Pruning-based OOD Detection via Activation Consistency Under Extraneous Shifts

Tim Krieg
TU Darmstadt
tim-michael.krieg@stud.tu-darmstadt.de

Paul Rabich
TU Darmstadt
paul.rabich@stud.tu-darmstadt.de

Marc Saghir
TU Darmstadt
marc.saghir@stud.tu-darmstadt.de

Abstract

Out-of-distribution (OOD) detection is a crucial test setting for ensuring the reliability of deep learning models deployed in real-world applications. A variety of different approaches to this problem have come up in recent years, employing different solution angles to this problem. Yet, very few papers deeply dive into the different characteristics at hand regarding OOD data, namely label and domain shifts. In this work, we introduce Activation Pruning under Extraneous Shifts (APES), a novel post-hoc activation shaping OOD detection technique that identifies and prunes activation channels sensitive to domain shifts in data. By leveraging the observation that visually perturbed in-distribution (ID) data exhibits score distributions similar to OOD data, APES selectively removes channels highly influenced by domain shifts. This allows the model to focus on label-specific activations. Our method is lightweight, easily integrates with existing activation shaping techniques, and enhances OOD detection performance. Extensive benchmarking on standard datasets demonstrates that APES, when combined with other activation shaping methods, achieves state-of-the-art results, improving the separation between ID and OOD score distributions.

1. Introduction

The field of deep learning has made notable advancements in recent years, towards more capable and higher-performing models. Aside from a purely research driven standpoint, many approaches were designed to be applicable in real-world scenarios. However, for most applications, the model training can not cover every possible input. Further, if deployed in an open environment, inputs outside the original class taxonomy are to be expected. In most cases, when encountering such an input, the model is prone to fail or make poor predictions due to its inexperience with such data. This limits the actual number of critical real-world use cases to employ these deep learning techniques. To give a rudimentary example, an autonomous driving system should be able to steer and react to cars breaking in front of it, but might not be able to recognize a tree blocking the road. This problem is addressed by the field of out-of-distribution (OOD) detection. Here, methods try to detect inputs outside the model's label space, for example by scoring the generated output of the model [9], modifying the entire training process [13, 26] or making post-hoc adjustments to the model itself [15]. For these methods, we consider data encountered during the training phase and that

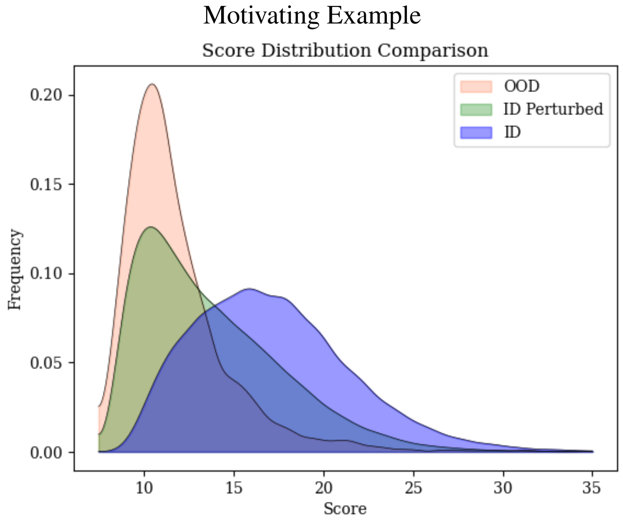

Figure 1. **Exemplary Score distributions**. Comparison for in-distribution (ID), perturbed in-distribution (ID perturbed) and out-of-distribution (OOD) data calculated on the Textures [4] dataset.

belongs to our known classes to be in-distribution (ID) data. Unseen data points, that do not conform to this designated class distribution, constitute OOD data. We can consider this as a binary classification problem with classes ID and OOD, in which we are trying to decide based on the calculated score. A better separation of the score distributions of the two classes (Fig. 1) results in a better classification of a given input.

Selectively lowering confidence on OOD data and separating the distributions is no trivial task to be left to the model. The confidence produced is often artificially raised, due to phenomena such as single neurons overcompensating for low activation sums [19]. Numerous approaches have been derived to address this problem, such as reshaping the internal activations of models to achieve an easier separable result, which is called activation shaping [10, 19, 27, 32].

Similarly, our approach follows this method by pruning specific channels. For this, we employ a unique strategy to select channels for pruning. This paper's novel method, Activation Pruning under Extraneous Shifts (APES), utilizes the difference in characteristics or domains between normal and perturbed ID data. An exemplary score difference is shown in Fig. 1. Previous methods in this area mainly focused on distinguishing data by reducing activation compensations or over-representation. However, differences in domains offer further vital information to be used for the classification process.

The motivation of the APES method originates from a key observation: Applying visual perturbations to ID data shifts the score distribution towards the score distribution of OOD data as shown in Fig. 1. This observation allows us to conclude that the domain of the images may influence the model's confidence. This is undesirable, as domain shifts may also occur in-distribution, and OOD data is primarily defined as the change in label from known classes to unknown ones. Therefore, the model should not utilize such unreliable and potentially noisy information about domain shifts when performing OOD detection. APES identifies the channels that are susceptible to such information and prunes them. This results in noise reduction and allows higher weighting of channels focusing on the detection of different labels.

The key contribution presented in this paper is a novel, post-hoc and lightweight activation pruning method that utilizes underlying OOD data characteristics to effectively separate ID and OOD score distributions. APES easily allows joint employment together with other OOD methods, which also utilize activation shaping. It achieves state-of-the-art results in combination with methods like Ash [10] or SCALE [26]. Furthermore, we provide an extensive benchmark against other high-performing OOD detection methods [10, 19, 26] and insights into detailed ablation studies on hyperparameters of this method. Lastly, we also include some new insights into the difference between far-OOD and near-OOD detection datasets, both on performance and domain information

2. Related Work

At first, this section defines preliminary information regarding this work, which includes insights into OOD as a whole and its methods.

