as E-SSL relies on a 2+4-view strategy (2 global and 4 local views), while Equimod employs a 3-view strategy (3 global views). To ensure fairness, we also implemented our method under the 2+4-view setting. Specifically, following the "local-to-global" design from DINO (Caron et al., 2021), we do not pass all four local views through the MoCo momentum encoder, avoiding loss computation among local views. For the equivariance loss, we form one global pair and two local pairs, with losses computed only within each pair. The results of this 2+4-view variant are reported separately under the "2+4-view" row in the tables.

Table 5: **Transformation prediction.** Evaluation of transformation label prediction from the learned representation of different layers (see Section B for more detail)

Tasks	Methods	Layer 1	Layer 3	Layer 6	Layer 9	Layer 12
Rotation Prediction (%)	MoCo + Ours	79.97	93.34	99.52	99.73	96.59
	MoCo + STL	79.56	92.8	99.46	99.76	98.08
	MoCo + augself	81.55	96.71	99.71	99.13	68.74
HFlip Prediction (%)	MoCo + Ours	63.02	78.68	92.87	96.59	75.95
	MoCo + STL	62.91	82.09	88.83	97.22	83.27
	MoCo + augself	63.24	85.66	96.48	91.51	62.68

B.2 Ablation Studies

In Section 4.3, we examine the current practice of imposing equivariance loss concurrently with invariance loss at the encoder's final layer. Our results show that applying equivariance loss either too early or too late leads to suboptimal downstream performance. Instead, peak accuracy is achieved by applying equivariance regularization at an intermediate stage (in our study, we found the third layer to be optimal); by using the intermediate representation, equivariance can avoid conflict with invariance loss and can be facilitated with group action to operate as an objective function. Similarly, we observe that the insertion of the `[CLS]` token critically affects the effectiveness of equivariance regularization as described in Figure 3. Early insertion can impede the ability of the model to learn equivariant representations at intermediate layers. Conversely, inserting the `[CLS]` token too late deteriorates the ability to learn invariance.

In Table 4, we examined the changes when shifting the equivariance loss layer closer to the final representation. Here, we describe how we measure the equivariance score. Specifically, because our method leverages the token features instead of `[CLS]`, we measured the equivariance score in a similar manner. Following (Zhang, 2019), we sampled the Transformation parameter from Rotation90°, horizontal flip, and scaling, and measured the equivariance by computing the following:

$$\text{Equivariance} = \mathbb{E}_{x,\,(g_1,g_2)\sim\mathcal{G}}\left[d\left(\rho_g(f(x)),\ f(\rho_g(x))\right)\right], \quad g = g_2 g_1^{-1},$$

Note that we use cosine similarity for the distance function d and measure equivariance at the last layer, thereby replacing $f^{(1)}$ to f.

Furthermore, in Table 5, we measured the transformation label following the implementation from (Garrido et al., 2023). Note that this is a classification score instead of R^2 regression, e.g., HFlip is a binary classification task. Though our method trains the sensitivity towards transformation at mid-layer, its representation at late layers holds sensitivity towards transformation.

B.3 Object Detection

Equivariance is expected to be particularly beneficial for tasks requiring finer-grained spatial sensitivity than classification. To further examine the impact of equivariant regularization on transfer learning, we evaluate frozen-encoder object detection on the COCO dataset Lin et al. (2014). As illustrated in Table 6, our method achieves the highest detection accuracy across all metrics, indicating that equivariance regularization leads to more spatially informative representations, which transfer better to object detection than both invariance and prior equivariant baselines. Note that we did not aim to achieve a high score but to show that our approach benefits task that demands more spatial sensitivity than classification and outperforms other approaches, as in classification. Therefore, following the protocol of Oquab et al. (2023), we froze the encoder weights and only train the rest. We trained for 45000 iteration with a mini-batch size of 32. We trained with the COCO2017 train set and report the performance on the COCO2017 validation set. Importantly, all methods are trained under an identical setup, varying only the encoder weights.

B.4 Trivial Invariant Intermediate Representation

In this section, we explain that our method does not collapse to a trivial solution. Minimizing $\mathcal{L}_{\text{equiv}}$ corresponds to minimizing $d(\rho_g(f^{(1)}(x)), f^{(1)}(\rho_g(x)))$. First, trivial invariance does not result in

Table 6: COCO object-detection results with a frozen backbone (higher is better).

Metric	MoCo	MoCo + Ours	MoCo + STL	MoCo + AugSelf
mAP	0.225	**0.242**	0.221	0.197
mAP@50	0.404	**0.428**	0.400	0.359
mAP@75	0.222	**0.244**	0.218	0.192

Table 7: Top-1 accuracy comparison on ImageNet-C, including 15 types of common corruptions, for our method and other equivariant representation learning methods built upon the invariant representation learning baseline MoCo (He et al., 2020).

