Multi-Symmetry Ensembles: Improving Diversity and Generalization via Opposing Symmetries

Charlotte Loh [1 2] **Seungwook Han** [1] **Shivchander Sudalairaj** [2] **Rumen Dangovski** [1] **Kai Xu** [3] **Florian Wenzel** [4] **Marin Soljačić** [5] **Akash Srivastava** [2]

Abstract

Deep ensembles (DE) have been successful in improving model performance by learning diverse members via the stochasticity of random initialization. While recent works have attempted to promote further diversity in DE via hyperparameters or regularizing loss functions, these methods primarily still rely on a stochastic approach to explore the hypothesis space. In this work, we present Multi-Symmetry Ensembles (MSE), a framework for constructing diverse ensembles by capturing the multiplicity of hypotheses along symmetry axes, which explore the hypothesis space beyond stochastic perturbations of model weights and hyperparameters. We leverage recent advances in contrastive representation learning to create models that separately capture opposing hypotheses of invariant and equivariant functional classes and present a simple ensembling approach to efficiently combine appropriate hypotheses for a given task. We show that MSE effectively captures the multiplicity of conflicting hypotheses that is often required in large, diverse datasets like ImageNet. As a result of their inherent diversity, MSE improves classification performance, uncertainty quantification, and generalization across a series of transfer tasks. Our code is available at `https://github.com/clott3/multi-sym-ensem`

1. Introduction

The field of computer vision has seen significant progress in various tasks such as classification and semantic segmentation in recent years. This success can be attributed to the advancements in model architectures, learning methods, and the availability of large-scale datasets (Dosovitskiy et al., 2020; Sun et al., 2017; Chen et al., 2020). Large and diverse datasets have proved crucial in improving performance, yet they present new challenges. The increased diversity of datasets makes it more difficult for a single dominant hypothesis to capture all semantic classes. To overcome this problem, model ensembling (Hansen & Salamon, 1990; Breiman, 1996) can be utilized to combine multiple networks. A popular approach is Deep Ensembles (DE) (Lakshminarayanan et al., 2016), which combines networks with different random initializations and relies on the non-convexity of the loss landscape (Fort et al., 2019) and stochasticity of the training algorithm to arrive at different solutions. They often significantly improve model performance and uncertainty quantification (Ovadia et al., 2019).

Their success can be attributed to the diversity amongst the ensemble members (Rame & Cord, 2021); ensemble perfor-

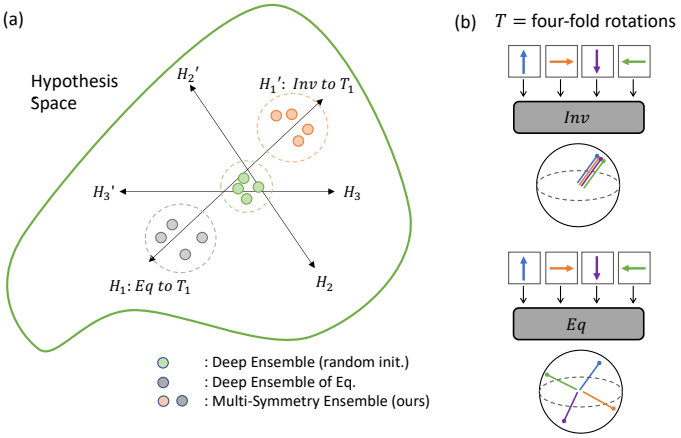

Figure 1. **(a)** A comparative illustration of the diversity in the hypothesis space that traditional deep ensembles and our Multi-Symmetry Ensembles can achieve. While deep ensembles are effective at capturing different solutions around one hypothesis, Multi-Symmetry Ensembles can learn diverse solutions around inherently opposing hypotheses. **(b)** Schematic visualization of invariance (top) v.s. equivariance (bottom) for the four-fold rotation. The spheres denote the representation space of the models.

[1]MIT EECS [2]MIT-IBM Watson AI Lab [3]Amazon (work done outside of Amazon) [4]AWS (work done outside of Amazon) [5]MIT Physics. Correspondence to: Charlotte Loh <cloh@mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

mance can be significantly improved relative to the individual models when the members are diverse and their errors are uncorrelated (i.e. when the members make mistakes on different samples). However, purely relying on the stochasticity in the random initialization and the training algorithm can only provide a limited amount of diversity (Rame & Cord, 2021) and previous works have attempted to promote diversity further by training models with different data augmentations (Stickland & Murray, 2020), hyperparameters (Wenzel et al., 2020), or explicitly encouraged via loss functions (Pang et al., 2019; Rame & Cord, 2021). Nonetheless, these methods primarily still rely on a stochastic approach to explore the hypothesis space.

In this work, we present a framework for constructing ensembles that are inherently diverse with respect to certain symmetry groups and thus in this regard, are *non-stochastic in exploring the hypothesis space*. We argue that current ensembling approaches are not effective in capturing the multiplicity of hypotheses, particularly along symmetry axes, which are necessary for large vision datasets. We motivate this with an intuitive example of rotational symmetry on the ImageNet (Deng et al., 2009) dataset. Recent works (Gidaris et al., 2018; Dangovski et al., 2021) have demonstrated the effectiveness of encoding rotational equivariance [1] on ImageNet. In this work, the term "equivariance" is used to explicitly refer to non-trivial equivariances, i.e. not encompassing the trivial instance, invariance (see footnote 1). Empirically, we found equivariance to be useful in images with a clear stance (e.g. dogs, where an upside-down dog is never observed in the dataset) and thus encoding information about its pose (i.e. rotation) aids their characterization. However, in large datasets like ImageNet, there also exist images like flowers that contain rotational symmetry and thus encoding rotational invariance, i.e. the removal of pose information, may be more desirable (see Figure 1b for an illustration of invariance versus equivariance).

Given the opposing nature of these hypotheses (see footnote 1 and Figure 1b), a stochastic ensembling approach cannot capture both simultaneously; i.e. a deep ensemble of rotational equivariant classifiers cannot be made rotational invariant by simply perturbing hyperparameters or model weights at initialization. We visually illustrate this point in Figure 1a. To address this problem, we leverage recent advances in contrastive representation learning (Chen et al., 2020; Dangovski et al., 2021) to create models that separately capture opposing invariant and equivariant hypotheses around a given symmetry group. In particular, in contrast to task-specific diversity promoting mechanisms of previous works (Pang et al., 2019; Rame & Cord, 2021), our approach aims to learn diverse representations that individually respect different symmetries and such task-agnosticity is desirable when transferring to new downstream tasks.

We present a practical, greedy ensembling approach that efficiently combines appropriate hypotheses for a given set of tasks. We provide extensive empirical results and analyses to demonstrate the superior performance of our method in classification performance, uncertainty quantification and transfer learning on new datasets. Our contributions can be summarized as follows:

- We empirically show that large, diverse datasets like ImageNet inherently have multiple and conflicting dominant hypotheses for classification.
- We propose Multi-Symmetry Ensemble (MSE), an ensembling method to train and combine models of opposing hypotheses with respect to certain symmetry groups. In contrast to previous works that rely on stochasticity created via random initializations or hyperparameters, we directly guide diversity exploration along the axes of symmetry.
- We demonstrate that MSE can leverage weaker models from the opposing hypothesis that improve performance more than the ensemble of higher-accuracy models corresponding to the leading hypothesis. To this end, we conduct a detailed empirical study to show that MSE improves classification performance and uncertainty quantification, and better generalizes across a series of transfer tasks.
- We also show that our method applies to different symmetry groups and that opposing hypotheses across multiple axes of symmetries further improve diversity.

