Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

Robustifying Routers Against Input Perturbations for Sparse Mixture-of-Experts Vision Transformers

Masahiro Kada[1], Ryota Yoshihashi[1], Satoshi Ikehata[1, 2] (Member, IEEE), Rei Kawakami[1] (Member, IEEE), and Ikuro Sato[1, 3] (Member, IEEE)

[1]Tokyo Institute of Technology, Tokyo, 152-8550 Japan
[2]National Institute of Informatics, Tokyo, 101-0003 Japan
[3]Denso IT Laboratory, Tokyo, 152-8550 Japan

Corresponding author: Masahiro Kada (email: mkada@d-itlab.c.titech.ac.jp).

This work was an outcome of a research project, Development of Quality Foundation for Machine-Learning Applications, supported by DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Tokyo Tech.). This work was also supported by JSPS KAKENHI Grant Number JP22H03642.

ABSTRACT Mixture of experts with a sparse expert selection rule has been gaining much attention recently because of its scalability without compromising inference time. However, unlike standard neural networks, sparse mixture-of-experts models inherently exhibit discontinuities in the output space, which may impede the acquisition of appropriate invariance to the input perturbations, leading to a deterioration of model performance for tasks such as classification. To address this issue, we propose Pairwise Router Consistency (PRC) that effectively penalizes the discontinuities occurring under natural deformations of input images. With the supervised loss, the use of PRC loss empirically improves classification accuracy on ImageNet-1K, CIFAR-10, and CIFAR-100 datasets, compared to a baseline method. Notably, our method with 1-expert selection slightly outperforms the baseline method using 2-expert selection. We also confirmed that models trained with our method experience discontinuous changes less frequently under input perturbations. The code will be released upon acceptance.

INDEX TERMS Mixture of Experts, Dynamic Neural Network, Image Classification, Vision Transformer

I. Introduction

MIXTURE of experts (MoE) [1] has been introduced to expand the expressivity of neural networks with multiple expert modules, each of which typically comprises a few layers adapted to a specific type of data. MoE networks generally have routers and experts, where routers typically selects one or a few experts based on the input features, and only the selected experts process these features. In recent years, deep neural network models that integrate MoE modules inside have been given much attention across various research fields. While initially popularized by large language models in natural language processing [2]–[8], this trend has extended to diverse visual tasks, such as image recognition [9]–[12], novel view synthesis [13], [14], image generation from text [10], [15], and motion prediction [16].

A key factor contributing to the growing interest in MoE is its sparse connectivity. It is now widely recognized that larger models trained on a broader range of data tend to generalize better [17], but they require greater computational costs in return. Sparse MoEs can address this issue in two ways. First, MoE allows the model size to increase by adding experts with negligible increase in computational cost by selecting a constant number of experts[1]. Second, given empirical observations that a learned expert captures specific semantics shared among a subset of training data [9], an MoE model with more experts can potentially accommodate more semantic variations for improved expressivity.

While sparse MoEs offer the advantages, the mechanism for selecting a limited number of experts can sometimes

[1]Typically, one or a few experts are selected.

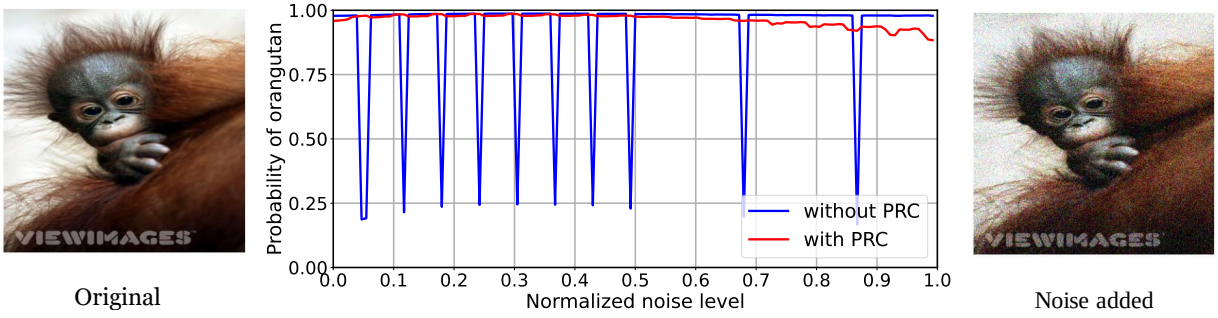

FIGURE 1: Illustration of the potential issue in sparse MoE models and how the proposed PRC method alleviates the issue, by gradually adding noises to an image of orangutan (Image taken from the ImageNet dataset [18]). Outputs of a sparse MoE model's can undergo abrupt changes under noise imposition (blue line). This is caused by the expert selection mechanism, which allows discontinuities in the output space. The proposed PRC method regularizes the routing function to better suppress the abrupt changes (red line).

compromise model performance. Unlike standard neural networks, sparse MoE models inherently exhibit discontinuities in their output space. This means that even small changes in input can result in different sets of experts being selected by the router, leading to abrupt changes in the model's output. Figure 1 illustrates how the prediction probability of a vision transformer (ViT) [19] with expert layer blocks [9] changes as Gaussian noises are gradually added to an input image. Once the noise level reaches a certain point, a different expert is selected, causing the model's output to change abruptly. At this point, the model becomes overly sensitive to the routing function. It is worth noting that discontinuities themselves could indeed bring benefits to certain tasks such as novel view synthesis for a single large-scale scene [13], where boundaries of different objects may be well represented by discontinuous functions. However, tasks that require robustness against input perturbations, such as classification, can potentially suffer from the discontinuous nature.

