1. Updated figure and table with more baselines

We have now added 3 more baselines to our study, 1) ensembles of Invariant-only models (Inv) , 2) supervised deep ensembles (DE supervised) and 3) deep ensembles of SimCLR models. The latter two are only half completed and will be updated asap. In particular, (1) provides strong evidence towards the effectiveness of our method by showing that including poorer invariant models in the ensemble gave better performance than ensembles of strong performing equivariant models.

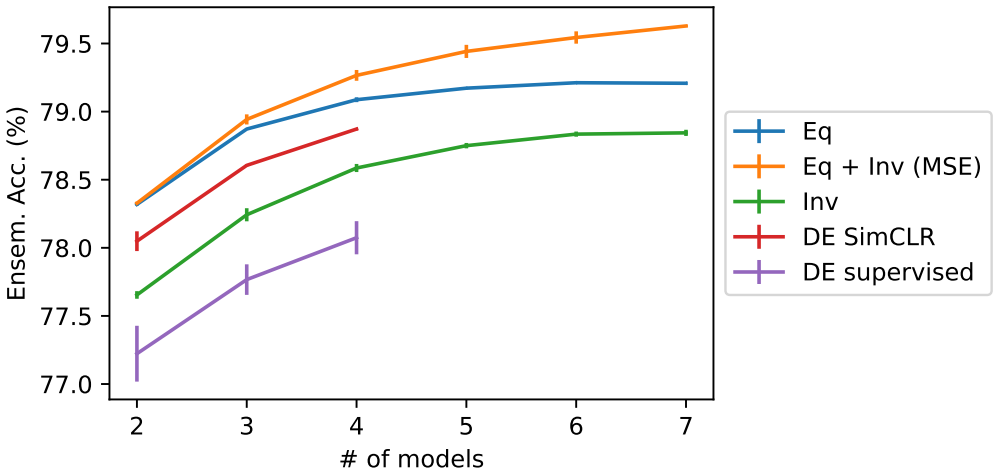

Figure 1. **Ensembles with opposing hypotheses have significantly larger potential.** Ensembles constructed only from a single hypothesis very quickly give marginal ensembling gains from adding more members.

We have also included new numbers for the new Inv baselines to our tables;

Table 1. **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

Table 2. **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

2. 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 (2) 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 2. 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.

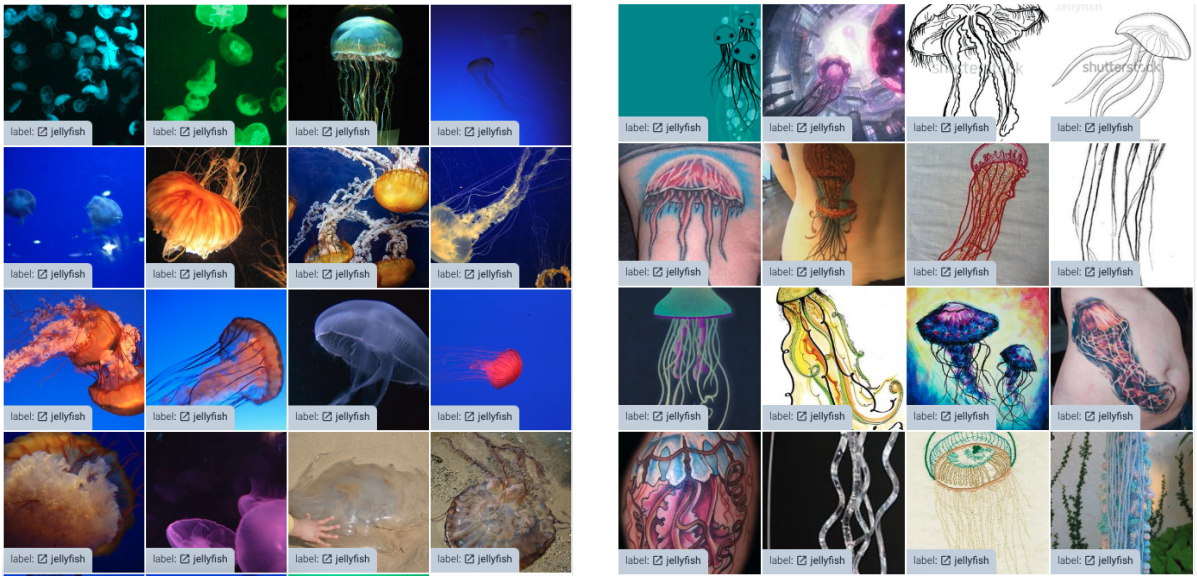

Figure 2. **Examples of images from the "jellyfish" class in ImageNet (left) and ImageNet-R (right).** Samples visualized using https://knowyourdata-tfds.withgoogle.com/

3. Including Additional Training Details in Appendix

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.

4. Improved organization of methods and results

We significantly improved our organization of sections 3 and 5.1 to improve the clarity and presentation. Changes are indicated in blue.

3. Multi-Symmetry Ensembles

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

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

where $\hat{\mathbf{z}^m_i}$ and $\hat{\mathbf{z}^m_j}$ are the ℓ_2-normalized representations of two views of an input $\mathbf{x}_i$ and $\hat{\mathbf{z}^m_i} = p_m \circ f_m(\mathbf{x}_i)/||p_m \circ f_m(\mathbf{x}_i)||$, 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,

$$\mathcal{L}^m_{eq} = \sum_{i=1}^{B} \sum_{g}^{|G|} H(h_m \circ f_m(T_g(\mathbf{x}_i)), g) \tag{2}$$

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* $\bar{f}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} f_i(\mathbf{x})$.

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.

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 1 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) consistently across all ensemble sizes. Figure 1 further highlights the gap between the ensemble accuracy of Eq + 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 (Eq + 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 the Appendix. Ensemble diversity is an important criterion since higher ensembling performance is derived when individual models make mistakes on different samples. Table 2 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.

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.

References

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.

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

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

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.

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

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.

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.

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