2. Background and Related Work

Neural network ensembles and diversity. Using an ensemble of neural networks to improve performance and generalization is a well known technique in machine learning that existed decades ago (Hansen & Salamon, 1990; Breiman, 1996). Deep ensembles (Lakshminarayanan et al., 2016) create an ensemble of networks by using different random initializations, and Gal & Ghahramani (2015); Wen et al. (2020); Havasi et al. (2020) improve upon this by making it more computationally efficient. Diversity is an important feature in ensembles, since averaging many models that give the exact same prediction is no better than using a single model. Pang et al. (2019); Lee et al. (2016); Dvornik et al. (2019) create diversity by changing the losses or the architecture. Wenzel et al. (2020) create

[1] Equivariance can be best understood when contrasted with invariance – while invariance requires the outputs to be unchanged when the inputs are transformed, equivariance requires the outputs to transform *according to the way inputs are transformed*. While invariance is a trivial instance of equivariance (where T'_g in Equation (1) is the identity), in this work we use "equivariance" to refer specifically to non-trivial equivariances.

ensembles using different hyperparameters and Stickland & Murray (2020); Hendrycks et al. (2019) leverage data augmentation strategies. All of these methods rely on the stochasticity from the architecture, random initialization, or hyperparameters to generate different solutions. However, our work differs in that, we learn diverse solutions by leveraging opposing functional classes along certain symmetry groups. Moreover, the supervised learning settings do not focus on the transferability of representations, and therefore we propose an ensembling method that transfers better to a series of new downstream datasets. Along the line of learning diverse representations, Lopes et al. (2021); Wortsman et al. (2022) both conducted large-scale empirical studies of ensembling representations and models across architectures, training methods and datasets. Our work differs from these in that instead of a large-scale study of ensembling representations, our work introduces and focuses on a new technique of creating diversity in representations by using equivariances and invariances.

Contrastive Learning and augmentations. Contrastive representation learning (He et al., 2019; Chen et al., 2020) is an effective method for learning transferable representations with self-supervised learning. The role of augmentations in contrastive learning has been extensively studied (Chen et al., 2020; Tian et al., 2020; Xiao et al., 2020; Reed et al., 2021; Dangovski et al., 2021) with the objective of discovering useful augmentations to improve performance on downstream tasks. In contrast, our work takes a general approach of creating more robust classifiers by ensembling models of opposing equivariances. Xiao et al. (2020) designed a training objective that simultaneously computes a contrastive loss on a variety of projected representations, and each loss is associated with leaving out one augmentation from the complete set of augmentations. Our contribution is different from (Xiao et al., 2020), because we use equivariance, instead of removing augmentations. In addition, rather than a joint or concatenated latent space that is specialized to the removed augmentation, we create independent latent spaces and use an ensembling approach to accumulate the predictions of each member. A growing field of contributions introduce equivariance to models via self-supervised learning (Dangovski et al., 2021; Devillers & Lefort, 2022).

Equivariant neural networks. Let f be a continuous function (parameterized with an encoder network) and $\mathbf{x}$ be the input; equivariance to a group G of transformations is mathematically defined as

$$\forall \mathbf{x} : f(T_g(\mathbf{x})) = T'_g(f(\mathbf{x})) \tag{1}$$

where T_g denotes the transformation associated with a group element $g \in G$. In this formulation, invariance can be understood as a particular (trivial) instance where T'_g is the identity function, i.e. $f(T_g(\mathbf{x})) = f(\mathbf{x})$ and the output of the network does not change after a transformation to the input. Instead, equivariance requires the network output to change in a well-defined manner according to the way the input has been transformed. Intuitively, the difference between invariance and non-trivial equivariance can be understood as follows; while invariance encourages representations to remove information about the way they are transformed, non-trivial equivariance encourages the network to preserve this transformation information. This allows for a broader class of inductive biases that allows the model to make a decision on how to utilize this information during prediction.

Group equivariant neural networks (Cohen & Welling, 2016; Weiler & Cesa, 2019; Weiler et al., 2018) are usually designed by generalizing convolutional neural networks to arbitrary groups by constructing specialized kernels that satisfies the equivariance constraints. As such these networks usually require specialized architectures that are less commonly used on large-scale vision benchmarks. Dangovski et al. (2021) proposes a technique to encourage equivariance by using a prediction loss and show that approximate equivariance can be achieved by predicting the transformation. Along a similar line of work, (Zhang et al., 2016; Gidaris et al., 2018; Noroozi & Favaro, 2016) propose to learn visual representations by pretext tasks of predicting transformations. In our work, to avoid having specialized architectures and to keep the framework highly general and flexible, we adopt the method proposed in (Dangovski et al., 2021) to achieve *approximate equivariance* by a training objective that predicts the transformations applied to the input during the self-supervised learning stage. However, instead of learning better representations, our work focuses on the importance of creating ensembles containing members having opposing equivariances. Furthermore, (Dangovski et al., 2021) showed that rotational equivariance leads to better representations while rotational invariance is harmful; in this work, we show that while equivariance is useful for the majority of classes, there is a significant proportion of the data that can benefit from rotation invariance.

3. Multi-Symmetry Ensembles

We go beyond the typical deep ensembling approach by constructing ensembles that include opposing hypotheses along a set of symmetries. We start by pre-training representation learning models using contrastive learning methods (Chen et al., 2020; Dangovski et al., 2021). The pre-training step allows for encoding the necessary equivariances and invariances into the models. During fine-tuning, the pre-trained models are adapted into classifiers, and finally, these classifiers are combined into an ensemble. We demonstrate analytically in a simple setting via Proposition A.2 in Appendix A that the trained classifiers of the equivariant and invariant models capture different hypotheses.

3.1. Invariant and Equivariant Constrastive Learners

We now describe the paradigm to obtain the diverse ensemble members by inducing different equivariance and invariance constraints to the models. For ensemble member m, let $f_m(\cdot, \theta_m)$ denote the backbone encoder and $p_m(\cdot, \phi_m)$ the projector (here, a 3-layer MLP), parameterized by θ_m and ϕ_m respectively. Let T_{base} be the base set of transformations (e.g., RandomResizedCrop, ColorJitter). We realize the axis of symmetry through the transformations in SSL. Let T^m denote the transformation to which member m should be invariant or equivariant.

Contrastive learning operates by learning representations such that views of an image created via T_{base} are pulled closer together while pushed away from other images. In doing so, the model learns representations that are invariant to T_{base}. This is realized through the InfoNCE loss (Chen et al., 2020). Specifically, for a batch of B samples, the loss is

$$\mathcal{L}^m_{CL} = \sum_{i=1}^{B} -\log \frac{\exp(\hat{\mathbf{z}}^m_i \cdot \hat{\mathbf{z}}^m_j)/\tau}{\sum_{k \neq i} \exp(\hat{\mathbf{z}}^m_i \cdot \hat{\mathbf{z}}^m_k/\tau)} \tag{2}$$

where $\mathbf{z}^{\hat{m}}_i$ and $\mathbf{z}^{\hat{m}}_j$ are the ℓ_2-normalized representations of two views of an input $\mathbf{x}_i$ and $\mathbf{z}^{\hat{m}}_i = p_m \circ f_m(\mathbf{x}_i)/||p_m \circ f_m(\mathbf{x}_i)||$, and τ is a temperature hyperparameter.

Learning invariant models. Leveraging the contrastive learning framework, we learn an invariant model by adding T_m into the set of transformations, i.e. by optimizing the InfoNCE loss (Chen et al., 2020) with the augmentations set to $T = T_{base} \cup \{T_m\}$.

Learning equivariant models. We learn a model that is equivariant to T^m by initializing a separate prediction network $h_m(\cdot, \psi_m)$ and use a prediction loss as proposed in (Dangovski et al., 2021). Let G^m be a group to which member m is equivariant, i.e. its elements $g \in G^m$ transform the inputs/outputs according to Equation (1). The goal of $\mathcal{L}^m_{eq}$ is for the model to predict g from the representation $h_m \circ f_m(T_g(\mathbf{x}_i))$. By doing such, we encourage equivariance to G^m. In our work, we consider discrete and finite groups of image transformations (e.g., 4-fold rotations, color inversion (2-fold), and half-swaps (2-fold)). For discrete groups, $\mathcal{L}^m_{eq}$ takes the form of a cross-entropy loss,

$$\mathcal{L}^m_{eq} = \sum_{i=1}^{B} \sum_{g}^{|G|} H(h_m \circ f_m(T_g(\mathbf{x}_i)), g) \tag{3}$$

where H denotes the cross-entropy loss function and $|G|$ denotes the order or cardinality of the group, i.e. number of elements. As an example, for the group of 4-fold rotations, g takes on values in $\{0, 1, 2, 3\}$ corresponding to T_g in $\{0°, 90°, 180°, 270°\}$ rotation respectively. The sum over g is explained as follows; for every input, four versions are created for each of the 4 possible rotations and a cross-entropy loss is applied with their corresponding label in $\{0, 1, 2, 3\}$. The combined optimization objective of an equivariant model for a batch of B samples is $\mathcal{L} = \sum_{m=1}^{M} \mathcal{L}^m_{CL} + \lambda \mathcal{L}^m_{eq}$. Here, the InfoNCE loss $\mathcal{L}^m_{CL}$ encourages invariance only to T_{base}, i.e. T_m is not included in the set of augmentations.

