Self-Supervised Learning Based on Transformed Image Reconstruction for Equivariance-Coherent Feature Representation

Qin Wang[1,3], **Alessio Quercia**[1,3], **Benjamin Bruns**[1],
Abigail Morrison[2,3], **Hanno Scharr**[1], **Kai Krajsek**[4]

[1]Institute for Advanced Simulation (IAS-8): Data Analytics and Machine Learning
Forschungszentrum Jülich GmbH, Jülich, Germany
[2]Institute for Advanced Simulation (IAS-6): Computational and Systems Neuroscience
Forschungszentrum Jülich GmbH, Jülich, Germany
[3]Software Engineering, Faculty of Computer Science, RWTH Aachen University, Aachen, Germany
[4]Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich GmbH, Jülich, Germany
{qi.wang, a.quercia, b.bruns, a.morrison, h.scharr, k.krajsek}@fz-juelich.de

Abstract

Self-supervised learning (SSL) methods have achieved remarkable success in learning image representations allowing invariances in them — but therefore discarding transformation information that some computer vision tasks actually require. While recent approaches attempt to address this limitation by learning equivariant features using linear operators in feature space, they impose restrictive assumptions that constrain flexibility and generalization. We introduce a weaker definition for the transformation relation between image and feature space denoted as equivariance-coherence. We propose a novel SSL auxillary task that learns equivariance-coherent representations through intermediate transformation reconstruction, which can be integrated with existing joint embedding SSL methods. Our key idea is to reconstruct images at intermediate points along transformation paths, e.g. when training on 30 rotations, we reconstruct the 10 and 20 rotation states. Reconstructing intermediate states requires the transformation information used in augmentations, rather than suppressing it, and therefore fosters features containing the augmented transformation information. Our method decomposes feature vectors into invariant and equivariant parts, training them with standard SSL losses and reconstruction losses, respectively. We demonstrate substantial improvements on synthetic equivariance benchmarks while maintaining competitive performance on downstream tasks requiring invariant representations. The approach seamlessly integrates with existing SSL methods (iBOT, DINOv2) and consistently enhances performance across diverse tasks, including segmentation, detection, depth estimation, and video dense prediction. Our framework provides a practical way for augmenting SSL methods with equivariant capabilities while preserving invariant performance.

1 Introduction

Self-supervised learning (SSL) methods (Chen et al. 2020a; He et al. 2020; Chen et al. 2020b; He et al. 2022b; Assran et al. 2023; Oquab et al. 2023; Zhou et al. 2022) have become crucial for pretraining foundation models by leveraging unannotated images for representation learning. However, joint embedding SSL methods, currently one of the best performing SSL methods, face a fundamental trade-off: they are designed around surrogate tasks that promote invariance, i.e. learning features that remain unchanged under input transformations. This invariance bias is reinforced due to the most popular evaluation being linear probing on ImageNet-1K (Deng et al. 2009), where classification labels are inherently invariant to transformations used in data augmentation.

While invariance is valuable for classification tasks, many computer vision applications require equivariant features that preserve transformation information rather than discarding it. Equivariance means that when an input undergoes a transformation (e.g. rotation, translation, scaling), the learned representation changes in a predictable, recoverable way. This property is essential for dense prediction tasks like object detection, segmentation, or pose estimation, where knowing precise object positions and orientations is critical for understanding where objects are in a scene. Recent methods like the Split Invariant–Equivariant (SIE) framework (Garrido, Najman, and LeCun 2023a) attempt to address this limitation by learning both invariant and equivariant features through transformation-conditioned linear operators in latent space. While effective, SIE imposes restrictive architectural constraints and assumes equivariant mappings to be linear. Such assumptions may be too strict for complex transformations or non-rigid deformations.

Our key idea is to use augmentation information in an auxiliary task, rather than modelling equivariance in latent space.

Our method reconstructs images at intermediate points along transformation paths. For example, when training on a 30 rotation, we reconstruct images at 10 and 20 rotation angles. This intermediate reconstruction pushes the model to provide features containing information on the transformation and thus encourages equivariant learning. Compared to SIE, our approach offers two key advantages: (1) it removes linearity constraints on equivariant mappings, broadening the function space for learning unseen or complex transformations, (2) it simplifies architecture by avoiding hypernetworks and operator prediction modules.

Empirical results in Table 1 demonstrate superior perfor-

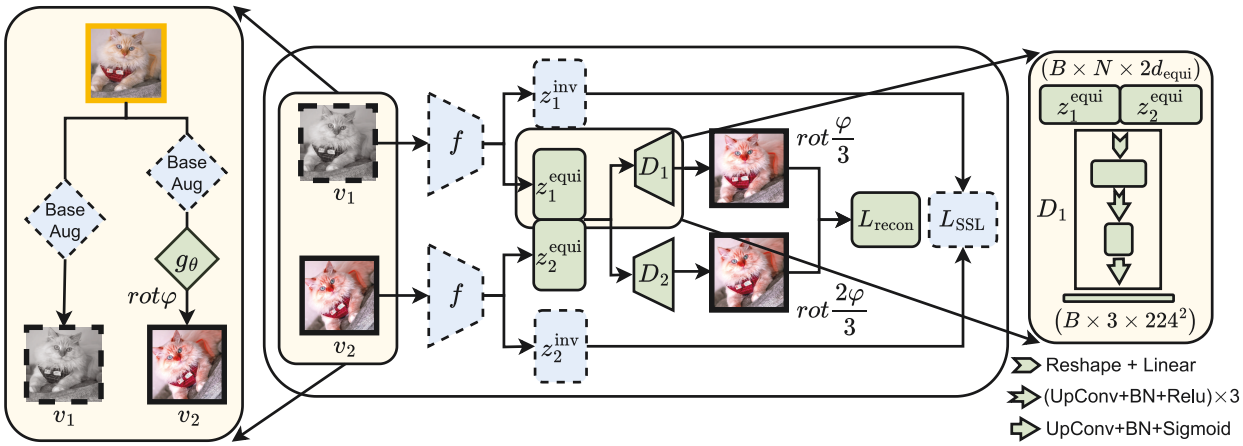

Figure 1: Overview of our framework for equivariance-coherent feature representation. As done in many joint embedding SSL methods like DINOv2, we apply two sets of image augmentations generating two different views of the original image. On the secondary view, we additionally apply a sequence of transformations g_i (e.g., three rotations by the angles $\frac{\varphi}{3}, \frac{2\varphi}{3}$ and φ as shown in the figure) to enforce equivariant structure during transformed image reconstruction.The last rotation serves as the second view for the joint embedding whereas the in-between transformed images are the targets of the reconstruction task.