To address this issue, we propose Pairwise Router Consistency (PRC) as a regularization method for the sparse MoE variants of ViT. The purpose of PRC is to alleviate the problem that output of a sparse MoE can undergo discrete changes due to changes in the expert selection even for small perturbations in the input. This regularization method, which is inspired by the idea of pairwise consistency in recent contrastive learning [20], brings the router's outputs from differently augmented versions of a training image closer. Despite its simple formulation, the PRC loss effectively penalizes discontinuities caused by input perturbations, while also preventing router collapse—ensuring the router returns varied outputs based on its input. This paper presents empirical evidence supporting the efficacy of PRC in significantly reducing the occurrence of discontinuities under input perturbations, along with performance improvements in image classification tasks across multiple datasets. Our main contributions can be summarized as follows.

- In a sparse MoE, the model's output can potentially change discretely under input perturbations, likely hindering generalization for tasks such as image classification. To alleviate this issue, we propose Pairwise Router Consistency (PRC) to regularize sparse MoE models. PRC effectively penalizes the discontinuities stemming from the router against input perturbations, while preventing the router collapse.
- We demonstrate that PRC improves image classification accuracy on ImageNet-1K [18], CIFAR-10, CIFAR-100 [21], and Oxford Flowers-102 [22] by 0.43%, 0.16%, 0.89% and 0.88%, respectively, compared to the baseline V-MoE [9] S model. Notably, our method with 1-expert selection slightly outperforms the baseline method using 2-expert selection; *i.e.*, PRC slightly improves the test accuracy while halving the computational cost at the expert layer.
- We empirically confirm that a model trained with the PRC loss experience discontinuous changes less frequently under input perturbations.

II. Related Work

A. Mixture of Experts (MoE)

The mixture of experts (MoE) is a type of dynamic neural network [23] comprising routers and experts. The routers work as gates to allocate input data / tokens to suitable experts, and each expert specializes in processing the allocated data / tokens after training. Although the concept of MoE dates back to 1990s [24], it has been in the second spotlight in the era of large models due to its ability to enhance network speed through sparsification [2]. MoEs have been extended and applied in various domains such as natural language processing [3]–[8], [25], image processing [9], [10], [13], [15], [26]–[30], multimodal learning [31], multitask learning [32], [33], knowledge transfer [34], speech recognition [35], and graph neural networks [36], among others.

There are several variations in the routing mechanism, which serves a key element in MoE models. The token choice router is a common approach, where the routers use softmax function to allow each token to select an expert [2], [9]. However, this can lead to inefficient utilization of computational resources when the routing is unbalanced. To mitigate this, router variance loss was introduced to balance the routers' output. An alternative approach is the expert choice router [37], where experts select tokens conversely to mitigate computational inefficiency caused by expert underutilization. There have been more attempts to refine routing formulation as follows. Lewis *et al.* [6] treated the token-expert matching as a linear assignment problem. Clark *et al.* [38] and Liu *et al.* [39] formulated routing algorithm as a transportation optimization problem. Roller *et al.* [40] proposed static routing using a hash function. Sander *et al.* [41] introduced a differentiable top-k operator by viewing it as a linear program over the convex hull of permutations.

Not all MoEs are necessarily sparse; some recent methods use dense MoE without functions such as top-k for discretization [24], [42]. Puigcerver *et al.* [43] proposed soft MoE, which compresses ViT tokens into a single token through a weighted average, processes it through MoE, and redistributes the output to the original tokens, bypassing discrete top-k routing. While they inherently avoid the problem caused by discretization, their dense utilization of experts somehow weaken the advantages of sparse MoEs; *i.e.*, sparse connectivity between routers and experts.

In our research, we adopt the token choice MoE as our basis of implementation. We apply the PRC loss in its routers, and this helps alleviate the problems associated with sparse MoE discretization without increasing computational load during inference.

B. Consistency-Based Learning

In recent years, the consistency-based learning, a type of unsupervised learning, has developed significantly. Methods of consistency-based learning encourages models to produce consistent predictions / representations among different transformations of a given input data, regardless of the existence of the label.

1) Semi-Supervised Learning

Semi-supervised learning utilizes both labeled and unlabeled data in the training. Typically, consistency regularization is applied for unlabeled data in such a way that the model returns consistent (similar) predictions / representations for two randomly augmented data from a common input data. Sajjadi *et al.* [44] first applied the consistency regularization for image classification task. Laine and Aila [45] proposed a method that keeps updating exponential moving average of stale models beside the main target model and calculates the consistency between the two models. Miyato *et al.* [46] proposed a method to calculate consistency against adversarial attacks.

2) Self-Supervised Learning with Contrastive Learning

Contrastive learning, a type of self-supervised representation learning, enhances pairwise consistency typically by bringing representations of positive input pairs closer together and pushing those of negative input pairs farther apart.

Hadsell *et al.* [47] first proposed contrastive learning for dimensionality reduction. More recently, contrastive learning attracted particular attention for self-supervised learning of neural representations. SimCLR [20] and MoCo [48] are ones of representative methods of self-supervised learning for images, exploiting two data-augmented versions of a single image as positives, and those of a different image pair as negatives. BYOL [49] performs self-supervised learning using only positive pairs without explicitly using negative pairs, since truly negative pairs are hard to be identified. Zbontar *et al.* [50] proposed a method that prevents the representation from collapsing to trivial solutions by bringing the cross-correlation matrix between network outputs closer to the identity matrix.

While our study is inspired by these contrastive representation learning approaches, it differs from them in two aspects. First, we measure the degrees of consistency in the output space of the router of the sparse MoE models, rather than general neural representations of the models. Second, not only bringing the router outputs close together for a positive input pair, our method encourages the pairwise consistency after the top-k discretization process along with other regularizations enhancing the balanced usage of different experts.