Forming the ensemble. The contrastive pretraining step ensures that the representation learners have the appropriate equivariance and invariances. The next step is to convert these pretrained models into classifiers. This can be done using two methods: linear-probing or fine-tuning. Linear-probing involves training a logistic regression model to map the learned representations to the semantic classes while keeping the pretrained models frozen. Fine-tuning, on the other hand, allows the pretrained models to be updated during training, often resulting in higher accuracies on the same dataset. In this work, we always use fine-tuning to convert the pretrained models to classifiers unless specified otherwise. We propose two strategies for ensembling these classifiers: (1) *Random* and (2) *Greedy*. In both cases, we start by selecting a random model from the leading hypothesis and sequentially add models until the ensemble has M members.

(1) *Random*: MSE under the *Random* strategy alternates between the two functional classes at every stage, where a random model from that functional class is sampled without replacement, i.e. MSE always consist of models from both hypotheses. The baselines under the *Random* strategy is equivalent to randomly selecting M models.

(2) *Greedy*: The *Greedy* strategy is inspired by the approach of (Wenzel et al., 2020). At each stage, the best model is chosen based on the validation set score by searching over all models.

We compute the ensemble prediction $\bar{f}(\mathbf{x})$ by taking the mean of the member's *prediction probabilities* $\bar{f}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} f_i(\mathbf{x})$.

4. Experimental Setup

We use the standard ResNet-50 architecture for the backbone encoder and follow the experimental setup in (Dangovski et al., 2021). Our main results consider four-fold rotation transformation as the primary hypothesis class. All contrastive learning models were trained for 800 epochs with a batch size of 4096. For the equivariant models, h_m is a 3-layer MLP and λ is fixed to 0.4. Additional training details can be found in the Appendix.

Evaluation Protocol. After contrastive pre-training, we initialized a linear layer for each backbone and fine-tuned

them end-to-end for 100 epochs using the SGD optimizer with a cosine decay learning rate schedule. We conducted a grid search to optimize the learning rate hyperparameter for each downstream task.

Transfer tasks. We evaluated the transfer learning performance on 4 natural image datasets. Through these experiments, we evaluated the generalization performance of Multi-Symmetry Ensembles on new downstream tasks and to show how models with opposing hypotheses can contribute to meaningful diversity. For each dataset, we randomly initialized a linear classifier for each encoder pretrained on ImageNet and fine-tuned both the encoder and the linear head for 100 epochs. Following the approach in (Kornblith et al., 2018), we performed hyperparameter tuning for each model-dataset combination and selected the best hyperparameters using a validation set. For the iNaturalist-1K dataset (Van Horn et al., 2018), due to its large size and computational limitations, we used the linear evaluation protocol (Wu et al., 2018; van den Oord et al., 2018; Bachman et al., 2019; Wenzel et al., 2022) which involves training a linear classifier on top of a frozen encoder.

5. Results

In the following sections, we provide empirical evidence to support our claim that the diversity of opposing hypotheses along the symmetry axes improves ensemble performance, both in terms of model accuracy and generalization. We begin by demonstrating that both the invariant and equivariant hypotheses along the rotational symmetry tend to be equally dominant in large datasets like ImageNet. Next, we show that MSE, which incorporates these hypotheses, outperforms strong DE-based baselines that do not. We then provide an analysis of diversity and uncertainty quantification of MSE. In Section 5.5, we evaluate MSE on a set of transfer tasks. Finally, we study the impact of exploring opposing hypotheses along different symmetry groups on model performance.

Dominance of hypothesis are class-dependent. In Table 1, we compare two models f_{roteq} and f_{rotinv} that respectively have trained to be invariant (Inv) and equivariant (Eq) to four-fold rotation as contrastive learners. Even though the invariant model falls behind quite significantly from the equivariant model by 0.9% in the overall performance on ImageNet, in contrast to the observation from (Lopes et al., 2021; Mania et al., 2019), we found the dominance of a hypothesis to be highly class-dependent, as opposed to the leading hypothesis performing better uniformly across all classes. While the leading equivariant hypothesis dominates in 47.7% of ImageNet classes, the invariant still proves to be more useful in a significant 36.3% of the classes. We repeat this experiment for a number of large and small datasets, as shown in Figure 4, and found that large datasets tend to follow this trend.

Table 1. **Most suitable functional class differs within a dataset.** The top-half shows the overall accuracy for models from the SimCLR baseline and each of the opposing hypotheses wrt 4-fold rotations. The bottom-half shows the proportion of classes within each dataset where each hypotheses dominate (i.e. averaged over all samples within the class), suggesting that hypotheses apart from the one with the highest individual accuracy are still beneficial.

Model Accuracy on ImageNet (%)	
Baseline	76.5
Eq	76.9
Inv	76.0
Proportion of Classes (%)	
Eq $>$ Inv	47.7
Eq $<$ Inv	36.3
Eq $==$ Inv	16.0

Table 2. **Multi-Symmetry Ensembles capturing opposing hypothesis outperform naive ensembles of the same hypothesis.** The top-half of the table compares the acccuracy of naive ensemble of a single hypothesis and a random ensemble of both equivariant and invariant hypotheses. We show that as the number of members in the ensemble grow, capturing weaker performing models from the opposing hypothesis outperforms the naive-counterpart. The lower half of the table shows that, the gains are further amplified when the ensembles are chosen in a greedy manner.

	$M=2$	$M=3$	$M=4$	$M=5$
Random Ensemble				
Inv	77.4±0.0	78.0±0.1	78.4±0.0	78.5±0.0
Eq	78.2±0.1	78.7±0.1	78.9±0.1	79.1±0.0
Eq + Inv	78.2±0.1	78.8±0.0	79.1±0.1	79.3±0.1
Greedy Ensemble				
Inv	77.65±0.02	78.24±0.04	78.59±0.02	78.75±0.01
Eq	78.32±0.00	78.87±0.00	79.09±0.01	79.17±0.00
Eq + Inv	78.32±0.01	78.94±0.03	79.28±0.01	79.43±0.05

5.1. MSE captures meaningful diversity that leads to improved performance

We now compare deep ensembles (DE) constructed with models from the leading hypothesis (Eq) against MSE, which combines models from both hypotheses (Eq + Inv), as shown in Table 2 for ImageNet. Intuitively, given that Eq outperforms Inv significantly by 0.9%, one might expect to get larger gains by adding high-accuracy models from the leading hypothesis to the ensemble. Instead, we found ensembles involving lower-accuracy models from the opposing hypothesis to be better, with MSE (Eq + Inv) outperforming DE of rotational equivariant models (Eq)

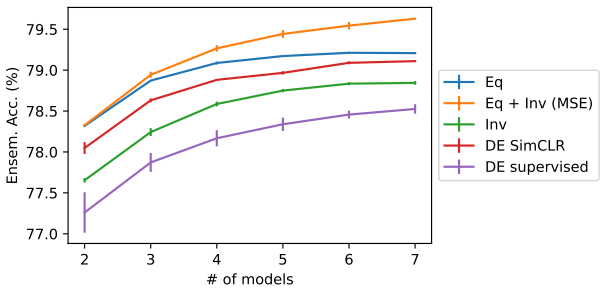

Figure 2. **Ensembles with opposing hypotheses have significantly larger potential.** Ensembles constructed only from a single hypothesis very quickly give marginal ensembling gains from adding more members. DE SimCLR and DE supervised refer to deep ensembles of baseline SimCLR (neither equivariant nor invariant) and supervised learning (without pretraining) respectively.

consistently across all ensemble sizes. Figure 2 further highlights the gap between the ensemble accuracy of $\mathrm{Eq} + \mathrm{Inv}$ and Eq. Ensembles constructed only from the leading hypothesis quickly result in marginal improvements gained from adding more members; by $M = 5$, the ensemble accuracy plateaus and does not benefit from further addition of more models. On the other hand, the ensemble accuracy of MSE demonstrates *greater potential and continues to benefit from increasing ensemble sizes*.