mance of our reconstruction-based approach across multiple benchmarks, with strong generalization capabilities that transfer effectively to unseen transformations. Our contributions are:

- **New task for joint invariant-equivariant learning**: We introduce a reconstruction-based task that learns generalised equivariant features, relaxing restrictive assumptions about linear equivariant mappings.
- **Strong empirical validation on synthetic benchmarks**: Our method achieves superior performance on all synthetic equivariance tasks from (Wang, Krajsek, and Scharr 2024), demonstrating clear advantages over existing approaches including SIE.
- **Consistent improvements across diverse real-world tasks**: Compared to strong baselines (iBOT (Zhou et al. 2022), DINOv2 (Oquab et al. 2023)), our approach improves performance on most evaluation tasks while maintaining competitive results on invariance benchmarks.
- **General framework compatible with existing SSL methods**: Our intermediate reconstruction approach can be integrated with various SSL frameworks, allowing to enhance existing methods with equivariant capabilities.

2 Related Work

2.1 Equivariant Neural Networks

Since the early days of neural networks research, the exploration of symmetries in the data has played a significant role, reduced model complexity, and improved inference quality of models (Fukushima 1980). One might argue that without convolutional neural networks, inherently implementing approximately translational equivariance, computer vision models could not have made this progress in the field. However, built-in permutation equivariance in transformer architectures has also been the object of study (Xu et al. 2023). Equivariance in deep learning can be split in two sub-categories: studies on models that inherit built-in equivariance (Cohen and Welling 2016; Jenner and Weiler 2022) or models that gain this property by experience (Xiao et al. 2020; Dangovski et al. 2021; Wang et al. 2024).

2.2 Self-Supervised Learning

State-of-the-art SSL methods learn feature representations by automatically labelling non-labeled data and applying supervised learning techniques. The assumption is that the learned feature representation is comprehensive enough to be used later in other tasks, denoted as downstream tasks. A large variety of different SSL methods have been proposed (He et al. 2020; Chen et al. 2020a; He et al. 2022a; Chen et al. 2020b; Zbontar et al. 2021; Caron et al. 2021b; Bardes, Ponce, and LeCun 2022; Lehner et al. 2023; Xie et al. 2024; Oquab et al. 2023). For an overview, we refer to (Balestriero et al. 2023).

For our purpose it is relevant, how models react to different transformations in the input space, i.e. if they maintain the information in the feature representation or if this information is suppressed. Older SSL methods proposed auxiliary tasks such as Jigsaw puzzle (Noroozi and Favaro 2016) or rotation estimation (Gidaris, Singh, and Komodakis 2018) that foster equivariance properties as the transformation properties need to be maintained in feature space. These methods have been overtaken by matching-type methods. They present different versions of the same semantic content to the model and motivate it to map them to nearby points in feature space. Semantically different images are mapped to points far apart. This is achieved either by contrastive learning approaches or by regularisation techniques like de-correlation (Zbontar et al. 2021; Bardes, Ponce, and LeCun 2022) or teacher-student architectures (Zhou et al.

Configurations	R^2(rot)	R^2(color)	R^2(blur)	R^2(trans)	Mean(R^2)
SIE(rot)	0.990	0.867	0.042	0.540	0.610
SIE(color)	0.078	0.890	0.097	0.355	0.355
SIE(blur)	0.153	0.883	0.941	0.189	0.542
SIE(trans)	0.213	0.885	0.023	0.978	0.525
SIE(all)	0.331 ± 0.007	0.899 ± 0.003	0.211 ± 0.005	0.925 ± 0.002	0.592
Cross Atten Recon (2024)	0.893 ± 0.004	0.921 ± 0.006	0.823 ± 0.030	0.875 ± 0.005	0.878
Ours(VICReg, rot, rot)	0.9975 ± 0.0005	0.9073 ± 0.0021	0.9310 ± 0.0020	0.9810 ± 0.0010	0.954
Ours(VICReg, all, rot)	$\mathbf{0.9983 \pm 0.0005}$	0.9231 ± 0.0005	0.9689 ± 0.0099	0.9801 ± 0.0004	0.968
Ours(VICReg, all, color)	0.9891 ± 0.0019	$\mathbf{0.9373 \pm 0.0013}$	0.9700 ± 0.0067	0.9699 ± 0.0022	0.967
Ours(VICReg, all, blur)	0.9981 ± 0.0001	0.9154 ± 0.0006	0.9392 ± 0.0106	0.9774 ± 0.0007	0.958
Ours(VICReg, all, trans)	0.9975 ± 0.0005	0.9288 ± 0.0012	$\mathbf{0.9747 \pm 0.0017}$	$\mathbf{0.9830 \pm 0.0004}$	**0.971**
Ours(VICReg, all, SE(2))	0.9980 ± 0.0001	0.9158 ± 0.0005	0.9520 ± 0.0050	0.9740 ± 0.0001	0.960

Table 1: Comparison of R^2 values across different configurations for synthetic tasks. Configuration naming: SIE methods show training transformation in brackets. Our methods use format: Ours(SSL loss, augmentation, reconstruction target) where 'SSL loss' is the $\mathcal{L}_{SSL}$ function used, 'augmentation' is the DINOv2 common augmentation pipeline (rot = rotation only, all = all common augmentations), and 'reconstruction target' is the target transformation for $\mathcal{L}_{\text{recon}}$. **Bold** values indicate overall best. We use VICReg (Bardes, Ponce, and LeCun 2022) as the invariance loss, consistent with SIE's approach.

2022; Oquab et al. 2023). Consequently, these matching-type methods learn feature representations that suppress the differences between the versions.