OOD Detection Suppose a classifier, with input space $\chi = \mathbb{R}^d$ with d the number of input dimensions and an output space of $\mathbb{R}^n$ with n as the number of classes of the training set. The training has been performed on a predefined data distribution $\mathcal{P}_{in}$ over χ. Data points from this distribution are considered ID. Furthermore, we define a scoring function $s : \chi \to \mathbb{R}_+$, which is used to score the model's output. The goal of OOD detection is to decide whether an input $x \in \chi$ belongs to $\mathcal{P}_{in}$ or belongs to a disjunct distribution. This can be done via formulating a simple decision criterion λ so that for an input $x \in \chi$

$$decision_\lambda(x) = \begin{cases} \mathcal{P}_{in} & s(x) \geq \lambda \\ \mathcal{P}_{out} & s(x) < \lambda \end{cases}$$

The goal is to separate $\mathcal{P}_{in}$ and $\mathcal{P}_{out}$ to allow for a clear decision.

OOD Detection Methods Recent OOD detection methods can be categorized by certain features. One distinction is whether the existing classifier is retrained, to achieve better OOD detection performance (training-time modification) [13, 26], or the model outputs are post-processed to facilitate the derivation of a decision criterion (post-hoc) [15, 26]. Another important distinction is where the method is utilized. Approaches like energy-based methods [13] modify the model output to distinguish OOD inputs. Other papers focus on tuning the inner model representation by shaping activations [10, 19, 32] or pruning weights [1, 18] of internal layers.

Post-Hoc OOD Detection Setup The post-hoc OOD Detection setup is defined by 4 steps. First, a model, such as a classifier, is trained on the ID data. After training, the parameters are frozen. Second, the model is inferring on OOD data, as opposed to previously. Using that, the third step is to use the model as a binary classifier for the classes ID and OOD. This is done by computing a designated detection score for an input, for example, the energy score [13] or entropy [6]. The key idea is to generate scores that differ between ID and OOD data, resulting in a separation of the score distributions. The more distinctly the two score distributions are separated, the better the model can detect OOD data. The fourth step evaluates the model's ability to detect OOD data. Commonly used metrics are the FPR95 and the AUROC. Some approaches that are commonly used in post-hoc settings are energy-based methods [13, 19], distance-based methods [11, 20, 24], but also activation shaping methods [10, 19, 32]. Other approaches include gradient-based methods [3, 17].

Activation Shaping A frequent and logical observation is that neuron activation behavior is highly dependent on the data distribution [10, 14, 19, 23]. This means that model activations or the inner representations are potentially acting in different ways with ID data compared to encountering OOD data. Therefore, augmenting neuron activations in certain ways may serve as an effective way to guide the model's decision process towards a better OOD detection performance. In particular, post-hoc methods may take advantage of the effectiveness this approach provides, as it can be done non-invasive after training and does not require any weight alteration.

The two main ways of applying activation shaping are called rectification and pruning-based approaches. Rectification approaches [19, 32] try to clip or norm activations of certain layers to adjust the unusual activation pattern resulting from OOD data. Activation pruning methods [10, 18], on the other hand, focus on the premise that the internal representation is overloaded and sparsification of it may help the utilization of vital information and benefit the OOD detection process. Djurisic et al. 2023 [10] with their proposed method, Ash achieves sparsification with a top-k approach on the activation magnitude of the penultimate layer in the network.

Full-Spectrum Experiments An insightful experimental setup is the ”full-spectrum experiment”. This experiment also consists of two bins of data that the model has to distinguish between. One classically contains the OOD data. The key distinction is that the other bin not only contains ID data, but also augmented versions of the same ID data points. This increases the difficulty of OOD detection, since the scores of perturbed ID and OOD images behave more similar than in normal OOD setting. Articles like Yang et al., 2023 [28] have argued for this to be a more realistic problem setting, presenting a greater challenge than isolated OOD environments.

3. Activation Pruning under Extraneous Shifts

APES is based on the aforementioned observation that if the score distributions of perturbed ID data approach the distribution of OOD data. This is shown in Fig. 1. Current methods fail to effectively distinguish between the two, even though the base ID data is the same.

Intuition To explain the idea behind our approach, we need to address two vital characteristics of how OOD data differs from ID data. Firstly, by definition, they possess different label distributions. In the training process a classifier is given a set of labels and learns a mapping from a given input to one of the defined labels. But when there is no correct label for a given input, the model can not produce a correct output due to its output label space being defined by the training set. This change from an ID to an OOD class is called a **label shift**. Secondly, for a given input is to be in a different domain. A data point of visual modality typically has characteristics that define it, for example, a scene with sunshine. If these features were to change, e.g. a dark and rainy scenery of the same landscape, it is called a **domain shift** or **covariate shift** [28]. While label shifts occur per definition between ID and OOD data, it is common for domain shifts to not only be present in this relation but also exist between data inside the ID distribution.

As stated in Sec. 2, we define a visual augmentation or perturbation of an image to be representative of a domain shift. Using our observation of distribution shift, we infer that applying domain shifts to ID data makes it harder for the OOD detection process to separate the data. Therefore the detection process also relies, at least partially, on domain shift data for differentiating the inputs. This is entirely undesirable, as a domain shift is not guaranteed to occur for OOD data, and also may occur inside the ID distribution. Furthermore, we are led to believe that such information may act as noise and lessen the importance given to information about label shifts, the latter being what we want to detect. Finally, our method aims at detecting channels that are sensitive to domain shifts and prunes them, ideally leaving mostly channels that convey label shift information. As a result, the remaining channels are more likely to encode label-relevant features, improving the model's ability to distinguish between different classes. Even in cases where a pruned channel contains some label-relevant information, its removal is still beneficial if it is also highly sensitive to domain shift. By eliminating such mixed-significance channels, our approach directs focus toward those channels that encode label information with minimal interference from domain shift, ultimately leading to more robust OOD detection.