Greedy search finds alternating sequences. Interestingly, the outcome of the greedy search produces the following sequence of models: [Eq, Inv, Eq, Eq, Inv, Eq, Inv], that almost alternates between adding an equivariant and an invariant model at every step. This result suggests that in order to best maximize ensemble accuracy, it is ideal to construct ensembles that contain opposing hypotheses.

MSE's performance can be attributed to greater ensemble diversity. To further analyze the effectiveness of MSE ($\mathrm{Eq} + \mathrm{Inv}$) over the DE of Eq hypotheses, we evaluate their diversity on commonly used metrics, such as the error inconsistency (Lopes et al., 2021) between pairs of models, variance in predictions (Kendall & Gal, 2017) and pair-wise divergence measures (Fort et al., 2019) of the prediction distribution. We use error inconsistency as the main measure of diversity given its intuitive nature, which can be described as the fraction of samples where only one of the models makes the correct prediction, averaged over all possible pairs of models in the ensemble. Other diversity measures are defined in Appendix D.1. Ensemble diversity is an important criterion since higher ensembling performance is derived when individual models make mistakes on different samples. Table 3 demonstrates that by including models from opposing hypotheses, MSE indeed achieves a greater amount of diversity compared to the DE of Eq, consistently across all the diversity metrics.

Table 3. **Diversity of ensembles.** We compare the diversity across several metrics for ensembles with $M = 3$ members: error inconsistency, variance of the logits, variance of the probabilities and KL-divergence between pair-wise predictions. In all metrics, higher the score, greater the diversity.

	INCONS.(%)	LOGITS	PROB (10^{-4})	KL-DIV
INV	17.0±0.1	0.88±0.02	2.85±0.04	0.332±0.012
EQ	15.6±0.1	0.82±0.01	2.64±0.00	0.287±0.001
EQ + INV	17.5±0.1	0.94±0.01	2.94±0.00	0.359±0.007

Comparison between ensembling methods. Figure 3 further compares Multi-Symmetry Ensembles across some alternative methods to creating ensembles: ensembling models trained with supervised learning (Lakshminarayanan et al., 2016) (Sup), models that are separately fine-tuned with randomly initialized linear head but using the same pre-trained backbone (SSL_FT), models trained with the baseline SimCLR (Chen et al., 2020) (SSL), models trained with Equivariant SSL (Dangovski et al., 2021) (E_SSL) and models with opposing equivariance (E+I_SSL). Apart from E+I_SSL, all other methods create models from a single hypothesis. Unsurprisingly, SSL_FT produces ensembles with particularly poor diversity due to the limited variance between members since they differ only in the initalization of the linear heads. In general, the ensemble diversity is directly correlated with the ensemble efficiency (defined as the performance improvement relative to the mean accuracy of all the models in the ensemble (Lopes et al., 2021)). However, larger ensemble diversity does not necessarily lead to greater ensemble accuracy, since it is also important for the individual models to be high performing. This is evident in ensembles of supervised models – while they demonstrate high diversity and ensemble efficiency, their ensemble accuracy is poorer than their SSL counterparts since SSL produces higher performing models.

5.2. MSE can quantify uncertainty better but may require more models

A strong motivation to using an ensemble of models is it provides a way to quantify uncertainty from a Bayesian perspective (Wilson, 2020; Ovadia et al., 2019). To evaluate the quality of the ensembles' uncertainty estimates, we use the negative log likelihood (NLL), which is a proper scoring rule and a popular metric used to evaluate *predictive uncertainty* (Lakshminarayanan et al., 2016; Quiñonero-Candela et al., 2006). As seen from Table 4, for ensembles consisting of 3 members or more, Multi-Symmetry Ensembles built from opposing hypotheses (Eq + Inv) performs slightly better in terms of NLL when compared to ensembles with members only sampled from a single hypothesis (Eq). However, when there are very few members in the en-

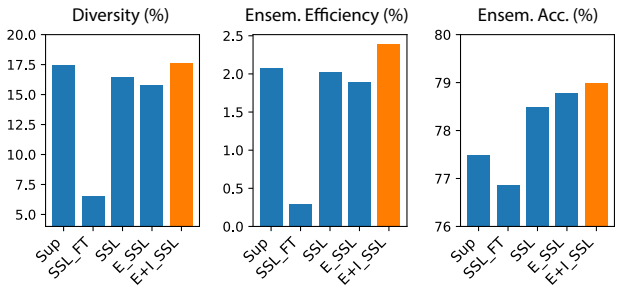

Figure 3. **Comparison between ensembling methods** for $M = 3$: supervised ensembles (Sup), models created from separate fine-tuning on the same backbone (SSL_FT), models pre-trained with SimCLR (Chen et al., 2020), models pre-trained with equivariant-SimCLR (Dangovski et al., 2021) (E_SSL) and Multi-Symmetry Ensembles of opposing hypotheses (E+I_SSL). Diversity is measured by pair-wise error inconsistency, ensemble efficiency is defined as the relative improvement over the mean accuracy of the members.

Table 4. **Uncertainty Quantification.** We evaluate the uncertainty quantification of our greedy ensembles, using the negative log likelihood loss (NLL) and the 'area under the uncertainty quantification curve' (AUUQC) which is obtained by sequentially removing the most uncertain samples and computing the area under the plot of ensemble accuracy versus fraction of samples removed. See Appendix G for results on our random ensembles.

	$M = 2$	$M = 3$	$M = 4$	$M = 5$
NLL ↓ (10^{-1})				
INV	8.77±0.03	8.49±0.01	8.35±0.00	8.26±0.00
EQ	8.40±0.00	8.18±0.00	8.06±0.00	7.99±0.00
EQ + INV	8.43±0.03	8.16±0.01	8.03±0.01	7.97±0.01
AUUQC ↑				
INV	0.919±0.000	0.924±0.000	0.925±0.000	0.926±0.000
EQ	0.921±0.000	0.925±0.000	0.927±0.000	0.928±0.000
EQ + INV	0.921±0.000	0.926±0.000	0.928±0.000	0.929±0.000

semble ($M = 2$), Multi-Symmetry Ensembles of opposing hypotheses performed slightly worse than single hypothesis according to the NLL metric. This is perhaps unsurprising, since the space of hypotheses is much larger with models from non-overlapping hypotheses and thus such an ensemble is likely to require more members in order to quantify the uncertainty surrounding each of these hypotheses. The NLL results show consistent trends the different ensembles.

To further evaluate the ensembles' ability to quantify *model uncertainty* (Gal, 2016), we also consider a different metric using an uncertainty-based prediction rejection setup, described as follows. We sequentially remove pools of test samples with the highest uncertainty from the ensemble and evaluate the ensemble accuracy on the remaining samples. This allows us to plot a curve of *fraction of samples removed* against *ensemble accuracy* which asymptotically approaches one when all samples are removed. An ensemble that "knows when it does not know" would produce a curve that is closer to the upper-left corner, since it can more accurately remove uncertain samples to give higher ensemble accuracies more quickly. We use the commonly used uncertainty measure BALD in the active learning framework (Gal et al., 2017; Houlsby et al., 2011), which is defined the information gained of the model parameters; see Appendix G.1 for a definition. Samples with large BALD would have the highest probability assigned to a different class on every stochastic forward pass (Gal et al., 2017) and thus have the highest model uncertainty. We compute and report the area under this curve and call it the "Area under the uncertainty quantification curve" (AUUQC) — higher AUUQC signifies better uncertainty quantification (see Figure 7 in Appendix G.2 for an illustration of this curve). Under this metric, we found that ensembles of opposing hypotheses (Eq + Inv) consistently outperforms ensembles of a single hypothesis across ensembles of different sizes.