Another state-of-the-art SSL branche includes mask-based approaches that remove parts of the input image and reconstruct them or a transformed version of them, either in the original image space (He et al. 2022a; Bandara et al. 2022), or in feature space (Assran et al. 2023). Contrastive learning has been combined with masked approaches, but only pixel-accurate translation has been applied as augmentation (Huang et al. 2022). As these methods do not, apart from masking or cropping, rely on other transformations, they are by construction more open for equivariance. Our approach is closely related to SIE (Garrido, Najman, and LeCun 2023a), combining the matching approach with an explicit model of transformations applied in the input space. Our approach can be seen as an extension not requiring knowledge about the transformation parameters.

In contrast to SIE, our approach reaches state-of-the-art results.

2.3 Equivariance vs. Invariance

A function f is denoted as equivariant with respect to a transformation t in the input space and a corresponding transformation $\hat{t}$ in the output space, if the function commutes with the transformations, i.e. $\hat{t}(f(x)) = f(t(x))$. Here, f is a deep learning model, input space is the space of images or videos, and output space is the feature space.

The definition includes the identity transformation in the output space such that invariance is always also equivariance. However, in the computer vision literature e.g. (Xiao et al. 2020; Dangovski et al. 2021; Devillers and Lefort 2022; Garrido, Najman, and LeCun 2023b; Park et al. 2022; Gupta et al. 2023; Wang et al. 2024), as we do in this paper, invariance is often opposed to equivariance to stress that all information about the transformation in the input space is still retrievable from the output space. The term equivariance is usually considered for a set of transformations, i.e. the function is said to be equivariant with respect to this set of transformations. Moreover, the set of transformations is considered a group transformation or, even stronger, a group representation of the transformation, i.e. a linear map, in the input and output space. However, not all transformations applied in computer vision can be modelled by group transformations like elastic distortions, crop-resize operations, or non-affine perspective transformations. In addition, transformations that can theoretically be formulated as group transformations might lose the corresponding group properties in practice. E.g. an image rotated around a general angle cannot be rotated back as parts of the image get lost during the forward transformation. In this paper, we do not restrict the model to be equivariant with transformations that belong to a certain further structure. We argue that equivariance is no value in itself but should serve as a property to help to learn a feature representation that contains all the information necessary for all kinds of downstream tasks. We do not restrict the transformation to form a group or require them to be linear in the feature space. Instead we motivate the model to maintain the information that is necessary to maintain input output relation such that transformed images in the input space can be retrieved by the representation. It shall be irrelevant if the transformation forms a linear transformation, a general group transformation, or even if the definition of equivariance is only approximately fulfilled in the feature representation. We denote this as equivariant-coherence.

3 Method

Our approach extends SSL frameworks with an auxiliary reconstruction task that learns equivariance-coherent features. The key innovation is intermediate reconstruction: rather than learning from just the original and final transformed images, we supervise the model to reconstruct images at multiple points along transformation trajectories. This design naturally encourages the model to retain information necessary to perform the considered transformations. Given transformation g_θ defined by a parameter vector $\theta = [\varphi; t_x; t_y; ...]$ of continuous parameters, we define K intermediate transformations $g_{\theta_1}, g_{\theta_2}, \ldots, g_{\theta_K}$ where $\theta_k := \frac{k\theta}{K+1}$ for $k \in$

$\{1, 2, \ldots, K\}$ are parameter vectors consisting of equidistant parameters values between no transformation besides the base transformation and the additional transformation defined by θ. In a first step to generate views as input for our joint embedding SSL approach a first view $v_1 = \mathcal{A}_1(I)$ is generated by means of a set $\mathcal{A}_1$ of augmentation transformations. The second view v_2 is generated by a second set $\mathcal{A}_2$ of augmentation transformations $u := \mathcal{A}_2(I)$ and, in contrast to the first view, it undergoes a sequence of additional transformations (the transformations are listed in Table 2)

$$u \to g_{\theta_1}(u) \to g_{\theta_2}(u) \to \cdots \to g_{\theta_K}(u) \to g_\theta(u) \quad (1)$$

where the end of the sequence constitutes the second view $v_2 := g_\theta(\mathcal{A}_2(I))$ in our joint embedding SSL method.

Intermediate images $u_k := g_{\theta_k}(u)$ are to be reconstructed by the SSL method acting as anchor points shaping the geometry of the feature space. We empirically determine that $K = 2$ yields optimal performance (Table 9).

3.1 Feature Splitting

In our proposed framework, depicted in Figure 1, we introduce transformation g as an additional augmentation applied exclusively to $u \in \mathbb{R}^{B\times3\times224\times224}$ where B denotes the batch size. Both views, v_1 and v_2, are processed through encoder f, which operates either with shared weights or within a student-teacher framework depending on the SSL method used. In the student-teacher setup, the teacher network is updated using an exponential moving average (EMA) of the student's parameters.

The resulting feature representations $z_i \in \mathbb{R}^{B\times N\times d_{\text{patch}}}$, $i \in \{1, 2\}$, are split into two complementary components along the feature dimension, where $d_{\text{patch}} = d_{\text{inv}} + d_{\text{equi}}$, N is the number of patches, and d_{patch} is the patch embedding dimension:

- **Invariant features** $z_i^{\text{inv}} \in \mathbb{R}^{B\times N\times d_{\text{inv}}}$ are supervised using standard SSL losses, e.g., the iBOT loss.
- **Equivariant features** $z_i^{\text{equi}} \in \mathbb{R}^{B\times N\times d_{\text{equi}}}$ are used to reconstruct intermediate transformed images.

The dimension of the equivariant feature component, denoted as d_{equi}, can be adjusted as needed. We perform sensitivity analyses on d_{equi} to examine its influence.

3.2 Intermediate Transformation Reconstruction

We employ K independent decoders, $D_1, D_2, \ldots, D_K$, that operate on the concatenated equivariant features $[z_1^{\text{equi}}; z_2^{\text{equi}}]$ to reconstruct intermediate transformed versions of the input image. To reduce computational cost and emphasize the encoder's role during pretraining, we deliberately design simple decoders consisting of a single linear layer followed by four convolutional layers (see Figure 1). Our objective is not to achieve high-fidelity reconstruction, but rather to provide sufficient supervisory signal for the encoder f to learn effective equivariant representations. The lightweight decoder design ensures that the primary learning occurs in the encoder without interference from complex reconstruction architectures. Each decoder produces reconstructions:

$$\hat{u}_k = D_k([z_1^{\text{equi}}; z_2^{\text{equi}}]) \quad \text{for } k = 1, 2, \ldots, K \quad (2)$$

Transformations g	**Parameters**	**Mag. range**
Rotation	angle φ	$[-45, 45]$
Color jittering	brightness, contrast,	$[-0.4, 0.4]$
	saturation S, hue H	$[-0.1, 0.1]$
Gaussian blur	radius σ	$[0.1, 2]$
Translation	displacement t_x, t_y	$[-10, 10]$
SE(2)	angle φ	$[-45, 45]$
	displacement t_x, t_y	$[-10, 10]$

Table 2: Considered transformations to generate the intermediate transformed images used as targets of the auxiliary reconstruction task, as well as the second view for the joint embedding.