3.1. Our Approach

Our proposed method consists of 2 main steps as displayed in Fig. 2. Firstly, identifying the channels that are more sensitive to domain changes and creating a mask that prunes them and thus diminishes their influence on the OOD score. This step is done before the detection takes place. The second step is performed during the OOD detection. While passing the input through the model, the mask is applied to a layer in the model, altering its activations. The placement in this work is the penultimate layer of our model structure,

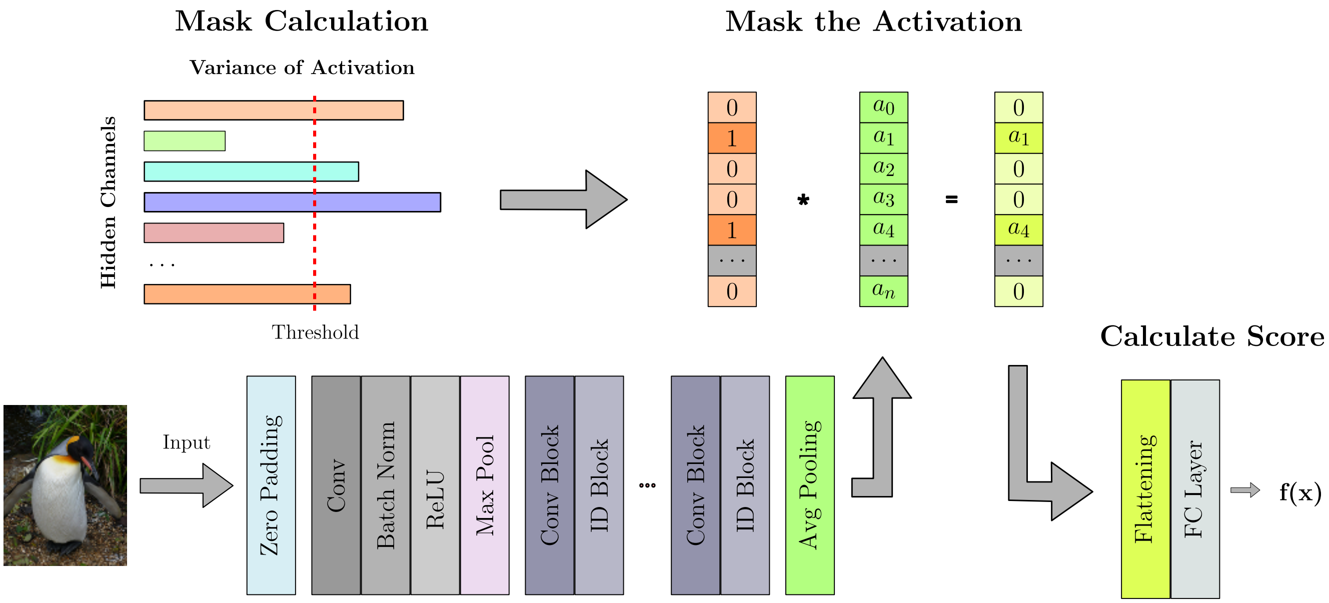

Figure 2. **APES method applied to the ResNet50 [7] architecture.** A mask is precomputed to remove the top p channels with the highest variance. The activations resulting from the penultimate layer are then multiplied by the mask, removing the activations from high domain-sensitive channels. The result of this multiplication is further processed with the fully-connected (FC) layer. Finally, the OOD score [13] is calculated.

inspired by the ablation study on activation shaping effectiveness in other papers [10, 19]. The activations are then processed as before, with the possibility of applying other activation shaping approaches afterwards. The next 2 sections describe the two steps in more depth.

Creating the Mask To create the mask, data with applied perturbations is required to find sensitive channels to the domain shift. The used perturbations and images are taken from the ImageNet-C [8] dataset and are also listed in the supplementary material. For the perturbations used for the final result, we refer to Sec. 4.3. Next, for each channel of the penultimate layer, the variance of activations for the perturbed images is calculated, resulting in a list of variances. If the variance is low, the perturbations applied to the image do not affect the channel significantly, and therefore, it is considered to be resistant to domain shifts. On the other hand, if the variance is high, some of the used perturbations affect the channel and therefore, the channel is sensitive to domain change. Figure 3 displays the channel activations for an arbitrary perturbation on the ID dataset. It is visible that some activations are now much more volatile than others, making these the targets for pruning. Next, the variances are used to create a mask. Therefore, the top p-percentile of the variances is calculated. A binary mask is created which has a 1 where the variance is lower than the p-percentile and a 0 where it is above. The mask is saved for the OOD detection process and not changed further.

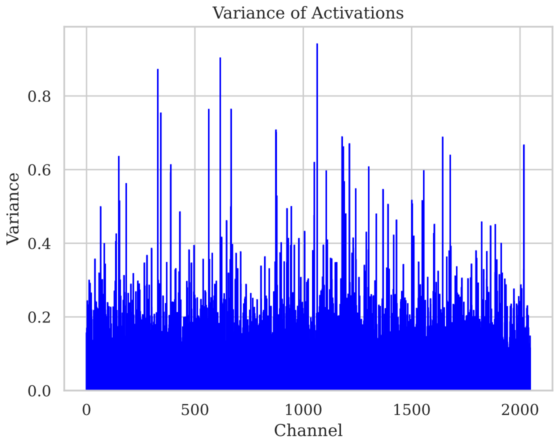

Figure 3. **Channel-wise variance of penultimate layer.** Variance in activations for the penultimate layer of the ResNet50 [7] architecture, trained on ImageNet-1k [5]. The variances were captured on the *brightness* perturbation from ImageNet-C [8]. Some channels show a high variance in activation and thus susceptibility to perturbed ID data.

4. Experiments

To show that APES improves on the OOD detection, we conducted three main experiments. First, we compared our OOD detection results of APES against other current approaches. Second, we repeated those experiments while combining APES with other activation shaping methods [10, 26], to investigate the joint performance on the same benchmark. Ultimately, we also repeated our experiments in a full-spectrum setup. Moreover, we can determine optimal pruning percentages as well as which perturbations to use. This is done in Sec. 4.3, where the effect of different perturbation masks is analyzed first, followed by an ablation study to find the best pruning percentage. The results shown use the values that were justified in the ablation section.

4.1. Experimental Setup

To test our method and the combination with different methods we had the following experimental setup.