III. Method

The proposed PRC imposes consistency regularization on the output of routing function to address the discontinuity issue in sparse MoE models discussed earlier. Taking inspiration from self-supervised contrastive learning, PRC generates a pair of randomly augmented images from a training image and update the model parameters to reduce the dissimilarity in the output space of the router, in addition to other loss terms. We implemented this unsupervised regularization loss on the existing Vision Mixture of Experts (V-MoE) framework [9].

A. Overview: Vision Mixture of Experts (V-MoE)

The Vision Mixture of Experts (V-MoE) model, proposed by Riquelme *et al.* [9], integrates MoE modules into Vision Transformer (ViT) [19]. Specifically, the Multilayer Perceptron (MLP) in the transformer block is replaced with the MoE block, as illustrated in the upper part of Fig. 2. Since ViT decomposes an image into patches, the MoE module is applied to the features of each patch independently. Denoting patch index as p and input image as x, the router $r_\theta(x, p)$

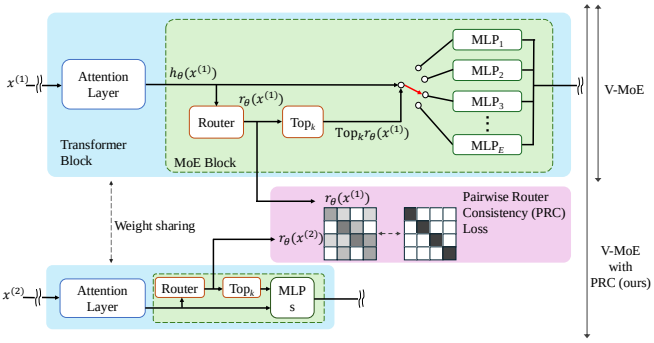

FIGURE 2: Visualization of the primary part of the V-MoE model architecture [9] (upper) and the extended part by our PRC method (upper and lower). PRC brings the router outputs of corresponding patches in two data-augmented images closer together. The patch index p is dropped for simplicity.

takes the feature vector $h_\theta(x, p)$ from the attention layer and applies a linear projection with learnable matrix θ_{r} followed by a softmax nonlinearlity as follows:

$$r_\theta(x, p) = \mathrm{Softmax}(\theta_{\mathrm{r}}\, h_\theta(x, p) + \epsilon) \in \mathbb{R}_+^E, \qquad (1)$$

where θ represents the set of model parameters, and E denotes the total number of experts for a given layer consisting of experts. The random noise vector ϵ is added to the softmax argument to give every expert a chance to be selected in the subsequent top-k operation. The output of the MoE module is given by

$$\mathrm{MoE}_\theta(x, p) = \sum_{i=1}^{E} \mathrm{Top}_k(r_\theta(x, p))_i \mathrm{MLP}_i(h_\theta(x, p)), \qquad (2)$$

The $\mathrm{Top}_k(v)$ returns a vector such that $\mathrm{Top}_k(v)_i = v_i$ if i-th component of v is within top-k and $\mathrm{Top}_k(v)_i = 0$ otherwise. With this formulation, only k experts are executed, depending on the input x.

V-MoE employs a specific type of regularization on the routing function aimed at ensuring balanced utilization of every expert within the MoE framework. The regularization seeks to enhance the expressive capabilities of the MoE block by encouraging even usage of experts. In practice, V-MoE computes the average of router outputs across different patches and samples, then calculates the variance across expert indices. This variance serves as a loss term to be minimized alongside other loss components. While achieving perfectly uniform expert usage is not guaranteed, minimizing this loss fosters a more balanced utilization of experts, thereby enhancing the overall performance and effectiveness of the MoE model.

Intriguingly, through the training method outlined above, Riquelme *et al.* empirically demonstrated that the trained router displays a correlation between experts and semantics [9]. This correlation implies that certain experts are selected with high probability for inputs associated with specific class labels. Such correlation likely contribute to the robustness of the classification capabilities exhibited by the model.

It has been empirically known that training MoE networks without regularizing routers often results in the situation where only one or a few specific experts are selected for all or majority of the input data [2]. This situation is obviously unwanted, because it fails to fully bring out the representation capabilities of the expert blocks. To address this issue, Riquelme *et al.* introduced what-they-call importance loss L_θ^{imp} and load loss L_θ^{load}. Roughly speaking, these loss terms are designed to enhance the balanced usage of experts by computing the variance of the sample-mean of the softmax outputs from the router.

B. Proposed Method: Pairwise Router Consistency (PRC)

To cope with the potential issues stemming from abrupt changes in the output of an MoE router, we propose Pairwise Router Consistency (PRC) to regularize the routing function by use of the pairwise loss term. The pairwise loss measures the discrepancy in the output space of the router function with respect to a pair of input images generated by random data augmentation from a common training sample. Combined with other regularization including collapse suppression, we formulate the PRC loss with a simple formula.

We outline the key requirements for effectively regularizing the routing function. First, the routing function should return nearly invariant vectors with respect to natural deformations of a common input. Second, the routing function should promote to the generation of 'k-hot' vectors, meaning that only k elements possess significantly larger values than the rest. Third, it should ensure diverse output of the router among data, ideally with a similar probability for each expert to be selected. The first requirement, with the aid of the second requirement, addresses the potential issue of sparse MoE routers undergoing abrupt changes under natural deformations of input, thereby safeguarding the model's output robustness. However, a collapsed router, which returns a constant output regardless of input, can trivially avoid the discontinuity issue, but will result in poor model performance, because the same top-k experts are always chosen and the expressivity of the expert module is underutilized. This triviality can be mitigated by fulfilling the third requirement.