The reconstruction loss $\mathcal{L}_{\text{recon}}$ is computed as the mean L_2 loss between each decoder's prediction $\hat{u}_k$ and the corresponding ground truth intermediate images u_k

$$\mathcal{L}_{\text{recon}} = \frac{1}{K}\sum_{k=1}^{K} \|\hat{u}_k - u_k\|_2^2 \quad (3)$$

This reconstruction objective is combined with the standard SSL loss $\mathcal{L}_{\text{SSL}}$ using a weighting hyperparameter λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SSL}}(z_1^{\text{inv}}, z_2^{\text{inv}}) + \lambda\mathcal{L}_{\text{recon}} \quad (4)$$

A sensitivity analysis for λ and d_{equi} is given in Section 4.5.

3.3 Transformation Types and Parameters

We introduce group transformations, approximate group transformation and non-group transformations in the reconstruction process to learn equivariant features. These group transformations include geometric transformations such as rotation, translation, and special Euclidean group transformations SE(2).

SE(2) defines the isometries in $\mathbb{R}^2$ that preserve the orientation, i.e. it combines 2d rotation and translation.

Additionally, we incorporate non-geometric transformations, including color jittering and Gaussian blur. Transformation parameter ranges are shown in Table 2.

4 Experimental Results

As our method builds on SIE (Garrido, Najman, and LeCun 2023a), SIE is our first natural baseline with synthetic evaluation tasks designed to benefit from equivariant features. To evaluate the generalization of our method against state-of-the-art approaches, we go beyond using the invariant loss function L_{SSL} as in SIE and explore integrating other methods. Specifically, we incorporate co-learning with iBOT (Zhou et al. 2022) and DINOv2 (Oquab et al. 2023), both augmentation-based techniques. They are also tested on the synthetic benchmark. Finally, we evaluate on a rich set of more realistic downstream tasks. We aim to enhance the baselines' performance on equivariance-related tasks while preserving strong results on invariance-related benchmarks, e.g. ImageNet linear probing.

4.1 Implementation Details

Architecture We use Vision Transformers (ViTs) (Dosovitskiy et al. 2020) with different configurations, specifically ViT-S/16 and ViT-L/16, as backbones for experiments. We incorporate a linear head on top of the backbone as originally done by the baseline methods to accommodate different representation dimensions d_{patch}, i.e. 8192 for iBOT, 512 for SIE, and 2048 for DINOv2.

Afterwards, a portion z^{equi} of these features z is allocated for reconstruction. We do not introduce more features than the baseline methods.

Pretraining Setup Our approach uses a baseline SSL loss L_{SSL} in addition to our new component L_{recon}. Each of the three baseline methods come with distinct training setups. The common training configuration includes ImageNet-1K as the dataset, optimizer AdamW (Loshchilov and Hutter 2017), and a cosine-scheduled learning rate. For the SIE-based method, we apply their invariant loss as L_{SSL}, and pretrain ViT-S/16 for 800 epochs with a batch size of 2048. The base learning rate is set to 10^{-4} and is linearly scaled with the batch size B: $lr = 10^{-4} \cdot B/256$. For the iBOT-based approach, we pretrain ViT-S/16 for 800 epochs and ViT-L/16 for 250 epochs, both with a batch size of 1024. The learning rate follows the same linear scaling strategy, with a base learning rate of $5e{-}4$. For the DINOv2-based training, we train ViT-L/16 with 100 epochs with batch size of 2048. The base learning rate is $4e{-}3$ with warmup 10 epochs. In the pretraining stage, we weight the reconstruction loss introduced by the auxiliary task using the optimal coefficient $\lambda = 1$, as determined by the sensitivity analysis in Section 4.5. For the equivariant feature dimension d_{equiv}, we also select the hyperparameter based on this sensitivity analysis, adopting a default value of 2048. For DINOv2, we retain the same proportional relationship between feature dimensions and therefore set $d_{equiv} = 512$. For VICReg, we follow SIE and evenly split the embedding dimension to determine d_{equiv}.

Computation Cost We conduct all pretraining experiments on the JUWELS Booster (Jülich Supercomputing Centre 2021). For VICReg, we adapt the method to a ViT-S/16 backbone and train on 4 nodes with a total of 16 A100 GPUs. Both iBOT and DINOv2 are pretrained on 16 nodes with 64 A100 GPUs. The computational cost of the different SSL methods, along with the additional cost introduced by our auxiliary tasks, is summarized in Table 3. Overall, our lightweight intermediate-transformation reconstruction task adds only a modest overhead to the baseline SSL methods.

4.2 Performance on Synthetic Tasks

Following (Wang, Krajsek, and Scharr 2024), we design synthetic tasks to evaluate the equivariant representations learned during pretraining, see Figure 2. These regression-based tasks assess the transformation parameters between original and transformed views. Following the evaluation metrics used in SIE (Garrido, Najman, and LeCun 2023a), we formulate transformation parameters prediction as a regression problem. To quantify alignment between predicted and true values, we use the coefficient of determination ($R^2 = 1 - \frac{RSS}{TSS}$), where RSS is the sum of squared residuals, and TSS is the total sum of squares. A higher R^2 value indicates more accurate transformation predictions.