Datasets and Models As is common in recent literature, we used ImageNet-1k [5] as our ID dataset. For OOD, we divided the experiments into two parts. First, testing the far-OOD datasets. This experiments used the datasets iNaturalist [21], SUN [25], Places365 [31] and Textures [4]. The second part was conducted on the near-OOD datasets. This experiment was performed on SSB-hard [22] and NINCO [2]. With respect to far-OOD datasets, we employed the same setup as ReAct[19] or Ash[10]. In addition to the near- and far-OOD datasets, we also applied perturbations to the ID images from ImageNet-C [8] for our full-spectrum detection experiment and to determine the hyperparameters of our model. All ablation studies that do not explicitly mention other datasets were performed on OpenImage-O [24] as our validation dataset. As for the model architectures, we used ResNet50 [7] and MobileNetV2 [16] to conduct the experiments. Both networks are pre-trained on ImageNet-1K, and the weights and parameters are frozen during OOD-detection. To compare with the other methods on near-ood datasets and test out joint performance, we reimplemented React, Ash and SCALE. In the result tables, we used the results presented in the respective papers if available, and otherwise utilized our recomputed values. Seeding is not as relevant in this project because there is no training or other random process involved may depend on randomness.

Metrics To be consistent with previous methods [10, 12, 13, 30], we evaluate APES by showing the false positive rate at a fixed true positive rate of 95% (FPR95). For this, lower scores represent better results. Additionally, we also use the area under the receiver operator characteristic (AUROC) curve to show the probability that any ID sample will have a higher score attributed to it than an OOD sample. Here, higher values are desirable.

4.2. Results

Far OOD When running the experiments described, APES as a standalone method performs mostly better than the baseline method (energy score). The improvement is 19.36 on the FPR95 average and 4.06 on the AUROC average. Another notable comparison is with Ash-P. It uses maximal activation pruning to increase the OOD detection performance and therefore has a similar approach. APES beats Ash-P by 10.96 on the FPR95 average and 1.19 on the AUROC average. This validates our approach of pruning channels based on maximal variance, actually increasing the OOD detection performance. However, when taking APES as a standalone method, the performance is worse than most of the state-of-the-art (SOTA) approaches. What furthermore stands out, is that APES significantly decreases the performance on the Textures dataset.

When combining our method with Ash-S, the performance increases drastically. In combination, the two methods outperform many of the other methods across all datasets. On the iNaturalist dataset, the combination performed best compared to current methods. The former best performance of 9.50 by SCALE is beaten by 2.21 with a value of 7.29. A similar improvement can be observed on the AUROC. For the SUN dataset, APES + Ash-S only falls short by 2.97 in the FPR95 and 0.14 on the AUROC to DDCS. Still, the combination stays a top-3 method for this dataset. On the Places dataset, the APES FPR95 is 1.46 worse than the LINE[1] performance but beats all other models on the AUROC, with an increase of 0.6 over DDCS[29]. This leads to a new best AUROC result of 93.49. As mentioned before, APES as a standalone method makes the performance on Textures worse. This trend can also be observed when combined with Ash-S. Together they have an FPR95 of 17.00 and an AUROC of 96.61, decreasing the performance for Ash-S alone. In terms of the average performance, Ash-S + APES in combination is highest among all SOTA methods. The average of the 4 datasets can be improved by 0.40 on the FPR95 and 0.44 on the AUROC.

Another combination tested was SCALE with APES. Here we see results similar to the previous combination, but improve over previously weaker performances such as the SUN AUROC or the performance on the Places dataset. The SUN average is increased by 0.38 and the Places AUROC by 0.40. Our metric average when utilizing SCALE jointly with our approach also beats all other approaches listed in this paper. The new best performances are an FPR95 average of 18.28 and an AUROC average of 96.11.

The same experiment was repeated with the MobileNet backbone. Here, none of the prior best results could be beaten. However, when comparing Ash-P against APES, Ash-S against Ash-S + APES, and SCALE against SCALE + APES, it can be observed that the results improve. As with the ResNet50 [7] backbone, APES performs better compared to the similar approach of Ash-P. When applying APES in combination with the other methods, the results show that both single results and the averages improve.

		OOD-Datasets									
	Methods	iNaturalist		SUN		Places		Textures		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet50	Energy score	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
	BATS	12.57	97.67	22.62	95.33	34.34	91.83	38.90	92.27	27.11	94.28
	Ash-P	44.57	92.51	52.88	88.35	61.79	85.58	42.06	89.70	50.32	89.04
	Ash-S	11.49	97.87	27.98	94.02	39.78	90.98	**11.93**	**97.60**	22.80	95.12
	fDBD	17.27	96.68	42.30	90.90	49.77	88.36	21.83	95.43	37.79	92.84
	VRA+	15.48	97.08	23.50	94.91	34.62	91.79	19.66	96.08	23.31	94.97
	LINe	12.62	97.56	19.48	95.26	28.52	92.85	22.54	94.44	20.70	95.03
	DDCS	11.63	97.85	**18.63**	95.68	28.78	92.89	18.40	95.77	19.36	95.55
	SCALE	9.50	98.17	23.27	95.02	34.51	92.26	12.93	97.37	20.05	95.71
	DICE + ReAct	18.64	96.24	25.45	94.55	36.86	90.67	27.25	92.90	27.05	93.40
	fDBD + SCALE	10.19	98.07	24.58	94.87	36.12	92.00	12.51	97.48	20.85	95.61
	APES	24.23	95.39	32.84	92.13	38.94	90.87	61.43	82.53	39.36	90.23
	Ash-S + APES	**7.29**	**98.54**	21.60	95.54	29.98	93.49	17.00	96.61	18.96	96.05
	SCALE + APES	**7.29**	98.48	19.63	**95.92**	**27.45**	**93.89**	18.77	96.14	**18.28**	**96.11**
MobileNet	Energy score	59.50	88.91	62.65	84.50	69.37	81.19	58.05	85.05	62.39	84.91
	ReAct	42.40	91.53	47.69	88.16	51.56	86.64	38.42	91.53	45.02	89.47
	DICE	43.09	90.83	38.69	90.46	53.11	85.81	32.80	91.30	41.92	89.60
	Ash-P	54.92	90.46	58.61	86.72	66.59	83.47	48.48	88.72	57.15	87.34
	Ash-S	39.10	91.94	43.62	90.02	58.84	84.73	13.12	97.10	38.67	90.95
	LINe	24.95	95.53	33.19	92.94	47.95	88.98	**12.30**	97.05	29.60	93.62
	SCALE	31.22	93.86	37.89	91.71	52.93	86.79	12.77	**97.34**	33.70	92.43
	DDCS	**17.44**	**96.87**	**17.42**	**95.83**	**30.49**	**91.80**	25.11	94.86	**22.61**	**94.84**
	APES	38.92	93.11	45.61	89.26	53.37	86.52	55.30	85.64	48.30	88.63
	Ash-S + APES	29.80	94.38	36.30	91.67	49.52	87.92	15.85	96.29	32.87	92.57
	SCALE + APES	25.79	94.89	31.81	92.77	46.43	88.38	12.91	97.24	29.74	93.32