To fullfill all three requirements for regularizing the routing function, we introduce the Pairwise Router Consistency (PRC) method. The PRC loss is constructed from asymmetric correlation matrices comprising pairs of output vectors of the routing function, $r_\theta(x^{(1)}, p_1)$ and $r_\theta(x^{(2)}, p_2)$. Here, $x^{(1)}$ and $x^{(2)}$ denote two images produced by random data-augmentation techniques from a common image x, while p_1 and p_2 denote patch indices of $x^{(1)}$ and $x^{(2)}$, respectively. In the following discussion, we assume p_1-th patch of $x^{(1)}$ and p_2-th patch of $x^{(2)}$ correspond each other. When geometrical transformations are adopted as data augmentation, one needs

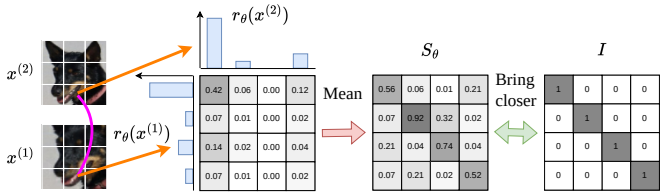

FIGURE 3: Overview of the computational procedure of Pairwise Router Consistency (PRC). PRC involves three steps: 1. Computation of the router outputs $r_\theta(x^{(1)})$ and $r_\theta(x^{(2)})$ between corresponding patches. 2. Computation of the asymmetric correlation matrix $r_\theta(x^{(1)})r_\theta(x^{(2)})^\top$. 3. Averaging over patches and samples followed by constant multiplication to compute S_θ. PRC encourages $S_\theta \simeq I$. The patch index p is dropped for simplicity. Images are taken from the ImageNet dataset [18].

to compute patch correspondence to identify appropriate pair of p_1 and p_2, as we describe in the following subsection.

We provide an overview of the entire process of PRC in Fig. 2 and an illustration of detailed procedure for computing the PRC loss in Fig. 3. Given the router outputs $r_\theta(x^{(1)}, p_1)$ and $r_\theta(x^{(2)}, p_2)$, the asymmetric correlation matrix S_θ is computed for the input dataset $\mathcal{X}$ as follows:

$$S_\theta = \frac{E}{\sum_{x\in\mathcal{X}} |\mathcal{U}_x|} \sum_{x\in\mathcal{X}} \sum_{(p_1,p_2)\in\mathcal{U}_x} r_\theta(x^{(1)}, p_1) r_\theta(x^{(2)}, p_2)^\top, \tag{3}$$

where $\mathcal{U}_x$ represents the set of corresponding patch pairs, with the p_1-th patch of $x^{(1)}$ corresponding to the p_2-th patch of $x^{(2)}$. In Eq. (3), $r_\theta(x^{(1)}, p_1)$ represents E-dimensional softmax output for the p_1-th patch. Again, E represents the total number of experts for a given layer consisting of experts. For simplicity, let us assume $k = 1$ for now. The second requirement states that the router should return nearly one-hot vectors, hence the product $r_\theta(x^{(1)}, p_1)r_\theta(x^{(2)}, p_2)^\top \in \mathbb{R}_+^{E\times E}$ will ideally have all zeros except for one element, which holds the value of one. In addition, the third requirement, ensuring equal selection probability for each expert, requires that every row has the same chance of having the non-zero elements. Meanwhile, the first requirement, encouraging router invariance under deformations, mandates that positive values should appear in the diagonal elements of the matrix. Combining these requirements, we define the PRC loss as follows:

$$L_\theta^{\mathrm{PRC}} = \frac{\lambda_{\mathrm{diag}}}{E} \sum_{i=1}^{E} (1 - S_{\theta\, ii})^2 + \frac{\lambda_{\mathrm{offdiag}}}{E(E-1)} \sum_{i=1}^{E} \sum_{j\neq i} S_{\theta\, ij}^2, \tag{4}$$

where hyperparameters $\lambda_{\mathrm{diag}}, \lambda_{\mathrm{offdiag}} > 0$ control the relative strengths for the diagonal and off-diagonal terms. We can further simplify L_θ^{PRC} by introducing following $E \times E$ coefficient matrix:

$$\Lambda_{ij} = \begin{cases} \sqrt{\frac{\lambda_{\mathrm{diag}}}{E}} & \text{for } j = i, \\ \sqrt{\frac{\lambda_{\mathrm{offdiag}}}{E(E-1)}} & \text{for } j \neq i. \end{cases} \tag{5}$$

Then, the PRC loss can be concisely rewritten with the Hadamard product $\odot$ as

$$L_\theta^{\mathrm{PRC}} = \|\Lambda \odot (I - S_\theta)\|_2^2, \tag{6}$$

where I represents $E \times E$ identity matrix.

So far, we restricted our discussion for $k = 1$ case; however, we confirmed in experiments that Eq. (4) works very well not only for $k = 1$ but for $k = 2$ settings[2]. With $k = 2$, PRC empirically shows clearer suppression of the sum of the bottom-$(E - 2)$ values of r_θ, meaning that r_θ is dominated by the largest and the second largest values, resulting in the superior performance of our regularization method.

By combining the proposed unsupervised pairwise loss L_θ^{PRC} with the standard cross-entropy supervised loss L_θ^{S}, the optimal model parameters $\theta^\star$ can be obtained as

$$\theta^\star = \arg\min_\theta \; L_\theta^{\mathrm{S}} + L_\theta^{\mathrm{PRC}}. \tag{7}$$

Beside regularization of the routing function r_θ, we employ the same computational procedure as in V-MoE, including the weighted sum of expert outputs as described in Eq. (2).