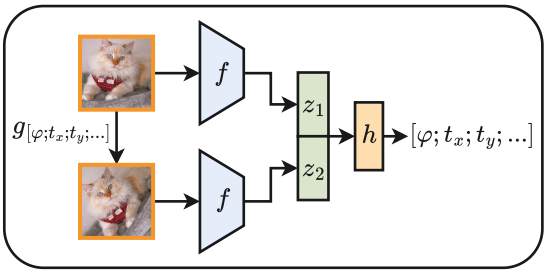

Figure 2: Synthetic tasks for evaluating equivariant representations. Transformation (g) is applied to the original image I. Both the transformed and original images are processed through a pretrained encoder f. A lightweight MLP h then predicts the parameters of the applied transformation.

Comparison with Other SIE-Based Methods Here, we use ViT-S/16 for all tests. Our approach adopts the SIE configuration and leverages VICReg (Bardes, Ponce, and LeCun 2022) as L_{SSL}, consistent with SIE (Garrido, Najman, and LeCun 2023a). We compare with SIE and a closely related cross-attention-based reconstruction method (Wang, Krajsek, and Scharr 2024), see Table 1. For SIE, the performance with single augmentations and prior knowledge is the best among the baseline methods, but for color jitter estimation, where all variants perform well. However, it generalizes poorly to other transformation evaluations (best mean $R^2 = 0.610$). Cross-attention reconstruction (Wang, Krajsek, and Scharr 2024) leads to much more balanced results (mean $R^2 = 0.878$). For our method, we tested different combinations of augmentation and intermediate transformation. All of them show a strong average performance increase (mean R^2 from 0.954 to 0.971).

The best ones, i.e. Ours(VICReg, all, rot) and Ours(VICReg, all, trans), outperform all competitors on all tasks individually, demonstrating the most versatile equivariant representations.

Enhancing Transformation Prediction of Augmentation-Based SSL Methods We test pretrained state-of-the-art

Config.	Backbone	Runtime s/epoch	Overhead
VICReg	ViT-S/16	156	-
+ Ours	ViT-S/16	179	14.7%
iBOT	ViT-L/16	230	-
+ Ours	ViT-L/16	249	8.3%
DINOv2	ViT-L/16	750	-
+ Ours	ViT-L/16	804	7.2%

Table 3: Computational overhead, where Ours refer to Ours(all, SE(2)), with all base augmentations and SE(2) transformation.

Configuration	R^2(rot)	R^2(color)	Mean(R^2)
SimCLR	0.288	0.575	0.432
+ Ours(all, rot)	0.637	0.712	0.675
+ Ours(all, SE(2))	0.705	0.723	0.714
IBOT	0.238	0.913	0.576
+ Ours(all, rot)	0.925	0.934	0.930
+ Ours(all, SE(2))	**0.937**	**0.943**	**0.940**
DINOv2	0.774	0.910	0.842
+ Ours(all, rot)	0.812	0.920	0.866
+ Ours(all, SE(2))	0.840	0.933	0.887

Table 4: Performance comparison on rotation and color prediction tasks with improvements of our methods (absolute values). See Table 1 for description of configuration names.

ViT-L/16 models on the same rotation and color jitter prediction task (as outlined in 4.2), see Table 4. The baseline iBOT (Zhou et al. 2022) pretrained model is taken from the official repository. For fair comparison, numbers shown for the DINOv2 baseline model are for a model we pretrained from scratch, as it performed better than with the weights from the official repository. For color prediction, iBOT and DINOv2 perform on par with the smaller ViT-S/16 models from Table 1. However, DINOv2 and iBOT perform significantly worse on the rotation estimation task. Training them from scratch using our approach improves their performance to best-in-class for color prediction (Ours(iBOT, all, SE(2)). For rotation prediction, improvements for iBOT are remarkably high from 0.238 to 0.937. DINOv2 performance is improved less and to a lower performance (0.84). Notably, using SE(2) as intermediate transform for iBOT and DINOv2 works slightly better than rotation, w.r.t. SIE, see Table 1.

4.3 Performance on Natural Images Tasks

We explore out method's impact on real-world imaging tasks commonly studied in self-supervised learning (SSL). We aim to be on par with state-of-the-art approaches, when downstream tasks are not to be expected to benefit from equivariance, like classification tasks. We strive to identify tasks where equivariant features are particularly beneficial.

Unless said differently, we use our method with iBOT and DINOv2 configurations performing best on the synthetic tasks from Table 4, i.e. Ours(iBOT, all, SE(2)) and Ours(DINOv2, all, SE(2)). Below, we call them Ours(iBOT) and Ours(DINOv2), respectively.

Linear Probing on Classification Tasks We follow the standard SSL evaluation pipeline, where the pretrained network is frozen, and only the linear head is fine-tuned on downstream tasks. The results, as shown in Table 5, are based on models pretrained on ImageNet1k. We report the performance using Top-1 accuracy, which measures the proportion of test samples for which the model's most confident prediction matches the ground-truth label.

Surprisingly, we found that our method improved performance compared to the baselines on average and across most datasets and tasks. Specifically, Ours(DINOv2) consistently achieved superior performance, with notable improvements in CIFAR10 (98.91% vs. 98.47%), CIFAR100 (90.37% vs. 89.28%), and Aircraft (71.64% vs. 70.89%), surpassing DINOv2 and other baselines. In contrast, only on the Food dataset our method fell slightly behind Ours(iBOT) (88.66% vs. 87.80%), which still represented a competitive result. Furthermore, SIE methods, which focus on equivariant features, did not perform well on natural image classification tasks (not shown). As a result, we focused on iBOT and DINOv2-related methods in later experiments.

Transfer Learning Tasks We investigate multiple downstream tasks that leverage equivariant features, including semantic segmentation, object detection, keypoint detection, homography estimation, monocular depth estimation, video object segmentation, semantic part propagation (Bhat et al. 2023; Quercia et al. 2025; Yang et al. 2024a,b; Pont-Tuset et al. 2017; Zhou et al. 2018). For semantic segmentation, we fine-tune our pretrained model using UPerNet (Xiao et al. 2018). For instance segmentation and object detection, we employ Mask R-CNN (He et al. 2017) with our pretrained model. For homography estimation, we design a 3-layer convolutional head to output the displacement map. For monocular depth estimation we fine-tune our pretrained models using the DepthAnything (Yang et al. 2024a) pipeline, based on ZoeDepth (Bhat et al. 2023). Lastly, for video object segmentation and semantic part propagation, we apply our pretrained models using the CropMAE (Eymaël et al. 2024) evaluation pipeline.