Table 1. **Far OOD Detection Results.** FPR95 refers to the False Positive Rate at 95% True Positive Rate (lower is better), and AUROC refers to the Area Under the Receiver Operating Characteristic curve (higher is better). All values refer to percentages. Both ResNet [7] and MobileNet [16] were trained on the ID data (Imagenet-1k [5]) only. For MobileNet, we implemented "SCALE" [26] ourselves. Other values were taken directly from the respective papers.

	OOD-Datasets					
Methods	Ninco		SSB Hard		Average	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
Energy Score	77.61	79.69	84.43	72.08	71.02	75.89
ReAct	71.47	80.00	78.93	72.80	75.20	76.40
Ash-S	64.09	83.22	80.67	74.36	72.38	78.79
SCALE	**51.80**	**85.37**	67.72	77.35	**59.76**	81.36
APES	57.11	83.20	72.74	77.20	64.93	80.20
Ash-S + APES	54.99	85.08	66.84	78.00	60.92	81.54
SCALE + APES	61.07	84.31	**64.91**	**79.38**	62.99	**81.82**

Table 2. **Near OOD detection results.** FPR95 refers to the False Positive Rate at 95% True Positive Rate (lower is better), and AUROC refers to the Area Under the Receiver Operating Characteristic curve (higher is better). All values are percentages. Here, ResNet50 [7] was used as a backbone, trained on the ID dataset ImageNet-1k [5].

Near OOD As with the far-OOD results, we can see that APES outperforms the AUROC energy baseline by 4.31. When it comes to Ash-S + APES, the combination improves the results over the Ash-S results, which is reflected by an AUROC average increase of 2.75. However, when combining APES with SCALE, only the results on the SSB Hard dataset improve. The improvement is 2.03 on the SSB Hard AUROC, which leads to an AUROC average improvement of 0.46. The new AUROC average high is 81.82.

Full-Spectrum Setting The results of the full-spectrum setting are given in Sec. B in the supplemental material. It is visible how other approaches have a significant decrease in performance in this experimental setting. As supposed, this may result from the alteration of the ID set through data that is only domain shifted, with the same labels. This further solidifies the point of SOTA methods gaining a performance boost by considering the huge domain gaps between ID and OOD data. It is also visible that our novel method, APES, helps lessen these effects and leads to an increase in performance of all displayed methods when combined. This is especially visible for comparing maximum activation pruning from Ash-P to our maximum variance pruning, as we did in the other result sections.

4.3. Ablations

The APES method has 2 main hyperparameters, influencing its performance: The pruning percentage p, as well as the perturbations used for computing the mask. To choose the best combination of pruning percentage with the best perturbations, several analyses and ablation studies were conducted. The next sections are divided into a deeper analysis of the perturbation characteristics and an ablation of the pruning percentage.

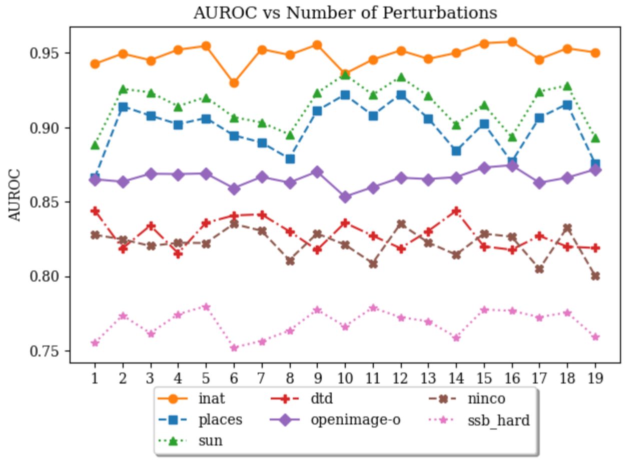

Figure 4. **AUROC for individual perturbations.** AUROC values captured for each ImageNet-C [8] perturbation individually across the OOD datasets. For the actual names of the perturbations, refer to Sec. B (”Arbitrary”) in the appendix. As the backbone architecture, we used ResNet50 [7] trained on ImageNet-1k [5].

Setup To evaluate the performance of the different combinations we kept ImageNet1k as ID dataset [5]. Firstly, we display an observation regarding each perturbation's domain shift on the OOD datasets, and secondly, we use the OpenImage-O [24] to conduct an ablation study. Further, all experiments regarding the hyperparameters were conducted using the ResNet50 [7] backbone. In terms of the evaluation metric we mainly used the AUROC because it gives a better overview of the performance compared to the FPR95. In some cases the latter can be found in the supplementary material in the appendix. ImageNet-C [8] provides 5 levels of severity for each perturbation. For simplicity and to limit the number of experiment runs, we limited our ablation study to severity 1.

Perturbation Analysis For choosing the perturbations several options are available. For instance, only a single perturbation, or even a combination of different perturbations, can be used for calculating a mask. However, when choosing multiple perturbations, the question arises which to combine.

To come up with a strategy on how to choose perturbations, it is first vital to remember what they are and how they are represented. As mentioned before, in our terminology, an applied perturbation equates to a domain shift. Our method tries to infer domain shift information through exemplary shifts through said perturbations. We also mentioned that OOD datasets are likely but not required to contain domain shifts for the ID data. As there are various ways to realize such shifts, it is to be assumed that different perturbations provide distinctively useful information for each dataset. If this holds, it is hard to choose a good subset of perturbations judging by a validation set, as we cannot infer from the validation set performance to other OOD datasets.