In practice, we adopt the mini-batch sampling scheme for the approximate minimization of $L_\theta^{\mathrm{S}} + L_\theta^{\mathrm{PRC}}$. Specifically, L_θ^{PRC} and L_θ^{S} are computed approximately from samples in a mini-batch that is randomly sampled at each iteration. Note that the proposed PRC method does not use either importance loss or load loss.

We also investigated a parallel approach that uses distribution distance instead of the quadratic distance as in Eq. (6). As the router usually accompanies softmax activation function, one may regards the output of the router as the expert-selection probability; therefore, use of distribution distance function such as Jensen-Shannon Divergence might sound a natural choice. Here, we briefly report the empirical results for readers' reference. We implemented a Jensen-Shannon Divergence loss as the consistency regularization along with the load loss and the importance loss. The model performance was a little bit better than V-MoE model but worse than the proposed PRC loss defined by Eq. (6). We empirically found that the distribution distance version has a tendency where the router output shows slightly higher entropy; therefore, the consistency of top-k selection is slightly more vulnerable than the proposed PRC approach.

C. Computation of the patch correspondences

When applied to Sparse MoE vision transformers, one needs to identify patch correspondences to compute PRC loss, since the routing function is applied to features of patches independently. In the following, we describe how to identify patch correspondences between two augmented data, especially when geometrical transformations are adopted as data-augmentation techniques. In this study, for a given patch p_1 of a data-augmented image $x^{(1)}$, we simply identify the nearest patch center from the other data-augmented

[2] $k = 2$ was adopted in [9].

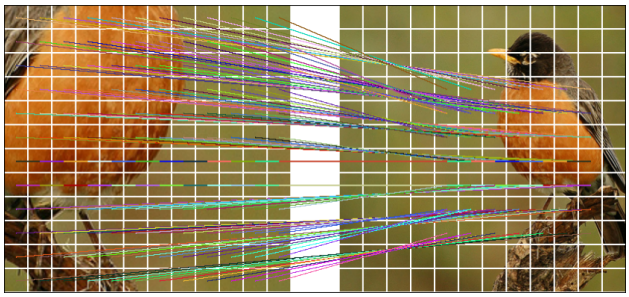

FIGURE 4: Visualization of patch correspondences between image 1 (left) and image 2 (right). Straight lines show corresponding patches. Note that we assign higher-resolution image as image 1 between a randomly augmented image pair. The geometrical transformation matrix that takes image 1 coordinates to image 2 coordinates is used to find corresponding patches. The image are taken from the ImageNet dataset [18].

image $x^{(2)}$ from the patch center of p_1. Specifically, when transformation matrices associated with image cropping and affine transformations are represented by R_1 and R_2 for image 1 and 2, respectively, the geometrical transformation from image 1 to image 2 is given by $R_1^{-1}R_2$. Using this transformation, we transform the center position of each patch of image 1 into the image 2 coordinate system to find the nearest patch center of image 2.

The above strategy is simple, but asymmetric under the exchange of image 1 and 2. To keep as many corresponding patches as possible, we use the following trick. Between two randomly augmented images, we assign the one with higher resolution to image 1. We illustrate an example of patch correspondences between the two augmented images in Fig. 4.

IV. Experiments

We conducted comparative evaluations between the baseline method (V-MoE) and the proposed method (V-MoE with PRC) using image classification datasets: ImageNet-1K [18], CIFAR-10, CIFAR-100 [21], and Oxford Flowers-102 [22]. We also evaluate the robustness of the models under $k = 1$ and $k = 2$ settings, where k is the number of experts used at the test time. We then analyze the statistics of the router output for both methods, exploring the robustness of the router in response to data augmentation.

A. Settings

Table 1 summarizes the hyperparameter setting used in this work. These hyperparameters were taken from the official implementation by the V-MoE authors [9]. The same hyperparameters were used for both the baseline and proposed methods. In the pre-training, shrunk images (size: 224×224) are used to save computational time, as was adopted in the previous work. During the fine-tuning, the original-size images (size: 384×384) are used. Since the number of patches

TABLE 1: Model and training hyperparameters used in the experiments of the proposed and baseline methods.

Number of Transformer blocks	8
Layer number of MoE	6, 8
Hidden layer size	512
Patch image size	32×32
Input image size (pre-training)	224
Input image size (fine-tuning)	384
Optimizer (pre-training)	Adam
Optimizer (fine-tuning)	SGD
Data augmentation (pre-training)	RandAugment [51]
Data augmentation (fine-tuning)	Random Crop, Random Flip
Initial learning rate	5e-4
Learning rate schedule	Cosine decay with warmup
Number of experts E	8
λ_{load} (baseline method only)	5e-3
λ_{imp} (baseline method only)	5e-3
λ_{diag} (PRC only)	5e-3
λ_{offdiag} (PRC only)	5e-2

changes during pre-training and fine-tuning, applying a naive way of positional encoding may be harmful because of the difference of the positional ranges. To alleviate this, we adopt 2D interpolation on the pre-trained positional encoding to generate better aligned positional encoding at the original image size.

Both the baseline and proposed methods employ the ViT-S model as their base architecture. Pre-training of the model was carried out on the ImageNet-1K dataset. The detailed model hyperparameters are presented in the appendix. While the baseline V-MoE method incorporates mixup [52] for data augmentation, the proposed method does not utilize this.

B. Results

1) Fine-tuning results

Table 2 shows the accuracy on the test sets of ImageNet-1K, CIFAR-10, and CIFAR-100 after performing fine-tuning on the corresponding training sets. Following the conditions adopted in [9], we used $k = 2$, where k is used in the top-k operator in Eq. (2). For the mean accuracy, the proposed PRC method consistently outperformed the baseline method on all three datasets. Taking the error bars into account, the mean gaps are clearly greater than the error bars on ImageNet-1K, CIFAR-100 and Oxford Flowers-102.