Table 6 shows our method's strong performance across diverse dense prediction tasks on standard benchmarks: semantic segmentation on ADE20k (Zhou et al. 2017), object detection and instance segmentation on COCO (Lin et al. 2015), keypoint detection on MPII (Andriluka et al. 2014), homography estimation on S-COCO (DeTone, Malisiewicz, and Rabinovich 2016), monocular depth estimation on NYU (Nathan Silberman and Fergus 2012) and KITTI (Geiger et al. 2013), video object segmentation on DAVIS 2017 (Pont-Tuset et al. 2017), and semantic part propagation on VIP (Zhou et al. 2018).

Our DINOv2-based approach consistently improves baseline methods across multiple tasks (but CoCo mAP, where it is on par). For semantic segmentation, we achieve notable improvement on ADE20k (54.12 vs. 53.49 mIoU compared to DINOv2). In homography estimation, our method excels with a Mean Corner Error of 1.42 on S-COCO, substantially better than both iBOT (1.76) and DINOv2 (1.68). For monocular depth estimation, our DINOv2 variant achieves the best RMSE and AbsRel scores on both NYU and KITTI datasets, while our iBOT variant improves on the original iBOT on NYU and matches its performance on KITTI.

For video tasks, our methods show strong improvements. Our iBOT-based approach achieves the best results on both DAVIS 2017 and VIP datasets, while our DINOv2 variant performs comparably to iBOT. Notably, the original DINOv2 performs poorly on DAVIS 2017, but our equivariant approach achieves reasonable performance levels.

These comprehensive results demonstrate that equivariant features effectively enhance performance across computer vision applications, providing consistent improvements over existing state-of-the-art SSL techniques.

Configuration or Method	CIFAR10	CIFAR100	Aircraft	Pet	Food	Flowers	INat18	ImageNet
iBOT	97.60	86.96	55.43	92.30	88.39	90.64	57.30	81.00
+Ours(all, SE(2))	98.08	87.36	57.55	94.34	**88.66**	96.03	57.99	81.44
DINOv2	98.47	89.28	70.89	94.82	87.92	96.39	69.42	82.60
+Ours(all, SE(2))	**98.91**	**90.37**	**71.64**	**95.42**	87.80	**96.81**	**70.41**	**82.73**

Table 5: Performance comparison on classification datasets given in percentage Top-1 accuracy. **Bold** values indicate the best performance across all methods for each dataset. Please see Table 1 for the naming convention.

Method	ADE20K	COCO			MPII	S-COCO ↓	NYU ↓		KITTI ↓	
	mIoU	AP^b	AP^m	mAP	PCKh	MCE	RMSE	AbsRel	RMSE	AbsRel
iBOT	52.17	0.5158	0.4448	0.7364	0.8697	1.76	0.3606	0.1001	2.7469	0.0672
+ Ours (all, SE(2))	52.38	0.5192	0.4478	0.7371	**0.8742**	1.53	0.3514	0.0985	2.7551	0.0671
DINOv2	53.49	0.5303	0.4574	**0.7512**	0.8728	1.68	0.3454	0.0965	2.7037	0.0664
+ Ours (all, SE(2))	**54.12**	**0.5332**	**0.4596**	0.7498	0.8736	**1.42**	**0.3413**	**0.0940**	**2.6578**	**0.0640**

Table 6: Performance comparison on dense prediction datasets. The symbol ↓ indicates that lower values are better. All the experiments are ran three times and report the mean value.

4.4 Comparison with Augmentation-Free Methods

We compare our approach with reconstruction-based self-supervised learning (SSL) methods that require minimal augmentations, such as MAE (He et al. 2022b), and those that require no augmentations, such as I-JEPA (Assran et al. 2023).

From Figure 3, we observe that all augmentation-based feature-matching methods (DINO (Caron et al. 2021a), DINOv2 (Oquab et al. 2023), MoCo (Chen et al. 2020b)) perform poorly on rotation prediction tasks, yielding worse results compared to reconstruction-based methods (MAE, I-JEPA). However, our approach enhances the performance of augmentation-based invariance matching methods on both tasks.

We see that the reconstruction-based methods in Figure 3 (MAE, I-JEPA) perform well on transformation prediction tasks. However, they are based on larger models such as ViT-H and ViT-G and/or pretraining on large-scale datasets like ImageNet-22K. In contrast, our approach uses ViT-S and still achieves results comparable to these larger models.

One limitation of reconstruction-based methods is their weaker performance on invariance-related tasks, as shown in Table 8.

When evaluated with linear probing, MAE and I-JEPA perform worse than or at best on par with augmentation-based methods iBOT and DINOv2, even though MAE and I-JEPA models are much larger. Our method improves slightly but consistently on the iBOT and DINOv2 baselines with the same model and training data.

We conclude that by incorporating transformation reconstruction, our method preserves equivariant representations like other reconstruction approaches, and even slightly outperforms augmentation-based methods on invariant tasks. Thus it combines the best of both worlds.

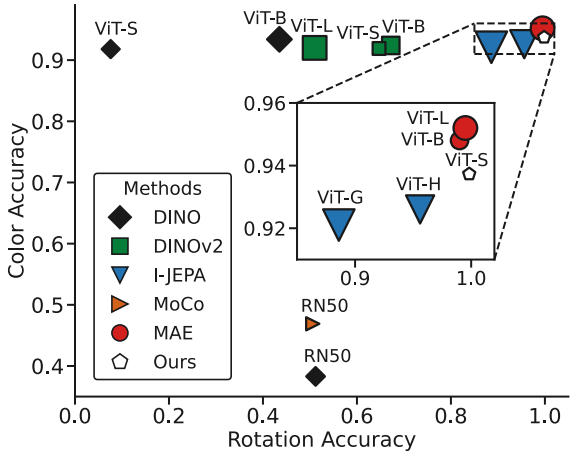

Figure 3: Synthetic tasks results among SSL methods

Method	DAVIS 2017			VIP
	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	mIoU
iBOT	63.4	62.3	64.6	40.7
+ Ours	**65.4**	63.9	**67.0**	**42.1**
DINOv2	29.8	27.7	31.9	40.5
+ Ours	65.2	**64.0**	66.3	41.5

Table 7: Performance comparison on the DAVIS and the VIP datasets for video object segmentation. "Ours" corresponds to Ours(all, SE(2)) using all base augmentations and the SE(2) transformation. Results are averaged over three runs.