To show this, we conducted an in-depth analysis, creating masks using only a single perturbation each. Next, we ran the OOD detection process and compared the results. The results are shown in Fig. 4 and the respective perturbation combinations are detailed in Sec. B. It is visible that while some perturbations behave similarly for some datasets, like the ones at index 19, the overall pattern of performance and relative performance to its neighbors fluctuates between datasets, like for instance perturbations 6, 10 and 16. This is also supported by Fig. 6, which showcases the averaged AUROC performances in near and far OOD datasets respectively, as well as Fig. 7 and Fig. 9, which display the same effect for the FPR95 metric. Here it is even more telling that there seems to be no clear behavior for most of the perturbations, each OOD class average having different perturbations that would showcase the highest results. This strongly supports our hypotheses from earlier. We can also see this by the example that the SUN [25] and Places [31] datasets have nearly identical perturbation behavior, as they are very close together domain-wise and visually, but they completely differ from OpenImage-O [24] for instance. The difference in behavior across near and far OOD datasets furthermore supports our findings, indicating that not only the type of domain shift but also the severity of it may factor in. This conclusion also transcends our testing datasets, as the whole point of OOD detection is to prepare for novel inputs from unknown distributions. Hence, this effect is likely to be stronger if we add even more OOD datasets. As we want to find a generally applicable setting, in this case, it is rational to use as many perturbations as we can in this setting to diminish the trade-off in performance across different domains. This also increases the chance of robustness for other unknown data distributions potentially encountered in the future. Consequently, the findings of this analysis suggest a need for further ablation on the accumulation of perturbations in this case, giving us a reason to directly use all 19 present.

Pruning Percentage Ablation The parameter p describes the percentage of activations pruned. A pruning percentage of 0, therefore, corresponds to no pruning at all, keeping 100% of the activations. A pruning percentage of e.g., 35% means that 65% of the activations are kept and the channels with the highest 35% activation variance are pruned.

In order to choose the best percentage for our method, we conducted this ablation study on the OpenImage-O [24] to investigate the influence p has on the performance. We tested the performance for $p \in \{0, 5, 10, \ldots, 95\}$ over all 19 perturbations from ImageNet-C [8]. The results are presented in Fig. 5. For every pruning percentage, we mea-

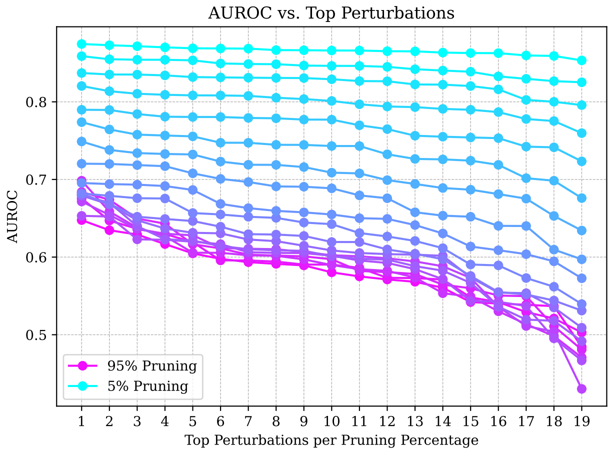

Figure 5. **Ablation study on pruning percentage** p. For each p, we list the top performing perturbations in descending order on the x axis. Lower pruning percentages correspond to cyan values, while higher pruning percentages correspond to magenta. Lower pruning percentages perform better. ResNet50 [7] trained on ImageNet-1k [5] was used as the backbone and the OpenImage-O [24] as the OOD dataset.

sured the AUROC over each perturbation in ImageNet-C. The perturbations under lower pruning percentages (cyan) perform generally better compared to applying high pruning percentages (magenta). This can be observed across all perturbations. The best overall performance could be achieved with $p = 5$, being the pruning percentage used in the presented results below.

4.4. Discussion

As is visible, our novel method helped achieve a new state-of-the-art on average over multiple OOD datasets. This was especially true on far-OOD but also on near-OOD datasets. The results show that domain shifts are an interesting characteristic of each OOD dataset, which should not be discarded when trying to find optimal solutions for OOD detection. Moreover, we showcased the partial reliance on domain shift information of other popular methods, as well as the ability of APES to overcome and improve on those drawbacks, in our full-spectrum experiment setting.

Limitations The results show that the APES method performs significantly worse on the textures dataset compared to the energy score baseline and other methods. This decline in performance is explainable by the unique nature of the dataset. The textures dataset consists of images depicting various textures with differences in colors, shapes, and patterns. In this dataset, label shift information closely resembles the effects of applying perturbations to the images, resulting in domain shifts being closely related to label shifts. This similarity creates a fundamental challenge for the APES method. Since APES identifies and removes channels sensitive to domain shift, it inadvertently removes channels that also encode vital label information. In other words, the method does not effectively separate domain shift from label shift in such datasets, as both are represented similarly. As a result, pruning these channels leads to a loss of essential label information, preventing APES from improving performance.
Additional drawbacks to this method are not severe but relevant. It stands to argue that the analysis of perturbations done in Sec. 4.3 was only the start of an investigation into a potentially larger problem that persists. As shown, the nature of perturbations and their varying interaction with different datasets make it hard to know how to select the best ones. Still, there is, a danger of declining performance if we choose too many perturbations, and thus create too much domain shift information in the ID dataset. Therefore, it may be possible that the results shown here can be elevated even more after an in-depth investigation of domain shift characteristics between OOD datasets and a refined selection process of perturbations, to optimize any trade-offs.