Table 5. **Ensemble performance on transfer tasks using the greedy approach.** Ensemble efficiency is defined as the relative improvement from the mean accuracy of all the models in the ensemble. All experiments are fine-tuned except iNaturalist-1k which is linear-probed. Note that by construct of the greedy approach, Eq + Inv searches over possible Eq and Inv models and thus will be *at least as good as* Eq, i.e. datasets with equal performance for Eq and Eq + Inv do not benefit from the opposing hypothesis. See Appendix F for results on our random ensembles.

	iNATURALIST-1K	FLOWERS-102	CIFAR-100	FOOD-101
SINGLE MODEL ACCURACY				
EQ	55.1 ±0.3	91.9 ±0.0	85.5 ±0.1	87.9 ±0.1
INV	56.3 ±0.2	91.2 ±0.1	84.0 ±0.1	87.9 ±0.1
ENSEMBLE ACCURACY ($M = 2$) (ENSEMBLE EFFICIENCY)				
EQ	58.4 ±0.0 (3.3)	92.7 ±0.0 (0.8)	86.6 ±0.0 (1.1)	89.3 ±0.1 (1.4)
EQ + INV	59.9 ±0.2 (4.2)	93.1 ±0.1 (1.5)	86.6 ±0.1 (1.4)	89.5 ±0.1 (1.6)
ENSEMBLE ACCURACY ($M = 3$) (ENSEMBLE EFFICIENCY)				
EQ	59.8 ±0.0 (4.7)	92.9 ±0.1 (1.0)	87.1 ±0.0 (1.6)	89.9 ±0.0 (2.0)
EQ + INV	61.4 ±0.1 (5.5)	93.2 ±0.1 (1.4)	87.2 ±0.0 (1.8)	90.1 ±0.1 (2.2)

5.3. Different tasks may have different leading hypotheses and thus MSE transfers better

Another important axis to evaluate is the generalization of the learned representations in MSE. To this end, we conduct transfer learning experiments using pre-trained MSE on four downstream tasks. As shown in Table 5, MSE improves transfer performance in majority of the cases. In the largest and most diverse dataset iNaturalist-1K, we see consistent improvements of 1.5% and 1.6% from MSE in the respective cases of $M = 2$ and $M = 3$. Also, across

the four transfer tasks, it is evident that ensemble efficiency, the change in performance of the ensemble relative to the mean accuracy of the individual models in the ensemble, always improves significantly with our method except in one case. In Section 5.4, we further empirically analyze the circumstances under which our Multi-Symmetry Ensembles prove to be more useful. An interesting phenomenon to highlight in these results is that the dominant hypothesis can change depending on the downstream task. In the pre-training dataset (ImageNet), the equivariant model always proved to be the dominant hypothesis, outperforming the invariant model by 0.9%. However, after transfer learning on iNaturalist-1K, for example, the invariant model switches to become the dominant hypothesis, outperforming the equivariant model by 1.2%. This result emphasizes that different downstream tasks encompass different sets of hypotheses and therefore an ensemble of opposingly equivariant models can lead to better generalization.

5.4. Effectiveness of MSE depends on dataset diversity

In this section, we aim to provide empirical guidance on when the inclusion of opposing hypotheses in an ensemble is beneficial. We evaluate the proportion of classes dominated by each of the opposing hypotheses (invariant and equivariant symmetries) for different datasets, including iNaturalist-1k, CIFAR-100, ImageNet-V2, and ImageNet-R. These results are shown in Figure 4. Our findings indicate that on datasets such as iNaturalist-1k, the inclusion of opposing hypotheses in the ensemble improves performance. However, on datasets like CIFAR-100 and ImageNet-R, the opposing hypotheses do not provide significant gains. This is because these datasets have a high level of imbalance between the dominance of the two hypotheses, with one hypothesis dominating in a majority of the classes. For example, in ImageNet-R, the equivariant hypothesis dominates in 76.5% of the classes while the invariant hypothesis only dominates in 18% of classes. These datasets are poorly described by the opposing hypothesis and thus including them in the ensemble provides little to no improvement in performance.

5.5. Exploring different symmetry groups captures further meaningful diversity

So far, we have shown that capturing opposing hypothesis along the axis of rotational symmetry increases the diversity and performance of model ensembles. A natural question arises: can other symmetry groups be useful as well? Specifically, referring back to the illustration in Figure 1, is it sufficient to capture diversity around the opposing hypotheses of a single symmetry group, or would opposing hypotheses across symmetry groups further add meaningful diversity? To address this question, we conduct an ablation study with two additional transformations, half swap (random swapping of the upper and lower halves of an image) and color inversion (randomly inverting the color of an image). Due to computational limitations, we conduct this ablation study on ImageNet-100, a subset of ImageNet generated by randomly selecting 100 classes from ImageNet-1k and train the classifiers using linear-probing. This dataset contains about 129k samples and thus is still sufficiently diverse.

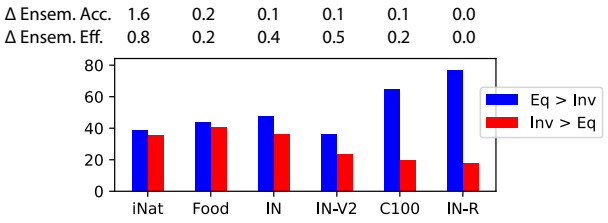

Figure 4. **Understanding the effectiveness of including the opposing hypothesis.** Plot shows the proportion of classes in each dataset where each hypothesis dominates. The remaining proportions (not shown) are classes where Eq and Inv are equally performant. Gains are minimal in datasets with a high level of imbalance between the leading and opposing hypothesis.

Table 6. **Capturing opposing hypotheses across transformations for $M = 6$.** The upper three rows are ensembles that consist of both equivariant and invariant learners with respect to a single transformation and the bottom row greedily searches over all models across the three transformations.

ENSEMBLE ACCURACY ON IMAGENET-100 (%)	
ROTATE	86.60
HALFSWAP	86.00
COLORINVERT	86.26
ROTATE + HALFSWAP + COLORINVERT	87.22

We present the results in Table 6. In the upper three rows, we create ensembles that consist of both equivariant and invariant learners with respect to a single axis of transformation. In the last row, we greedily search over the space of models that were trained across the three axes of transformations (rotation, half swap, and color inversion). By exploring multiple symmetry groups, we find additional diversity that improves the performance by up to 1.2%. This result bolsters the value of exploring multiple groups of opposing hypotheses and highlights the potential for future research directions to more effectively combining these models.

6. Conclusion and Limitations

In this work, we have showed that many large vision datasets benefit from a multiplicity of hypotheses, particularly along different axes of symmetries. To address this, we proposed to ensemble members from opposing hypotheses, disregarding the fact that models from the opposing hypothesis are significantly poorer performing. We showed that despite

their lower accuracies, ensembles containing the opposing hypotheses are meaningfully diverse and outperform current ensembling approaches of exploring the leading hypothesis class in multiple metrics of ensemble performance, ensemble potential, uncertainty quantification and generalization across transfer tasks.

While we explored a simple deep ensembling approach to combine multiple hypotheses, in principle one could also combine these hypotheses in alternative model combination approaches such as stacking (Wolpert, 1992) and mixture of experts (Riquelme et al., 2021; Mustafa et al., 2022; Allingham et al., 2022). Furthermore, since equivariance and invariance are invoked in the pre-training stage, the construction of these ensembles have higher computational costs compared to supervised deep ensembles that are trained from scratch (but on par with deep ensembles of contrastive learners), further work could look into more efficient methods to invoke equivariance and invariance during fine-tuning to mitigate this. Finally, while we found MSE to be highly effective in diverse, natural vision datasets, its effectiveness is dependent on dataset diversity (for e.g. less effective in ImageNet-R); we provide some intuition for these cases in Appendix B. We hope the findings from our work can motivate future research in these directions.

7. Acknowledgements

This work was sponsored in part by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

C.L. acknowledges fellowship support from the DSO National Laboratories, Singapore.

References

Allingham, J. U., Wenzel, F., Mariet, Z. E., Mustafa, B., Puigcerver, J., Houlsby, N., Jerfel, G., Fortuin, V., Lakshminarayanan, B., Snoek, J., Tran, D., Ruiz, C. R., and Jenatton, R. Sparse moes meet efficient ensembles. In *Transactions on Machine Learning Research*, 2022.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019.