Algorithm	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defo.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Cont.	Elas.	Pixel	JPEG	
MoCo-v3	39.18	37.81	36.09	33.51	13.85	31.49	25.86	30.61	30.03	35.02	62.81	52.00	54.65	55.78	53.37	39.47
+ AugSelf	34.91	31.81	31.44	35.06	**17.12**	34.28	**27.67**	**31.99**	28.50	35.01	61.99	50.64	55.48	54.88	53.17	38.93
+ STL	17.78	16.26	14.65	29.51	15.13	27.33	25.77	29.50	27.60	34.40	61.81	48.84	54.31	46.40	50.63	33.33
+ Ours	**39.42**	**38.30**	**36.85**	**36.23**	15.91	**34.90**	27.35	30.71	**30.12**	**36.04**	**63.60**	**52.81**	**55.85**	**56.63**	**54.03**	**40.58**
+ EquiMod†	34.33	32.59	31.95	31.81	15.37	31.76	27.38	29.08	25.73	31.38	61.94	47.41	55.07	53.58	52.40	37.45
+ E-SSL‡	**43.80**	**42.59**	**40.80**	**38.44**	16.40	**36.73**	28.20	34.22	32.24	37.93	65.50	**55.89**	56.12	55.80	55.21	42.66
+ Ours‡	39.88	39.27	37.18	36.21	**19.58**	34.16	**31.90**	**36.07**	**35.87**	**40.74**	**66.76**	54.90	**57.87**	**56.42**	**56.79**	**42.91**

Table 8: Top-1 accuracy comparison on ImageNet-P, including 14 perturbation types, for our method and other equivariant representation learning methods built upon the invariant representation learning baseline MoCo (He et al., 2020).

Algorithm	Noise			Blur			Weather			Digital					Avg.
	Gau. N.	Shot	Speck.	Motion	Zoom	Gau. B.	Snow	Spatter	Bright.	Trans.	Rot.	Tilt	Scale	Shear	
MoCo-v3	67.85	67.85	67.97	57.31	68.30	68.28	56.57	66.57	63.43	68.05	64.55	67.58	45.18	65.32	63.91
+ AugSelf	67.44	67.47	67.47	57.51	67.76	67.81	56.77	66.15	60.77	67.39	64.41	67.10	**47.15**	64.96	63.58
+ STL	66.19	66.14	66.15	54.97	66.29	66.32	54.38	64.71	61.16	66.00	62.07	65.86	42.53	63.17	61.85
+ Ours	**68.97**	**69.01**	**69.00**	**59.09**	**69.46**	**69.35**	**58.02**	**67.86**	**64.58**	**69.04**	**65.76**	**68.60**	46.78	**66.34**	**65.13**
+ EquiMod†	68.60	68.70	68.78	57.08	69.17	69.13	56.97	67.51	60.88	68.75	65.18	68.27	47.04	66.17	64.44
+ E-SSL‡	70.26	70.19	70.22	**61.49**	70.65	70.54	60.60	68.87	65.82	70.27	66.92	69.82	48.86	67.59	66.58
+ Ours‡	**71.68**	**71.65**	**71.67**	60.36	**71.87**	**71.87**	**61.92**	**70.47**	**67.18**	**71.62**	**68.62**	**71.34**	**52.74**	**68.99**	**68.00**

zero loss, and therefore our equivariance loss is not collapsed toward trivial invariance; under invariance ($f^{(1)}(\rho_g(x)) = f^{(1)}(x)$), the loss simplifies to $d(\rho_g(f^{(1)}(x)), f^{(1)}(x))$, which is nonzero unless $f^{(1)}(x)$ is invariant under ρ_g (e.g., spatially constant map). Second, our contrastive L_{equiv} not only avoids model collapse, but it also promotes uniformity among negatives, which are sampled features from all positions of non-anchor images, encouraging uniformity on the hypersphere, thus preventing spatial constancy, as intra-image features must diversify to minimize the loss. Please refer to (Wang & Isola, 2020) for more details. Third, joint optimization with L_{inv} (e.g., MoCo) further promotes rich, non-constant representations to discriminate instances. Last, our method can predict the transformation information with a comparable accuracy to other equivariance algorithms, as shown in Table 5.

B.5 Latent Space Visualization

Beyond quantitative metrics, we also conduct additional qualitative analysis by comparing latent space features extracted from MoCo (trained with invariance loss alone) and MoCo + Ours. Due to ImageNet's large class count of 1000, we randomly sample 20 classes for analysis. As shown in Figures 4 and 5, we confirm that incorporating equivariance through our method also benefits downstream tasks that require invariance by promoting better class clustering; this provides novel evidence supporting our claim that equivariance and invariance layers should be decoupled.