5. Conclusion

With our novel method, APES, we have further improved the performance of the OOD detection of common near- and far-OOD datasets, providing a simple post-hoc approach of activation shaping. With that we also further explored the concept of domain and label shifts in OOD settings. As demonstrated, APES combined with the energy score [13] increases OOD performance, compared to only using the latter. This warrants the inclusion of this useful concept of distinguishing domain and label shift information when designing future approaches. Moreover, our approach is highly compatible with other activation shaping methods, enabling even higher OOD detection performance. The combination of Ash [10] or SCALE [26] and APES outperforms current state-of-the-art methods on most OOD datasets, constituting the highest score on average and in most cases. Lastly we hope to encourage a new focus on the underlying characteristics of common OOD test datasets, as our analysis and subsequent results suggest them to be relevant in the performance of OOD methods.

References

[1] Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. Line: Out-of-distribution detection by leveraging important neurons. In CVPR, pages 19852–19862, 2023. 2, 5

[2] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In ICML, pages 2471–2506, 2023. 5

[3] Jinggang Chen, Junjie Li, Xiaoyang Qu, Jianzong Wang, Jiguang Wan, and Jing Xiao. Gaia: Delving into gradient-based attribution abnormality for out-of-distribution detection. NeurIPS, 36:79946–79958, 2023. 3

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In CVPR, pages 3606–3613, 2014. 1, 5

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248–255. Ieee, 2009. 4, 5, 6, 7, 8, i

[6] Krzysztof Gajowniczek, Leszek J Chmielewski, Arkadiusz Orłowski, and Tomasz Zabkowski. Generalized entropy cost function in neural networks. In ICANN, pages 128–136. Springer, 2017. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. 4, 5, 6, 7, 8, i

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. ICLR, 2019. 4, 5, 7, i

[9] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. ICL, 2016. 1

[10] Sunwoo Kim, Jaehong Yoon, and Sung Ju Hwang. Extremely simple activation shaping for out-of-distribution detection. In ICLR, 2023. 2, 3, 4, 5, 8

[11] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NeurIPS, 31, 2018. 3

[12] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In ICML, pages 31728–31746, 2024. 5

[13] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. NeurIPS, 33: 21464–21475, 2020. 1, 2, 3, 4, 5, 8

[14] Yibing Liu, Chris Xing Tian, Haoliang Li, Lei Ma, and Shiqi Wang. Neuron activation coverage: Rethinking out-of-distribution detection and generalization. CVPR Spotlight, 2023. 3

[15] Aitor Martinez-Seras, Javier Del Ser, and Pablo Garcia-Bringas. Can post-hoc explanations effectively detect out-of-distribution samples? In FUZZ-IEEE, pages 1–9. IEEE, 2022. 1, 2

[16] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In CVPR, pages 4510–4520, 2018. 5, 6

[17] Sina Sharifi, Taha Entesari, Bardia Safaei, Vishal M Patel, and Mahyar Fazlyab. Gradient-regularized out-of-distribution detection. In ECCV, pages 459–478. Springer, 2024. 3

[18] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In ECCV, pages 691–708. Springer, 2022. 2, 3

[19] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. NeurIPS, 34: 144–157, 2021. 2, 3, 4, 5

[20] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In ICML, pages 20827–20840. PMLR, 2022. 3

[21] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In CVPR, pages 8769–8778, 2018. 5

[22] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? 2021. 5

[23] Weilin Wan, Weizhong Zhang, Quan Zhou, Fan Yi, and Cheng Jin. Out-of-distribution detection using neural activation prior. arXiv, 2024. 3

[24] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In CVPR, pages 4921–4930, 2022. 3, 5, 7, 8

[25] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In CVPR, pages 3485–3492. IEEE, 2010. 5, 7

[26] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In ICLR, 2024. 1, 2, 5, 6, 8

[27] Mingyu Xu, Zheng Lian, Bin Liu, and Jianhua Tao. Vra: variational rectified activation for out-of-distribution detection. NeurIPS, 36:28941–28959, 2023. 2

[28] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. IJCV, 131(10):2607–2622, 2023. 3

[29] Yue Yuan, Rundong He, Yicong Dong, Zhongyi Han, and Yilong Yin. Discriminability-driven channel selection for out-of-distribution detection. In CVPR, pages 26171–26180, 2024. 5

[30] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. DMLR, 2024. 5

[31] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6):1452–1464, 2017. 5, 7

[32] Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue, Xiang Tian, Yaowu Chen, et al. Boosting out-of-distribution detection with typical features. NeurIPS, 35:20758–20769, 2022. 2, 3

Pruning-based OOD Detection via Activation Consistency Under Extraneous Shifts

Supplementary Material

A. Ablation Material

In this section, additional material on the ablation can be found. This includes FPR95 plots, corresponding to the AUROC plots from the main work.

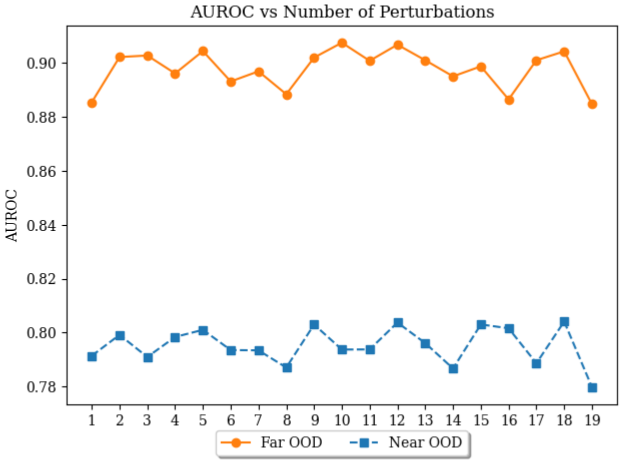

Figure 6. **Average AUROC for perturbations.** Average AUROC performance of each individual perturbation taken over each near and far OOD dataset. We used ResNet50 [7], trained on ImageNet-1k [5] only, as the backbone architecture.