Breiman, L. Bagging predictors. *Mach Learn*, 24:123–140, 1996. doi: https://doi.org/10.1007/BF00058655.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Dangovski, R., Jing, L., Loh, C., Han, S., Srivastava, A., Cheung, B., Agrawal, P., and Soljačić, M. Equivariant contrastive learning, 2021. URL https://arxiv.org/abs/2111.00899.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devillers, A. and Lefort, M. Equimod: An equivariance module to improve self-supervised learning, 2022. URL https://arxiv.org/abs/2211.01244.

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. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

Dvornik, N., Schmid, C., and Mairal, J. Diversity with cooperation: Ensemble methods for few-shot classification. *CoRR*, abs/1903.11341, 2019. URL http://arxiv.org/abs/1903.11341.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective, 2019. URL https://arxiv.org/abs/1912.02757.

Gal, Y. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2015. URL https://arxiv.org/abs/1506.02142.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data, 2017. URL https://arxiv.org/abs/1703.02910.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations, 2018. URL https://arxiv.org/abs/1803.07728.

Hansen, L. and Salamon, P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. doi: 10.1109/34.58871.

Havasi, M., Jenatton, R., Fort, S., Liu, J. Z., Snoek, J., Lakshminarayanan, B., Dai, A. M., and Tran, D. Training independent subnetworks for robust prediction, 2020. URL https://arxiv.org/abs/2010.06610.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning, 2019. URL https://arxiv.org/abs/1911.05722.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty, 2019. URL https://arxiv.org/abs/1912.02781.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning, 2011. URL https://arxiv.org/abs/1112.5745.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision?, 2017. URL https://arxiv.org/abs/1703.04977.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2018.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles, 2016. URL https://arxiv.org/abs/1612.01474.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, S., Purushwalkam, S., Cogswell, M., Ranjan, V., Crandall, D. J., and Batra, D. Stochastic multiple choice learning for training diverse deep ensembles. *CoRR*, abs/1606.07839, 2016. URL http://arxiv.org/abs/1606.07839.

Lopes, R. G., Dauphin, Y., and Cubuk, E. D. No one representation to rule them all: Overlapping features of training methods. *CoRR*, abs/2110.12899, 2021. URL https://arxiv.org/abs/2110.12899.

Mania, H., Miller, J., Schmidt, L., Hardt, M., and Recht, B. Model similarity mitigates test set overuse. *ArXiv*, abs/1905.12580, 2019.

Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., and Houlsby, N. Multimodal contrastive learning with limoe: the language-image mixture of experts, 2022. URL https://arxiv.org/abs/2206.02770.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles, 2016. URL https://arxiv.org/abs/1603.09246.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019. URL https://arxiv.org/abs/1906.02530.

Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. Improving adversarial robustness via promoting ensemble diversity, 2019. URL https://arxiv.org/abs/1901.08846.

Quiñonero-Candela, J., Rasmussen, C. E., Sinz, F., Bousquet, O., and Schölkopf, B. Evaluating predictive uncertainty challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F. (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 1–27, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33428-6.

Rame, A. and Cord, M. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation, 2021. URL https://arxiv.org/abs/2101.05544.

Reed, C., Metzger, S., Srinivas, A., Darrell, T., and Keutzer, K. Evaluating self-supervised pretraining without using labels. In *CVPR*, 2021.

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A. S., Keysers, D., and Houlsby, N. Scaling vision with sparse mixture of experts, 2021.

Stickland, A. C. and Murray, I. Diverse ensembles improve calibration, 2020. URL https://arxiv.org/abs/2007.04206.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. K. Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852, 2017.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Weiler, M. and Cesa, G. General $E(2)$-Equivariant Steerable CNNs. *arXiv:1911.08251*, November 2019. URL http://arxiv.org/abs/1911.08251.

Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. *arXiv:1807.02547*, October 2018. URL http://arxiv.org/abs/1807.02547.

Wen, Y., Tran, D., and Ba, J. Batchensemble: An alternative approach to efficient ensemble and lifelong learning. 2020. doi: 10.48550/ARXIV.2002.06715. URL https://arxiv.org/abs/2002.06715.

Wenzel, F., Snoek, J., Tran, D., and Jenatton, R. Hyperparameter ensembles for robustness and uncertainty quantification, 2020. URL https://arxiv.org/abs/2006.13570.

Wenzel, F., Dittadi, A., Gehler, P. V., Simon-Gabriel, C.-J., Horn, M., Zietlow, D., Kernert, D., Russell, C., Brox, T., Schiele, B., Schölkopf, B., and Locatello, F. Assaying out-of-distribution generalization in transfer learning. In *Neural Information Processing Systems*, 2022.

Wilson, A. G. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

Wolpert, D. H. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(05)80023-1. URL https://www.sciencedirect.com/science/article/pii/S0893608005800231.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. URL https://arxiv.org/abs/2203.05482.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Xiao, T., Wang, X., Efros, A. A., and Darrell, T. What should not be contrastive in contrastive learning. *CoRR*, abs/2008.05659, 2020. URL https://arxiv.org/abs/2008.05659.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization, 2016. URL https://arxiv.org/abs/1603.08511.

A. Formalism and Intuition

We show analytically that the functional classes of invariant and equivariant contrastive learners are different in a simple setting. As our assumptions are strong and simplistic, we only aim to provide intuition through a simple formalism. Our experiments in Section 5 support this intuition without the strong assumption and demonstrate the diversity from equivariance using real-world examples.

Assumption A.1. Consider a linear model class from (Kumar et al., 2022), which is $f_{v,B}(x) = v^T Bx$, where $B \in \mathbb{R}^{k\times d}$, is a linear encoder, $v \in \mathbb{R}^k$ is a linear head and $x \in \mathbb{R}^d$ is a datapoint. Consider an invariant model, $f_v^{\text{inv}}(x) := v^T Bx$, such that $BT_g(x) = Bx$ for every $g \in G$ and $x \in X$. Let $f_v^{\text{equiv}}(x) := v^T B'x$ be an equivariant model, such that $B'T_g(x) = T'_g(B'x)$ for all $g \in G$. Here we assume that B and B' are pretrained and fixed encoders and so we are only training v. Thus, we can represent $v \equiv v(B)$ as a function of the backbone B. Let $\tilde{X} = [X^T \mid T_g(X)^T]^T \in \mathbb{R}^{2n\times d}$ be our training input data for some $X = \{x_i\}_{i=1}^n \in \mathbb{R}^{n\times d}$, a fixed group element $g \in G$, and $T_g(X) := \{T_g(x_i)\}_{i=1}^n$. Assume the labels are $\tilde{y} = [y^T|y'^T]^T \in \mathbb{R}^{2n\times 1}$ where y are the labels for X and y' are the corresponding labels for $T_g(X)$. Here we assume that the data contains all input images from X and their transformation by T_g. Finally, assume an ordinary least squares (OLS) problem for learning v with $(\tilde{X}B^T, \tilde{y})$ training data for the invariant case and $(\tilde{X}B'^T, \tilde{y})$ for the equivariant.

Proposition A.2. *Under Assumption A.1, the solutions v^{inv} and v^{equiv} to the ordinary least squares problem for the corresponding f^{inv} and f^{equiv} with $(\tilde{X}B^T, \tilde{y})$ training data for the invariant case and $(\tilde{X}B'^T, \tilde{y})$ for the equivariant are:*

$$v^{\text{inv}}(B) = \frac{1}{2}(BX^TXB^T)^{-1}BX^T(y + y')$$

$$v^{\text{equiv}}(B') = (B'X^TXB'^T + T'_g(XB'^T)^T T'_g(XB'^T))^{-1}(BX^Ty + T'_g(XB'^T)^T y').$$

Proof. The proof is a simple combination of the OLS solution and the equivariance property. Namely, if the input data is A and the target is b, then the OLS solution is $(A^TA)^{-1}A^Tb$. Now, it suffices to replace the placeholder b with $\tilde{y}$, and the placeholder A with $\tilde{X}B'^T$ in the invariant case, and $\tilde{X}B'^T = [B'X^T|T'_g(XB'^T)^T]^T$ in the equivariant case. For the invariant case, we use the invariance property, which yields $T_g(X)B^T = XB^T$. For the equivariant case, we use the equivariance property, which yields us $T_g(X)B'^T = T'_g(XB'^T)$. Simplifying the algebra completes the proof. □

As functions of the pretraining backbones (B and B'), the two models in Assumption A.1 yield different functional classes (or hypotheses) as it can be seen by the forms of solutions in Proposition A.2. This analytical example provides us with further motivation to leverage on self-supervised models with opposing equivariances to capture diversity around multiple hypotheses. We choose to ensemble these different members instead of training one model because a single model cannot be simulataneouly invariant and equivariant to the same transformation due to conflicting objectives (i.e., a model cannot be invariant to a transformation and still change its representations according to the transformation).