Table 9: Experiments with various SSL algorithms. Top-1 accuracy (%) on **ImageNet-P**. All models are trained with the setting addressed in Section 4.1. See Table 8 for the results from MoCo.

Algorithm	Noise			Blur			Weather		Digital / Geometric						Avg.
	G.Nse	Shot	Spkl	Mot.	Zoom	G.Blr	Snow	Spat	Brt.	Tran	Rot	Tilt	Scal	Shear	
DINO	66.69	66.67	66.70	52.74	67.02	66.94	55.83	65.60	61.27	66.70	62.73	66.49	**42.96**	63.51	62.27
+ **Ours**	**67.39**	**67.31**	**67.41**	**54.41**	**67.69**	**67.66**	**56.19**	**66.06**	**62.13**	**67.22**	**63.33**	**67.02**	42.76	**64.31**	**62.92**
Barlow Twins	60.09	60.08	60.12	42.25	60.53	60.53	46.37	58.84	52.92	60.22	55.47	59.37	33.15	56.79	54.77
+ **Ours**	**63.85**	**63.93**	**63.91**	**49.18**	**64.29**	**64.19**	**50.09**	**62.34**	**57.93**	**63.89**	**59.63**	**63.38**	**37.29**	**60.56**	**58.89**

Table 10: Top-5 accuracy comparison on ImageNet-C, including 15 types of common corruptions, for our method and other equivariant representation learning methods built upon the invariant representation learning baseline MoCo (He et al., 2020).

Algorithm	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defo.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Cont.	Elas.	Pixel	JPEG	
MoCo-v3	63.30	61.68	59.76	56.22	28.46	53.52	45.98	52.52	50.54	58.64	84.47	76.74	77.38	79.44	77.76	61.76
+ AugSelf	58.96	55.21	54.61	59.06	34.52	57.55	**48.71**	**55.05**	49.48	59.30	84.39	76.26	78.25	79.15	78.02	61.90
+ STL	37.10	34.36	32.04	52.71	31.08	48.45	46.35	50.96	47.60	59.05	84.07	74.74	76.96	71.37	75.66	54.83
+ Ours	**64.16**	**62.70**	**61.21**	**59.92**	**32.28**	**58.17**	48.22	53.23	**51.26**	**60.49**	**85.68**	**77.94**	**78.55**	**80.52**	**78.58**	**63.53**
+ EquiMod†	58.63	56.28	55.34	55.28	31.48	54.45	48.20	51.11	45.88	55.49	84.63	73.66	78.51	78.46	77.62	60.33
+ E-SSL‡	**68.70**	**67.42**	**65.64**	**63.00**	33.09	**60.35**	49.83	57.64	53.72	62.83	86.80	**80.33**	79.05	**80.32**	80.05	**65.92**
+ Ours‡	64.00	62.89	60.60	60.01	**37.02**	56.65	**53.40**	**58.81**	**57.43**	**65.06**	**87.36**	79.14	**79.60**	79.82	**80.42**	65.48

Table 11: Top-5 accuracy comparison on ImageNet-P, including 14 perturbation types, for our method and other equivariant representation learning methods built upon the invariant representation learning baseline MoCo (He et al., 2020).

Algorithm	Noise			Blur			Weather			Digital					Avg.
	Gau. N.	Shot	Speck.	Motion	Zoom	Gau. B.	Snow	Spatter	Bright.	Trans.	Rot.	Tilt	Scale	Shear	
MoCo-v3	87.75	87.81	87.82	80.34	87.91	87.94	79.22	86.75	84.54	87.72	85.16	87.48	69.08	85.84	84.67
+ AugSelf	87.58	87.50	87.59	80.81	87.73	87.67	79.70	86.50	83.07	87.66	85.25	87.31	**71.44**	85.83	84.69
+ STL	86.67	86.65	86.66	78.33	86.82	86.77	77.62	85.59	83.21	86.50	83.81	86.50	66.31	84.75	83.30
+ Ours	**88.62**	**88.59**	**88.60**	**82.01**	**88.66**	**88.68**	**80.56**	**87.58**	**85.59**	**88.52**	**86.01**	**88.17**	71.25	**86.78**	**85.69**
+ EquiMod†	88.78	88.79	88.75	80.86	88.93	88.98	80.16	87.95	83.41	88.68	86.38	88.51	71.89	87.13	85.66
+ E-SSL‡	89.60	89.58	89.64	**83.94**	89.78	89.72	82.95	88.80	86.88	89.47	87.25	89.31	73.39	87.79	87.01
+ Ours‡	**89.96**	**89.93**	**90.00**	82.50	**90.10**	**90.08**	**83.23**	**89.15**	**87.30**	**89.93**	**87.80**	**89.68**	**75.80**	**88.29**	**87.41**