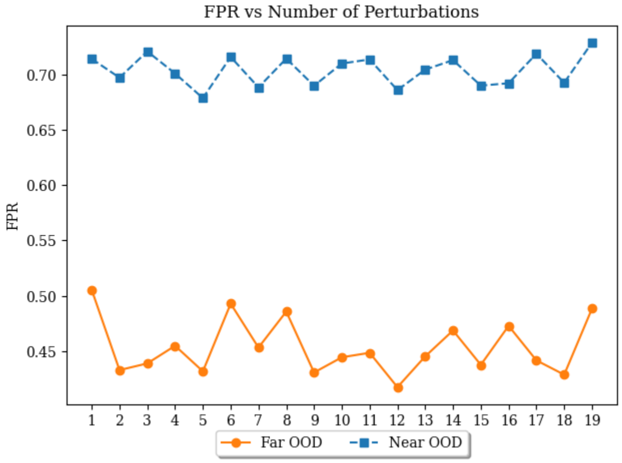

Figure 7. **Average FPR for perturbations.** Average FPR performance of each individual perturbation taken over each near and far OOD dataset. We used ResNet50 [7], trained on ImageNet-1k [5] only, as the backbone architecture.

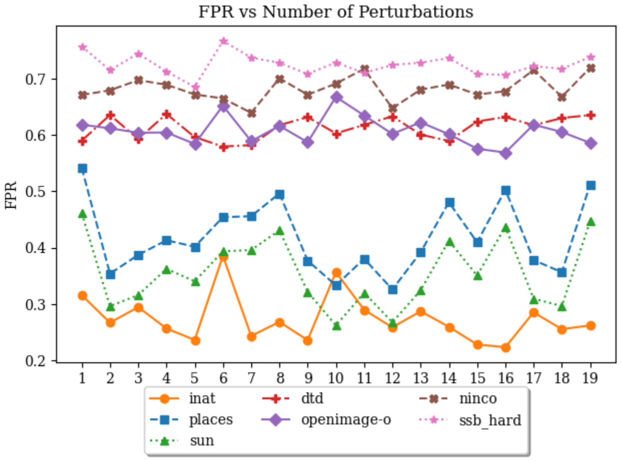

Figure 8. **FPR for individual perturbations.** FPR values captured for each ImageNet-C [8] perturbation individually across the OOD datasets. For the actual names of the perturbations, refer to Sec. B ("Arbitrary"). We used ResNet50 [7], trained on ImageNet-1k [5] only, as the backbone architecture.

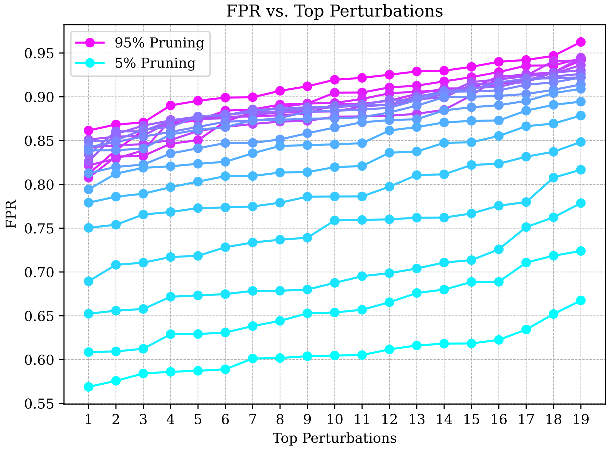

Figure 9. **Ablation study on pruning percentage** p**.** For each p, we list the top performing perturbations in descending order on the x axis. Lower pruning percentages correspond to cyan values, while higher pruning percentages correspond to magenta. Lower pruning percentages perform better. We used ResNet50 [7], trained on ImageNet-1k [5] only, as the backbone architecture.

B. Additional Results

https://github.com/PaulRabich/RectPruneOOD

OOD Dataset	**iNaturalist**		**SUN**		**Places**		**Textures**		**Ninco**		**SSB Hard**		**Average**	
Metric	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
Baseline	92.24	75.37	91.82	70.62	93.71	66.86	84.17	72.41	97.12	60.56	97.55	52.44	92.77	66.38
Ash-S	44.28	91.11	66.95	82.31	77.37	76.27	**40.87**	**91.00**	89.35	62.90	95.13	50.49	68.99	75.68
Ash-P	97.93	71.07	97.30	65.47	97.29	62.09	83.53	74.65	97.91	59.25	98.67	49.35	95.44	63.65
SCALE	40.97	92.03	62.05	84.57	73.44	78.86	43.17	90.41	88.16	63.57	94.65	53.24	67.07	77.11
APES	65.48	85.31	69.65	80.58	77.73	77.68	86.54	65.82	91.14	65.06	**93.28**	**58.48**	80.64	72.15
Ash-S + APES	35.06	93.29	59.11	85.63	70.56	81.20	51.65	88.21	86.92	65.95	93.84	55.14	66.19	78.24
SCALE + APES	**34.88**	**93.41**	**54.71**	**87.12**	**66.05**	**82.83**	52.71	87.62	**86.22**	**65.99**	93.52	57.76	**64.68**	**79.12**

Index Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Arbitrary	Br	DBl	Fog	GaBl	GlBl	JPEG	Pix	ShN	Sp	ZoBl	Con	ET	Fr	GaN	IN	MBl	Sat	Sn	SN
Analysis	IN	JPEG	Fog	MBl	GlBl	Sn	SN	GaN	ET	Fr	Sat	Sp	Con	ShN	GaBl	DBl	Pix	Br	ZoBl
5 Percent Pruning	DBl	GaBl	ZoBl	ET	GlBl	Fog	MBl	SN	ShN	Br	Sp	Con	IN	JPEG	GaN	Pix	Sat	Sn	Fr

Table 3. **Order of ImageNet-C perturbations in the experiments.** This table shows the names of the imagenet-c perturbations used in the experiments and also the respective order for different ablation study settings. The name on the left most column encodes the methods used, as well as the pruning percentages. The perturbations are abbreviated with the following symbols: Brightness=Br, Defocus Blur=DBl, Fog=Fog, Gaussian Blur=GaBl, Glass Blur=GlBl, JPEG Compression=JPEG, Pixelate=Pix, Shot Noise=ShN, Spatter=Sp, Zoom Blur=ZoBl, Contrast=Con, Elastic Transform=ET, Frost=Fr, Gaussian Noise=GaN, Impulse Noise=IN, Motion Blur=MBl, Saturate=Sat, Snow=Sn, Speckle Noise=SN