4.5 Sensitivity Analysis

Our method involves several hyper-parameters: the split dimension d_{equi} of z^{equi}, i.e. the the portion of the feature vector used to reconstruct the intermediate images u_k, the weighting hyper-parameter λ in (4) for the equivariance-coherent loss L_{recon} and the number K of intermediate images for reconstruction. We used iBOT and the small ViT-S/16 for this sensitivity analysis to minimize computational load. Specifically, for the split dimension d_{equi} and loss weight λ, we performed pretraining for 100 epochs.

Method	Pretrain	Arch.	CIFAR100	iNat18	IN
MAE†	IN1k	H/14	77.3	32.9	77.2
I-JEPA†	IN1k	H/14	87.5	47.6	77.5
I-JEPA†	IN22k	H/14	89.5	50.5	79.3
I-JEPA†	IN22k	G/16	89.5	55.3	-
iBOT‡	IN1k	L/16	87.0	57.3	81.0
DINOv2‡	IN1k	L/16	89.3	69.4	82.6
DINOv2†	LVD	g/14	**94.0**	**81.6**	**87.1**
Ours (iBOT)	IN1k	L/16	**87.8**	**58.0**	**81.6**
Ours (DINOv2)	IN1k	L/16	**90.4**	**70.4**	**82.7**

Table 8: Linear probing performance on invariance tasks compared to models requiring minimal or no data augmentation. † denotes results reported in I-JEPA (Assran et al. 2023); ‡ denotes our reproduced pretrained model using the publicly available source code. IN: ImageNet, LVD: LVD-142M, H/14: ViT-H/14, G/16: ViT-G/16, L/16: ViT-L/16, g/14: ViT-g/14

For transformation analysis, we extended pretraining to 800 epochs. We use SE(2) transformation if not said differently.

Equivariant Dimension d_{equi} and Loss Weight λ We selected all combinations from $\lambda \in \{0.1, 1.0, 5.0\}$ and $d_{\text{equi}} \in \{256, 512, 1024, 2048, 4096\}$ and pretrained on ImageNet-1k as described above. We also tested $\lambda > 5.0$, and observed the training process to become unstable.

For the classification tasks, we provide in Figure 4 the mean and standard error of the accuracy for the hyper-parameter combination. For dense prediction tasks with different performance measures, we provide in Figure 5 the average rank for each hyper-parameter combination. For the classification tasks we observe for small $\lambda = 0.1$ and medium $\lambda = 1$ weighting parameter a rather stable behaviour of the mean accuracy around the baseline (dashed line) performance. For larger weighting parameter $\lambda = 1$ the performance decreases with larger portion of the feature vector used for the intermediate reconstruction task. The overall observation is reasonable, as classification tasks do not benefit from equivariance as dense prediction tasks do. Thus, increasing the weight parameter, i.e. focusing on the equivariant reconstruction task while giving less space portion of the feature vector for invariance, i.e. increasing d_{equi}, leads to a significant performance decrease. The qualitative behavior of our approach for the dense prediction tasks is different as, on one hand, our method outperforms the baseline for all hyper-parameters with respect to the average rank and on the other hand we observe an optimal parameter combination at $\lambda = 1$ and $d_{\text{equi}} = 2048$.

Number of Inbetween Images In Table 9, we investigate the effects of the number of in-between images. We observe that with only one in-between image, the performance is sub-optimal. Increasing the number of intermediate images to two significantly improves performance on synthetic tasks. Notably, adding more than two in-between images or incorporating the augmented view v_2 for reconstruction does not lead to further improvements. Considering both performance gains and GPU memory constraints, we select two in-between images for our experiments.

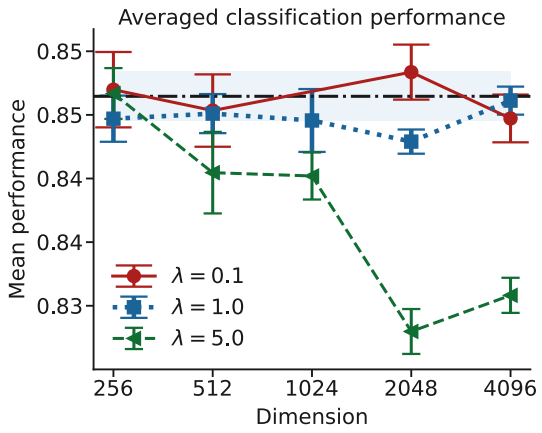

Figure 4: Mean performance across classification tasks.

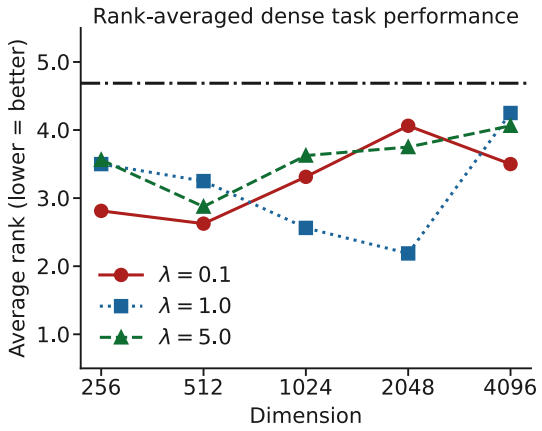

Figure 5: Mean performance across dense prediction tasks.

5 Summary and Conclusions

We propose a novel approach for augmentation-based SSL methods that enhances the learning of equivariant-coherent features by reconstructing intermediate representations of transformed images. Our method significantly boosts performance on equivariance-focused synthetic tasks and surpasses competitors like SIE. Moreover, we achieve comparable or superior results on real-world imaging tasks using iBOT and DINOv2 as base methods. This approach provides a promising direction for improving the generalization of SSL methods and can be easily adapted to other SSL frameworks.

# of Images	R^2(rot)	R^2(color)	R^2(blur)	R^2(trans)
1	0.2915	0.4708	0.8725	0.4230
2	**0.9983**	**0.9373**	**0.9689**	**0.9801**
3	0.9981	0.9215	0.9506	0.9795
2 + final	0.9981	0.9311	0.9562	0.9771

Table 9: Number of Inbetween images investigation. The model is Ours(VICReg, all, rot), cmp. Table 1.