B. More discussion towards Equivariance and Invariance

Comparison with Group Equivariant networks. The general notion of "Equivariance" typically emcompasses both non-trivial equivariance and invariance (i.e. trivial equivariance where T'_g of Equation (3) is the identity. For brevity and to maintain the convention used in (Dangovski et al., 2021), we however use the term "equivariance" to specifically refer only to non-trivial equivariance (i.e. excluding invariance) in our work. Equivariance in deep learning is most commonly known through the concept of Group Equivariant neural networks (Cohen & Welling, 2016; Weiler & Cesa, 2019; Weiler et al., 2018). There, non-trivial equivariance and invariance to a particular group are achieved through equivariant architectures, by generalizing convolutional kernels to respect the symmetries of that group. These are often implemented in the form of equivariant layers, where the trivial instance of invariance can be acheived by invoking a global pooling function after a series of equivariant layers. In our work, (non-trivial) equivariance and invariance to a particular transformation T_m are achieved purely via training objectives — invariance is achieved by adding T_m into the set of augmentations used in contrastive learning that encourages representations to be invariant to and equivariance is achieved by adding an auxiliary self-supervised task that predicts the transformation T_m applied to the input. The architecture we use for all models is a non-equivariant architecture, i.e. the common ResNet-50 model. In this setting and our definition of "equivariance" that refers only to non-trivial equivariance, a single model cannot be equivariant and invariant simultaneously and thus the two form a set of opposing hypotheses.

Empirical intuition. Equivariance to rotation has been known to be highly beneficial for learning visual representations (Gidaris et al., 2018; Dangovski et al., 2021), however the underlying reasons are not so clear. Empirically, we found the usefulness of rotation equivariance is generally related to pose or the existence of rotational symmetry in the dataset. We found that rotation equivariance is useful in image classes that often occur with a clear stance, for e.g. some classes of animals, where an upside-down dog is almost never observed in the dataset and thus the ability to recognize the rotation would require the features to encode information about its pose (Gidaris et al., 2018), aiding the characterization of dogs. On the other hand, we found that rotation invariance is useful in image classes that do not occur with a clear stance (for e.g. corkscrews that can be pictured in any orientation) or in images that have a clear rotation symmetry (e.g. flowers imaged from the front or analog clocks).

Empirical intuition on datasets where MSE are effective. In our work, we found the effectiveness of MSE to be highly dependent on dataset diversity. In particular, if the datasets are poorly described by the opposing hypothesis (i.e. ImageNet-R) as discussed in section 5.4, the gains from MSE would be negligible. Here, we provide some intuition on why this may be so. Following the intuition provided in the previous paragraph, we conjecture that this could be related to the existence of a dominant pose of images in the dataset. An example of the class of "jellyfish" in ImageNet (IN) and IN-R is shown in Figure 5. In IN-R which contains renditions of the images, such as in cartoon and art, many images assume a conventional "upright" pose of the jellyfish with its head on top and its tentacles trailing below vertically. However, in IN where real-life jellyfish are imaged, they often occur in multiple poses. We believe this is true for other classes as well, since artists often draw objects in their 'conventional pose'. Thus, for IN, invariant models are useful for 36.3% (v.s. equivariant models being useful in 47.7%). In contrast, for IN-R, invariant models are dominant only for 18% of the classes (v.s. equivariant models being dominant in 76.5%). Given the existence of an upright pose in IN-R, equivariant models that encode pose information are likely more useful than invariant models leading to this stark difference.

Figure 5. **Examples of images from the "jellyfish" class in ImageNet (left) and ImageNet-R (right).** Samples visualized using `https://knowyourdata-tfds.withgoogle.com/`

C. Additional Training Details

All pre-training. We use the SGD optimizer with a learning rate of 4.8 ($0.3 \times BatchSize/256$). We decay the learning rate with a cosine decay schedule without restarts. Following (Dangovski et al., 2021), T_{base} uses a slightly more optimal implementation that uses BYOL's augmentation (i.e. including solarization).

Equivariant pre-training. Following (Dangovski et al., 2021), the predictor for equivariance uses a smaller crop of 96×96. The predictor network uses a 3-layer MLP with a hidden dimension of 2048 to predict the corresponding

transformation (i.e. 4-way rotation).

Invariant pre-training. For invariant models, the transformation T_m is added to the base set of augmentations T_{base} with probability $p = 0.5$, i.e. with 0.5 probability, one of the possible transformations $(0°, 90°, 180°, 270°$ for the case of 4-fold rotations) are applied.

Explored hyperparameters for fine-tuning. For fine-tuning on ImageNet, we swept the learning rate ($lr \in \{0.1, 0.03, 0.01, 0.003, 0.004\}$ for both equivariant and invariant models. We found $lr = 0.003$ to consistently give the best performance for equivariant models and $lr = 0.004$ to consistently give the best performance for invariant models. For fine-tuning on transfer tasks, we swept the learning rate $lr \in \{0.003, 0.1, 0.2, 0.5, 1.0, 5.0\}$ for each equivariant/invariant model and picked the best learning rate. We set the weight-decay to 10^{-6} for all fine-tuning experiments.

D. Ensemble Diversity

D.1. Diversity measures

Error inconsistency. Following (Lopes et al., 2021), we use error inconsistency between pairs of models to quantify their diversity. For every sample and a pair of models, model A and model B, there are four possibilities: 1) both models are correct, 2) both models are wrong, 3) model A is correct and model B is wrong and 4) model B is correct and model A is wrong. Samples that fall into the case of (3) and (4) constitute to the error inconsistency. We report the percentage of samples in the test set that pairs of models make inconsistent errors on. For ensembles more than $M = 2$ members, we take the average over all possible pairs of models.

Variance of predictions. Another measure one could use to measure ensemble diversity is the variance of the predictions (Kendall & Gal, 2017):

$$\mathrm{Var}_{p(\mathbf{f})}[\mathbf{f}(\mathbf{x})] = \sum_{i=1}^{C} \mathrm{Var}_{p(\mathbf{f})}[f^{(i)}(\mathbf{x})] \tag{4}$$

where $f^{(i)}$ refers to the probability assigned by the model to the ith class and C is the total number of classes. We report both the variance of the probabilities (labeled 'prob' in Table 3) and the variance of the logits (before the softmax, labeled 'logits' in Table 3).

Divergence measures. One can also use divergence metrics to quantify ensemble diversity (Fort et al., 2019). We simple use the KL-divergence between the prediction probability distributions of a pair of models, and take the average over all possible pairs in the ensemble.

D.2. Visualization of diversity across selected classes

Figure 6 shows the accuracy per class for 10 randomly selected classes in ImageNet. The figure compares the performance of models trained with opposing equivariances (upper plot) and those with different random initializations (lower plot) and shows larger variances induced from opposing equivariant hypotheses. Further analysis of their diversity is presented in Section 5.1. The above results motivate the use of leveraging opposing equivariances as a method to induce diversity especially for large datasets like ImageNet.