2) Robustness about *k*

We compare model performance for $k = 1$ and $k = 2$ cases. To obtain $k = 1$ models, we took the pre-trained model on ImageNet-1K with $k = 2$ setting, and fine-tuned on ImageNet-1K, CIFAR-10, CIFAR-100, and Oxford Flowers-102 with $k = 1$ setting. Smaller k is generally preferred because of lower computational costs.

TABLE 2: Test accuracies on ImageNet-1K, CIFAR-10, CIFAR-100 and Flower datasets with $k = 2$ setting. The average and standard deviation of the test accuracies over three trials are shown. PRC clearly outperforms the baseline method on all datasets.

	ImageNet-1K	CIFAR-10	CIFAR-100	Flowers
V-MoE-S [9]	$75.84 \pm 0.04\%$	$95.20 \pm 0.07\%$	81.38 ± 0.03	$89.30 \pm 0.09\%$
PRC (ours)	$\mathbf{76.27 \pm 0.01\%}$	$\mathbf{95.36 \pm 0.02\%}$	$\mathbf{82.27 \pm 0.10\%}$	$\mathbf{90.18 \pm 0.01\%}$

TABLE 3: Test accuracies on ImageNet-1K, CIFAR-10, CIFAR-100 and Flower datasets with $k = 1$ setting. The average and standard deviation of the test accuracies over three trials are shown. PRC clearly outperforms the baseline method on all datasets. Comparing with Table 2, the mean accuracies of the proposed method with $k = 1$ are higher than those of the baseline method with $k = 2$ on ImageNet-1K and CIFAR100.

	ImageNet-1K	CIFAR-10	CIFAR-100	Flowers
V-MoE-S [9]	$75.23 \pm 0.01\%$	$94.81 \pm 0.06\%$	$81.18 \pm 0.07\%$	$90.21 \pm 0.09\%$
PRC (ours)	$\mathbf{75.92 \pm 0.02\%}$	$\mathbf{95.12 \pm 0.16\%}$	$\mathbf{82.12 \pm 0.18\%}$	$\mathbf{91.24 \pm 0.08\%}$

Table 3 shows the accuracy with $k = 1$ setting. Similar to $k = 2$ cases, PRC consistently outperforms the baseline on ImageNet-1K, CIFAR-10, CIFAR-100, and Oxford Flowers-102 with $k = 1$ setting. The performance gaps between PRC and the baseline for $k = 1$ cases are clearly larger than those for $k = 2$ cases. This implies that PRC performs well with $k = 1$ setting. More remarkably, ours with $k = 1$ even slightly outperforms the baseline with $k = 2$ on ImageNet-1K and CIFAR-100. This means that our PRC can safely reduce computational cost of the MoE blocks by half without compromising generalization abilities.

C. Analyses

1) Router robustness against data augmentation

We investigated the extent to which different sets of experts are selected by the routers under data augmentation procedures. In this experiment, we analyzed the ratios of matches in the expert selection using the sixth MoE layer of the fine-tuned V-MoE-S model with $k = 2$ setting. The ImageNet-1K validation dataset is used. To count the matched and mismatched pairs, we go through the procedure depicted in Fig. 3-1. Namely, we first compute $\mathrm{Top}_k(r_\theta(x^{(1)}, p_1))$ and $\mathrm{Top}_k(r_\theta(x^{(2)}, p_2))$, where $x^{(1)}, x^{(2)}$ represent randomly deformed images from validation sample x, and p_1, p_2 indicate corresponding patches. We then count the cases where top-k indices match.

Table 4 shows the results. The 'Top-1' column shows the average ratio at which the top-1 index matches, whereas the 'Top-2' column shows the average ratio at which both the highest matches and the second highest matches. The 'Top-2 (order-agnostic)' is the same as 'Top-2' except that orders of the top-1 and top-2 indices do not matter. In all three categories, our PRC clearly exhibits higher matching ratios for unseen data by large margins. This fact indicates that the PRC method reduces the frequency of discontinuous changes in the router's output for $k = 1$ and $k = 2$ settings. These observations likely support the superior performance of PRC.

TABLE 4: Ratios of matching in the expert selection under data augmentation. Higher matching ratios indicate that the expert selection is more robust against deformations of input images; therefore, the model undergoes less frequent discrete changes under such operations.

	Top-1	Top-2	Top-2 (order-agnostic)
V-MoE-S [9]	63.04%	34.77%	45.44%
PRC (ours)	**75.78%**	**48.54%**	**58.96%**

TABLE 5: Mean values of the highest, the second highest, and the sum of the rest in the output of the routing function r_θ. The results indicate that PRC encourages 1) top-1 components to have large values and 2) the sum of the 3rd and the rest components to be suppressed. These observations provide indirect evidence that the router function becomes more robust against small perturbations.

	Highest	2nd highest	Sum of the rest
V-MoE-S [9]	0.7586	0.1493	0.0921
PRC (ours)	0.8788	0.0915	0.0297

Figure 5 gives qualitative examples that show the robustness about the router output. When adding small Gaussian noise in input images, our PRC produces less changes in the top-1 expert selection, especially for the foreground objects, compared to the existing method [9].