Acknowledgments

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer JUWELS (Jülich Supercomputing Centre 2021) at Jülich Supercomputing Centre (JSC).

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 3686–3693.

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M. G.; LeCun, Y.; and Ballas, N. 2023. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15619–15629.

Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; Schwarzschild, A.; Wilson, A. G.; Geiping, J.; Garrido, Q.; Fernandez, P.; Bar, A.; Pirsiavash, H.; LeCun, Y.; and Goldblum, M. 2023. A Cookbook of Self-Supervised Learning. arXiv:2304.12210.

Bandara, W. G. C.; Patel, N.; Gholami, A.; Nikkhah, M.; Agrawal, M.; and Patel, V. M. 2022. AdaMAE: Adaptive Masking for Efficient Spatiotemporal Learning with Masked Autoencoders. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14507–14517.

Bardes, A.; Ponce, J.; and LeCun, Y. 2022. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. arXiv:2105.04906.

Bhat, S. F.; Birkl, R.; Wofk, D.; Wonka, P.; and Müller, M. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Caron, M.; Touvron, H.; Misra, I.; Jegou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021a. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630–9640.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021b. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*.

Cohen, T. S.; and Welling, M. 2016. Steerable CNNs. arXiv:1612.08498.

Dangovski, R.; Srivastava, R.; Cheung, B.; Agrawal, P.; and Soljačić. 2021. Equivariant Contrastive Learning. *ArXiv*, abs/2111.00899.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2016. Deep Image Homography Estimation. *ArXiv*, abs/1606.03798.

Devillers, A.; and Lefort, M. 2022. EquiMod: An Equivariance Module to Improve Visual Instance Discrimination. In *International Conference on Learning Representations*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Eymaël, A.; Vandeghen, R.; Cioppa, A.; Giancola, S.; Ghanem, B.; and Van Droogenbroeck, M. 2024. Efficient Image Pre-training with Siamese Cropped Masked Autoencoders. In *European Conference on Computer Vision (ECCV)*, 348–366.

Fukushima, K. 1980. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, 36: 193–202.

Garrido, Q.; Najman, L.; and LeCun, Y. 2023a. Self-Supervised Learning of Split Invariant Equivariant Representations. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Garrido, Q.; Najman, L.; and LeCun, Y. 2023b. Self-supervised learning of Split Invariant Equivariant representations. *ArXiv*, abs/2302.10283.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised Representation Learning by Predicting Image Rotations. *ArXiv*, abs/1803.07728.

Gupta, S.; Robinson, J.; Lim, D.; Villar, S.; and Jegelka, S. 2023. Structuring Representation Geometry with Rotationally Equivariant Contrastive Learning. *ArXiv*, abs/2306.13924.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollar, P.; and Girshick, R. 2022a. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 15979–15988. IEEE Computer Society. Publisher Copyright: © 2022 IEEE.; 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022 ; Conference date: 19-06-2022 Through 24-06-2022.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022b. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. arXiv:1911.05722.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Huang, Z.; Jin, X.; Lu, C.; Hou, Q.; Cheng, M.-M.; Fu, D.; Shen, X.; and Feng, J. 2022. Contrastive Masked Autoencoders are Stronger Vision Learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46: 2506–2517.

Jenner, E.; and Weiler, M. 2022. Steerable Partial Differential Operators for Equivariant Neural Networks. In *International Conference on Learning Representations*.

Jülich Supercomputing Centre. 2021. JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, 7(A183).

Lehner, J.; Alkin, B.; Fürst, A.; Rumetshofer, E.; Miklautz, L.; and Hochreiter, S. 2023. Contrastive Tuning: A Little Help to Make Masked Autoencoders Forget. arXiv:2304.10520.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Nathan Silberman, P. K., Derek Hoiem; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. *ArXiv*, abs/1603.09246.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. Q.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y. B.; Li, S.-W.; Misra, I.; Rabbat, M. G.; Sharma, V.; Synnaeve, G.; Xu, H.; Jégou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision. *ArXiv*, abs/2304.07193.

Park, J. Y.; Biza, O.; Zhao, L.; van de Meent, J.-W.; and Walters, R. 2022. Learning Symmetric Embeddings for Equivariant World Models. In *International Conference on Machine Learning*.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675*.

Quercia, A.; Yildiz, E.; Cao, Z.; Krajsek, K.; Morrison, A.; Assent, I.; and Scharr, H. 2025. Enhancing Monocular Depth Estimation with Multi-Source Auxiliary Tasks. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6435–6445. IEEE.

Wang, Q.; Krajsek, K.; and Scharr, H. 2024. Equivariant Representation Learning for Augmentation-based Self-Supervised Learning via Image Reconstruction. arXiv:2412.03314.

Wang, Y.; Hu, K.; Gupta, S.; Ye, Z.; Wang, Y.; and Jegelka, S. 2024. Understanding the Role of Equivariance in Self-supervised Learning. *ArXiv*, abs/2411.06508.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified Perceptual Parsing for Scene Understanding. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part V*, 432–448. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-01227-4.

Xiao, T.; Wang, X.; Efros, A. A.; and Darrell, T. 2020. What Should Not Be Contrastive in Contrastive Learning. *ArXiv*, abs/2008.05659.

Xie, J.; Lee, Y.; Chen, A. S.; and Finn, C. 2024. Self-Guided Masked Autoencoders for Domain-Agnostic Self-Supervised Learning. *ArXiv*, abs/2402.14789.

Xu, H.; Xiang, L.; Ye, H.; Yao, D.; Chu, P.; and Li, B. 2023. Permutation Equivariance of Transformers and its Applications. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5987–5996.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10371–10381.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. arXiv:2103.03230.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5122–5130.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. iBOT: Image BERT Pre-Training with Online Tokenizer. *International Conference on Learning Representations (ICLR)*.

Zhou, Q.; Liang, X.; Gong, K.; and Lin, L. 2018. Adaptive Temporal Encoding Network for Video Instance-level Human Parsing. In *Proc. of ACM International Conference on Multimedia (ACM MM)*.