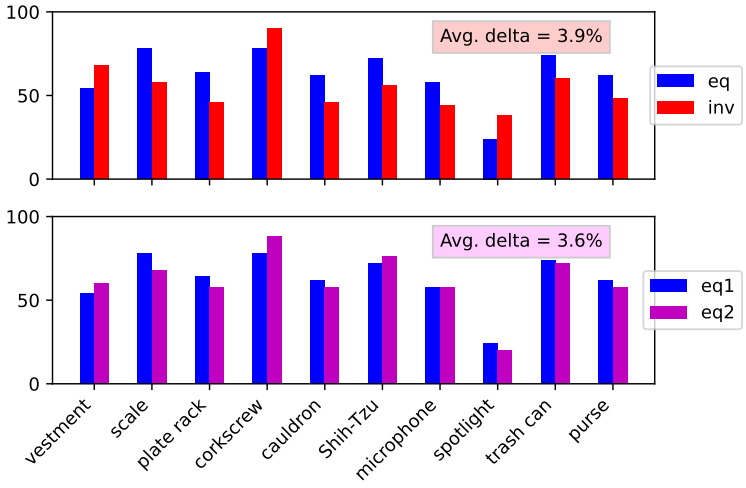

Figure 6. **Accuracy per class for 10 randomly selected classes in ImageNet.** Top panel compares the per class accuracy for a rotation equivariant model versus an invariant model and bottom panel compares the per class accuracy for two rotation equivariant models.

E. Uncertainty quantification results using random ensembles

Appendix G supplements the results in Table 4 in the main text. While Table 4 shows the results for the greedy ensembling approach, this table shows the results for the random ensembling approach. In both of the cases, we see general improvements in the uncertainty quantification metrics with additional models.

Table 7. **Uncertainty Quantification.** We evaluate the uncertainty quantification of the ensembles using the negative log likelihood loss (NLL) and the 'area under the uncertainty quantification curve' (AUUQC) which is obtained by sequentially removing the most uncertain samples and computing the area under the plot of ensemble accuracy versus fraction of samples removed.

	$M=2$	$M=3$	$M=4$	$M=5$
RANDOM ENSEMBLE				
NLL $\downarrow$ (10^{-1})				
EQ	8.43±0.03	8.20±0.03	8.09±0.02	8.03±0.02
EQ + INV	8.46±0.01	8.17±0.01	8.08±0.02	7.98±0.02
AUUQC $\uparrow$				
EQ	0.919±0.001	0.924±0.001	0.926±0.000	0.927±0.000
EQ + INV	0.921±0.001	0.926±0.001	0.927±0.000	0.928±0.000
GREEDY ENSEMBLE				
NLL $\downarrow$ (10^{-1})				
EQ	8.40±0.00	8.18±0.00	8.06±0.00	7.99±0.00
EQ + INV	8.43±0.03	8.16±0.01	8.03±0.01	7.97±0.01
AUUQC $\uparrow$				
EQ	0.921±0.000	0.925±0.000	0.927±0.000	0.928±0.000
EQ + INV	0.921±0.000	0.926±0.000	0.928±0.000	0.929±0.000

F. Transfer results using random ensembles

Table 8 supplements the results in Table 5 in the main text. While Table 5 shows the results for the greedy ensembling approach, this table shows the results for the random ensembling approach.

Table 8. **Transfer tasks for Random ensembles.** Ensemble efficiency is defined as the relative improvement over the mean accuracy of all the models in the ensemble. All experiments are fine-tuned except for iNaturalist-1k which is linear-probed.

	iNaturalist-1K	Flowers-102	CIFAR-100	Food-101
Single Model Accuracy				
Eq	55.1 ±0.3	91.9 ±0.0	85.5 ±0.1	87.9 ±0.1
Inv	56.3 ±0.2	91.2 ±0.1	84.0 ±0.1	87.9 ±0.1
Ensemble Accuracy ($M = 2$) (Ensemble efficiency)				
Eq	58.3 ±0.1 (3.2)	92.3 ±0.4 (0.4)	86.6 ±0.2 (1.1)	89.2 ±0.1 (1.2)
Eq + Inv	60.0 ±0.0 (4.3)	92.8 ±0.1 (1.3)	86.5 ±0.1 (1.8)	89.5 ±0.1 (1.6)
Ensemble Accuracy ($M = 3$) (Ensemble Efficiency)				
Eq	59.8 ±0.0 (4.7)	92.4 ±0.2 (0.5)	87.1 ±0.1 (1.6)	89.9 ±0.0 (1.3)
Eq + Inv	61.2 ±0.1 (5.5)	93.0 ±0.3 (1.3)	87.0 ±0.1 (2.0)	90.0 ±0.0 (2.1)

G. Uncertainty Quantification

G.1. Definition of BALD

In Section 5.2, we use the commonly used uncertainty measure BALD (Gal et al., 2017; Houlsby et al., 2011) to measure model uncertainty. It is defined as below

$$\mathbb{I}[y, \mathbf{w}|\mathbf{x}, \mathcal{D}] = \mathbb{H}[y|\mathbf{x}, \mathcal{D}] - \mathbb{E}_{p(\mathbf{w}|\mathcal{D})}\left[\mathbb{H}[y|\mathbf{x}, \mathbf{w}]\right]$$

where $\mathcal{D}$ refers to the training set, $p(\mathbf{w}|\mathcal{D})$ is the posterior our ensemble approximates, $\mathbf{w}$ are the model parameters, i.e. a member sampled from $p(\mathbf{w}|\mathcal{D})$, $\mathbb{H}[y|\mathbf{x}, \mathbf{w}]$ is the predictive entropy given model weights $\mathbf{w}$ and $\mathbb{H}[y|\mathbf{x}, \mathcal{D}] = -\sum_c p(y = c|\mathbf{x}, \mathcal{D}) \log p(y = c|\mathbf{x}, \mathcal{D})$ is the entropy of the ensemble's prediction.

G.2. Area under uncertainty quantification curve (AUUQC)

Figure 7 provides an illustration of the 'uncertainty quantification curve' described in Section 5.2, for ensembles of the leading hypothesis (rotation equivariant) with different ensemble sizes. As the ensemble size grows, the AUUQC increases as expected since a larger ensemble should be able to quantify uncertainty better.

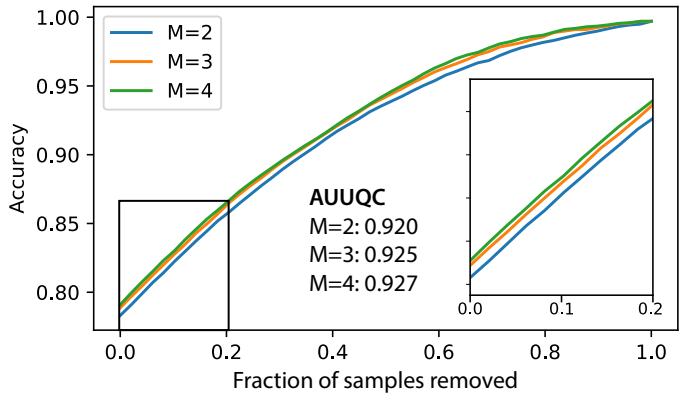

Figure 7. **Example of plot of the 'uncertainty quantification curve' used to generate AUUQC**.

H. Proportion of classes for hypothesis dominance

Table 9. **Proportion of classes and performance gains in Transfer datasets.** The top half of the table detail the proportion of classes captured by dominating hypothesis for each transfer dataset. The bottom half describes the accuracy and ensemble efficiency gained by capturing opposing hypothesis over a single hypothesis. This table is used to generate the bar plot (Figure 4) in the main text

	IN	iNat	Food	C100	IN-V2	IN-R
Eq > Inv	47.7	38.9	43.6	65.0	36.1	76.5
Eq < Inv	36.3	35.7	40.6	20.0	23.7	18.0
Δ_{acc} (EI - EE)	+0.1	+1.6	+0.2	+0.1	+0.1	0.0
Δ_{eff} (EI - EE)	+0.4	+0.8	+0.2	+0.2	+0.5	0.0

Table 10. **Proportion of classes and performance gains in ImageNet-100.** The top half of the table detail the proportion of classes captured by dominating hypothesis over different axes of transformations. The bottom half describes the accuracy and ensemble efficiency gained by capturing opposing hypothesis over a single hypothesis. This table supplements the results from Table 6 in the main text

	Rotate	HalfSwap	ColorInvert
Eq > Inv	65.0	28.0	48.0
Eq < Inv	15.0	44.0	26.0
Eq == Inv	20.0	28.07	26.0
Δ_{acc} (EI - EE)	0.0	0.04	0.14
Δ_{eff} (EI - EE)	0.0	0.27	0.24