2) Router confidence

We have already observed the remarkable experimental results: the proposed PRC with $k = 1$ setting (top-1 selection) ties with the baseline with $k = 2$ setting (top-2 selection) in accuracy. To further analyze the cause behind this observation, we investigated the characteristics of the routing function. It would be no doubt that a model performs worse if the router returns nearly uniform vectors, as this would

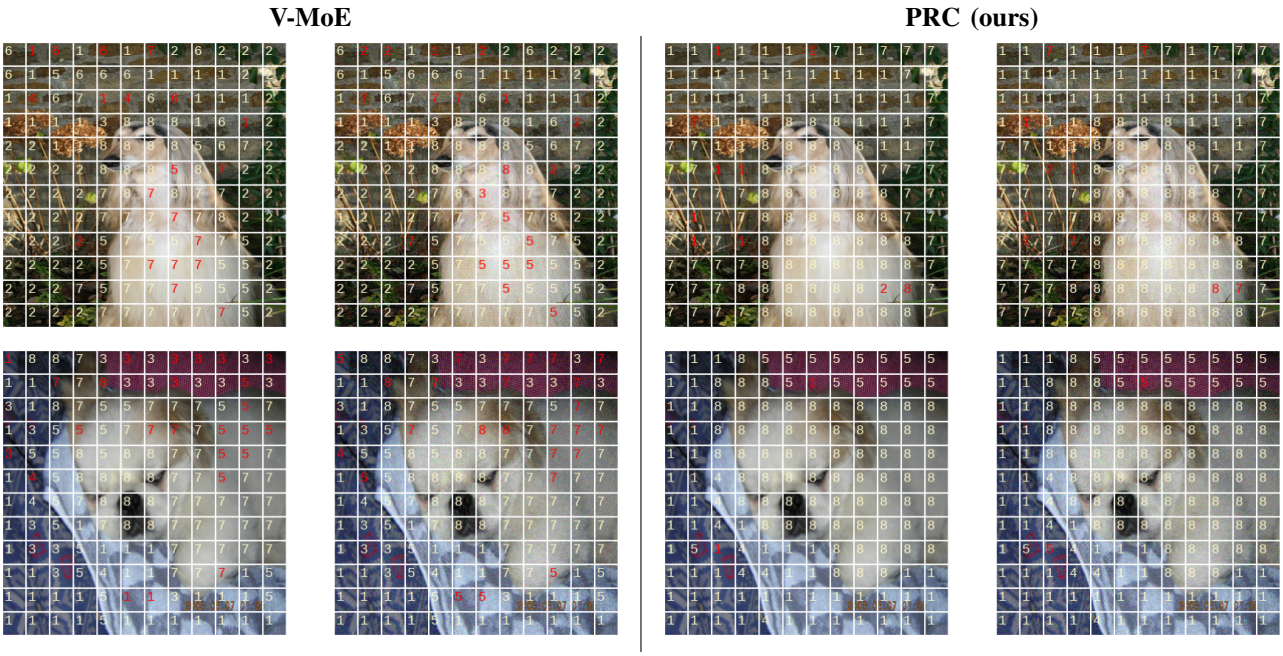

FIGURE 5: Demonstration of the robustness of the routers produced by V-MoE [9] (left) and our PRC (right). Images in the 1st and 3rd columns are the input (images taken from ImageNet [18]), and images in the 2nd and 4th columns are the same input images with small Gaussian noise. The number given in each patch indicates the selected expert ID. Red ones indicate the cases, where different experts are selected by adding noise. It is observed that ours makes the expert selection more robust against noise superposition, especially for the foreground objects.

lead to unstable expert selection even with very small input perturbations. Therefore, models with stronger peaks in the router's output are generally expected to perform better with $k = 1$ setting. Similarly, models exhibiting higher values in the sum of the top-2 scores would typically perform better with $k = 2$ setting.

We computed the highest and the second highest scores in the router's output for all samples in the ImageNet-1K validation set. Subsequently, we calculated the mean of the highest score, the mean of the second-highest score, and the mean of the sum of the rest of the scores. Table 5 presents these values. As expected, our PRC yields a higher score in the highest box, indicating that the router provides greater confidence in selecting an expert. This indirectly supports the superior performance of PRC with $k = 1$ setting. Additionally, when summing the highest and the second-highest scores in the table, PRC also exhibits a higher score compared to the baseline. Again, this indirectly suggests the superiority of the PRC performance with $k = 2$ setting.

V. Summary and Discussion

In this paper, we proposed Pairwise Router Consistency (PRC), an unsupervised regularization method designed to enhance the robustness of the routing function in sparse MoE models against input perturbations. Without the inclusion of the PRC loss term, sparse MoE models may exhibit discontinuous output patterns along axes of natural deformations for a given input image. Such discontinuities can impede the model's ability to acquire appropriate invariance necessary for effective classification. Our PRC method primarily calculates pairwise dissimilarities among the router's outputs for randomly deformed inputs. Through empirical evaluations, we confirmed that incorporating the PRC loss consistently improves the performance of a ViT-based sparse MoE across multiple image classification datasets. Notably, our method with 1-expert selection slightly outperforms the baseline method using 2-expert selection. Additionally, analyses conducted on various statistics related to the routing function further support the superior performance of the models trained with PRC.

Limitation. In this paper, we have demonstrated the effectiveness of the proposed method. However, it remains unclear how effective the proposed method is when applied to MoE models used in the field of natural language processing, where MoE exhibits great success. The impact on model performance in other visual tasks, such as novel view synthesis, image generation, and vision-and-language tasks, is yet to be investigated. These areas present promising directions for future research to explore the broader applicability and effectiveness of our proposed method beyond image classification tasks. In addition, the proposed method is fundamentally applicable to semi-supervised learning problems as the PRC loss is independent of labels, but this has not been evaluated in this paper.

REFERENCES

[1] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[2] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *ICLR*, 2017.

[3] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," in *ICLR*, 2020.

[4] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, and C. Cui, "GLaM: Efficient scaling of language models with mixture-of-experts," in *ICML*, 2022.

[5] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research (JMLR)*, vol. 23, no. 1, jan 2022.