Table 12: **Computation overhead.** Measured FLOPs includes both forward and backward pass with a 2-view augmentation policy, and "Relative overhead" is the relative FLOPs to vanilla MoCo-v3. FLOPs for ours were measured for the overall mini-batch computation and divided by the mini-batch sample number (including both b1 and b2 as illustrated in Figure 2)

Method	Per-image FLOPs	Relative overhead
MoCo-v3	18.48G	1.0x
+ Ours	18.63G	1.008x

C LIMITATIONS

Our method significantly advances equivariant representation learning but faces key limitations. Primarily, it relies on structured geometric transformations, such as rotations, scaling, and flips, limiting its use to image-based tasks where these transformations are meaningful. Extending the approach to modalities without clearly defined transformations (*e.g.*, text, audio, graphs) is challenging. Second, despite scalability, the added regularization introduces computational overhead, particularly significant in large-scale or resource-limited environments.

D USE OF LARGE LANGUAGE MODELS

We used large language models (LLMs) to provide writing assistance during the preparation of this manuscript. The LLMs were used in the following ways:

- Polishing and rephrasing sentences for clarity and readability, including parts of the introduction, background, and experiments.
- Condensing text to meet page limits.

Importantly, the LLMs were not used for research ideation, experimental design, implementation, or result generation. All conceptual contributions, algorithm development, theoretical analysis, and experimental work were conceived, conducted, and verified entirely by the authors.

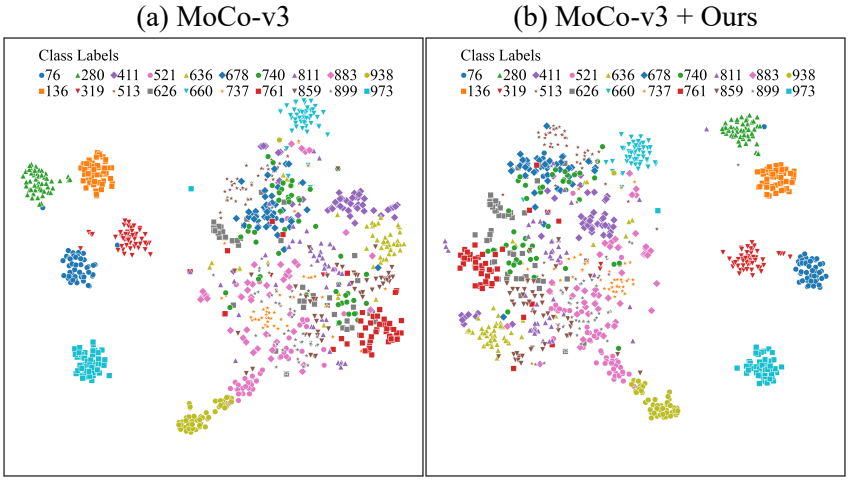

Figure 4: t-SNE visualization of latent space features from 20 randomly sampled ImageNet-1k classes, comparing (a) MoCo-v3 (trained with invariance loss alone) and (b) MoCo-v3 + Ours. Our method promotes better class clustering, demonstrating that incorporating equivariance benefits downstream tasks requiring invariance.

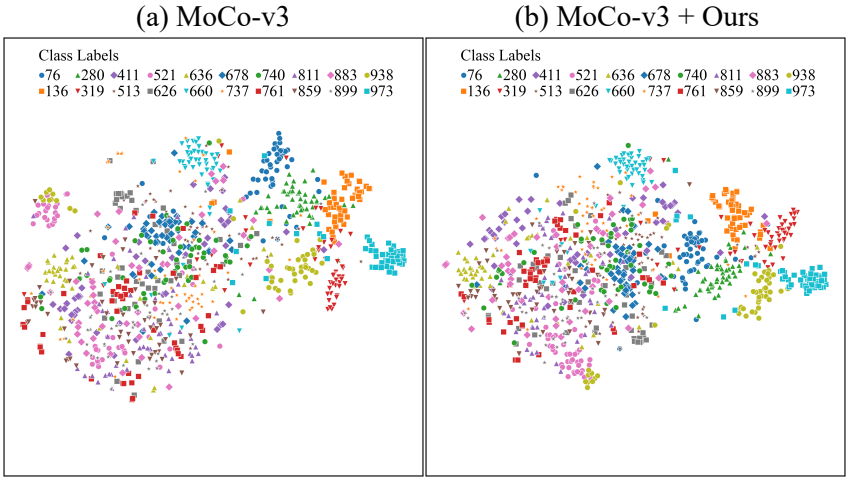

Figure 5: t-SNE visualization of latent space features from 20 randomly sampled ImageNet-C classes under defocus blur corruption, comparing (a) MoCo-v3 (trained with invariance loss alone) and (b) MoCo-v3 + Ours. Our method maintains better class clustering under corruption, demonstrating robustness benefits of incorporating equivariance.