[6] M. Lewis, S. Bhosale, T. Dettmers, N. Goyal, and L. Zettlemoyer, "Base layers: Simplifying training of large, sparse models," in *ICML*, 2021.

[7] D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. K. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, and W. Liang, "DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.

[8] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[9] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," in *NeurIPS*, 2021.

[10] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng, Y. Sun, L. Chen, H. Tian, H. Wu, and H. Wang, "Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts," in *CVPR*, June 2023, pp. 10 135–10 145.

[11] T. Chen, X. Chen, X. Du, A. Rashwan, F. Yang, H. Chen, Z. Wang, and Y. Li, "AdaMV-MoE: Adaptive multi-task vision mixture-of-experts," in *ICCV*, 2023.

[12] J. Zhu, X. Zhu, W. Wang, X. Wang, H. Li, X. Wang, and J. Dai, "Uni-perceiver-MoE: Learning sparse generalist models with conditional MoEs," in *NeurIPS*, 2022.

[13] Z. Mi and D. Xu, "Switch-NeRF: Learning scene decomposition with mixture of experts for large-scale neural radiance fields," in *ICLR*, 2023.

[14] Y. Sekikawa, C. Hsu, S. Ikehata, R. Kawakami, and I. Sato, "Gumbel-nerf: Representing unseen objects as part-compositional neural radiance fields," in *ICIP*, October 2024.

[15] W. Cong, H. Liang, P. Wang, Z. Fan, T. Chen, M. Varma, Y. Wang, and Z. Wang, "Enhancing nerf akin to enhancing llms: Generalizable nerf transformer with mixture-of-view-experts," in *ICCV*, 2023.

[16] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. J. Black, "Stochastic scene-aware motion prediction," in *ICCV*, October 2021, pp. 11 374–11 384.

[17] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[21] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[22] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[23] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7436–7456, 2022.

[24] S. Nowlan and G. E. Hinton, "Adaptive soft weight tying using gaussian mixtures," in *NeurIPS*, 1991.

[25] F. Xue, Z. Zheng, Y. Fu, J. Ni, Z. Zheng, W. Zhou, and Y. You, "Openmoe: an early effort on open mixture-of-experts language models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2025.

[26] J. Puigcerver, R. Jenatton, C. Riquelme, P. Awasthi, and S. Bhojanapalli, "On the adversarial robustness of mixture of experts," in *NeurIPS*, 2022.

[27] Y. Zhang, R. Cai, T. Chen, G. Zhang, H. Zhang, P.-Y. Chen, S. Chang, Z. Wang, and S. Liu, "Robust mixture-of-expert training for convolutional neural networks," in *ICCV*, 2023.

[28] Y. Jain, H. Behl, Z. Kira, and V. Vineet, "Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets," in *NeurIPS*, 2023.

[29] A. Royer, I. Karmanov, A. Skliar, B. E. Bejnordi, and T. Blankevoort, "Revisiting single-gated mixtures of experts," in *BMVC*, 2022.

[30] Y. Shi, S. N, B. Paige, and P. Torr, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[31] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," in *NeurIPS*, 2022.

[32] h. liang, Z. Fan, R. Sarkar, Z. Jiang, T. Chen, K. Zou, Y. Cheng, C. Hao, and Z. Wang, "M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design," in *NeurIPS*, 2022.

[33] S. Kudugunta, Y. Huang, A. Bapna, M. Krikun, D. Lepikhin, M.-T. Luong, and O. Firat, "Beyond distillation: Task-level mixture-of-experts for efficient inference," in *EMNLP*, 2021.

[34] X. Yang, D. Zhou, S. Liu, J. Ye, and X. Wang, "Deep model reassembly," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 25 739–25 753.

[35] Y. Kwon and S.-W. Chung, "Mole : Mixture of language experts for multi-lingual automatic speech recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[36] H. Wang, Z. Jiang, Y. You, Y. Han, G. Liu, J. Srinivasa, R. R. Kompella, and Z. Wang, "Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling," in *NeurIPS*, 2023.

[37] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, z. Chen, Q. V. Le, and J. Laudon, "Mixture-of-experts with expert choice routing," in *NeurIPS*, 2022.

[38] A. Clark, D. de Las Casas, A. Guy, A. Mensch, M. Paganini, J. Hoffmann, B. Damoc, B. A. Hechtman, T. Cai, S. Borgeaud, G. van den Driessche, E. Rutherford, T. Hennigan, M. J. Johnson, K. Millican, A. Cassirer, C. Jones, E. Buchatskaya, D. Budden, L. Sifre, S. Osindero, O. Vinyals, J. W. Rae, E. Elsen, K. Kavukcuoglu, and K. Simonyan, "Unified scaling laws for routed language models," in *ICML*. PMLR, 2022.

[39] T. Liu, J. Puigcerver, and M. Blondel, "Sparsity-constrained optimal transport," in *ICLR*, 2023.

[40] S. Roller, S. Sukhbaatar, J. Weston *et al.*, "Hash layers for large sparse models," in *NeurIPS*, 2021.

[41] M. E. Sander, J. Puigcerver, J. Djolonga, G. Peyré, and M. Blondel, "Fast, differentiable and sparse top-k: a convex analysis perspective," in *ICML*, 2023.

[42] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," in *NeurIPS*, 2019.

[43] J. Puigcerver, C. R. Ruiz, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *ICLR*, 2024.

[44] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.
[45] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning,," in *ICLR*, 2017.
[46] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2019.
[47] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, vol. 2, 2006, pp. 1735–1742.
[48] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.
[49] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *NeurIPS*, 2020.
[50] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *ICML*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 12 310–12 320.
[51] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *NeurIPS*, 2020.
[52] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.