Improving robustness against common corruptions with frequency biased models

Tonmoy Saikia
University of Freiburg
saikiat@cs.uni-freiburg.de

Cordelia Schmid
Inria
cordelia.schmid@inria.fr

Thomas Brox
University of Freiburg
brox@cs.uni-freiburg.de

Abstract

CNNs perform remarkably well when the training and test distributions are i.i.d, but unseen image corruptions can cause a surprisingly large drop in performance. In various real scenarios, unexpected distortions, such as random noise, compression artefacts or weather distortions are common phenomena. Improving performance on corrupted images must not result in degraded i.i.d performance – a challenge faced by many state-of-the-art robust approaches. Image corruption types have different characteristics in the frequency spectrum and would benefit from a targeted type of data augmentation, which, however, is often unknown during training. In this paper, we introduce a mixture of two expert models specializing in high and low-frequency robustness, respectively. Moreover, we propose a new regularization scheme that minimizes the total variation (TV) of convolution feature-maps to increase high-frequency robustness. The approach improves on corrupted images without degrading in-distribution performance. We demonstrate this on ImageNet-C and also for real-world corruptions on an automotive dataset, both for object classification and object detection.

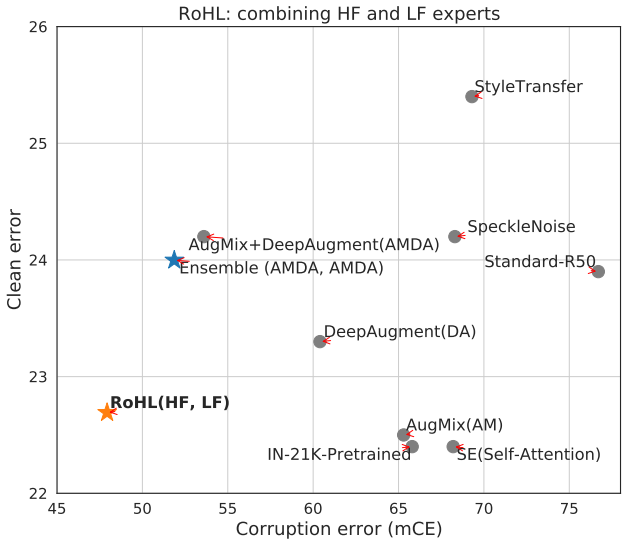


Figure 1: Improving clean and corruption errors. Each item shows the error of a model on ImageNet (y-axis) and on ImageNet-C (x-axis). All models use a ResNet50 backbone. Orange: The proposed **RoHL** approach – **Ro**bust mixture of a **H**F (high-frequency) and a **L**F (low-frequency) expert model. Blue: An ensemble trained with the state-of-the-art approach AugMix + DeepAugment. Gray: Other approaches.

1. Introduction

Robustness to distribution shift is possibly the core challenge in deep learning. CNNs show strong performance when training and test set samples are independent and identically distributed (i.i.d). This led to strong claims of obtaining superhuman performance on the challenging ImageNet dataset. However, such claims have somewhat diminished as the community, driven by practical applications, started testing on out-of-distribution (OOD) test sets. Unlike human vision, CNNs are affected even by small perturbations in the input. Simply adding random noise to the ImageNet test set is sufficient to almost triple the classification error [15].

Why does performance drop so severely under distribution shift? One explanation is that models rely on spurious, unstable correlations present in the i.i.d training and test dataset to obtain low training and test errors. When, due to distribution shift, these unstable correlations are missing, performance drops severely. Although there has been substantial prior work [12, 15, 24, 29, 35] investigating this problem, it is far from being fully understood, let alone solved. The most successful remedies to-date are well-chosen data augmentation schemes [7, 14, 17, 26, 11] and adversarial training [10, 26, 32]. Geirhos *et al*. [12] proposed the *texture hypothesis*, where they show that classification models learn feature representations biased towards textures. Many of these texture features are unstable and get destroyed, for example, due to weather effects or digital corruptions.

The *texture hypothesis* can also be regarded from a Fourier perspective [35]. Yin *et al*. [35] showed that models achieve reasonable performance (∼60% accuracy) on the i.i.d test set of ImageNet even with strong low or high pass filtering applied to the input images during training and testing.

This indicates the existence of many input-output correlations in low-frequency and high-frequency domains. They also showed that the performance degradation on corrupted data varies across the frequency spectrum. For instance, standard models trained on clean images are inherently biased to be more robust towards low-frequency corruptions compared to high-frequency ones. It might seem that such biases can be easily fixed with data augmentation. However, data augmentation comes with robustness trade-offs, i.e., many transformations improve performance on some types of corruptions but reduce performance on clean images. In realistic scenarios, the dominant fraction of data is typically clean and not corrupted. Therefore, clean performance must not be ignored.

To avoid such trade-offs, we propose **RoHL** — **Ro**bust mixture of a **H**F (high-frequency) and a **L**F (low-frequency) expert model. To build the HF expert model, we apply TV minimization [2] on the activations of the first convolutional layer, as well as generic augmentations that affect high-frequency components in the image. The HF expert is robust to high-frequency corruptions whereas the LF expert, based on plain contrast augmentation, is robust to low-frequency corruptions. We show that having such complementary models improves performance both on corrupted and clean images. Also compared to a standard two-member ensemble it adds robustness at no additional cost. An overview of its effectiveness is shown in Fig. 1.

In summary, we make two contributions: **(1)** We propose a new regularization scheme that enforces convolutional feature maps to have a low total variation (TV). We show that this boosts high-frequency robustness and is complementary to other high-frequency augmentation operations. **(2)** We introduce the idea of mixing two experts that specialize in high-frequency and low-frequency robustness. We show that this mixture is complementary to diverse data augmentation, such as AugMix [17] and DeepAugment [14].

2. Related work

Lack of robustness under distribution shift. Geirhos *et al*. [12] and Vasiljevic *et al*. [31] showed that models trained against certain distortions often fail to generalize to unseen distortions. Hendryks and Dietterich [15] proposed a synthetic benchmark (ImageNet-C) to study robustness against diverse image distortions. Recht *et al*. [24] recreated a new "ImageNetV2" validation set to benchmark naturally occurring domain shift over time and observed larger performance drops. Recent works evaluated performance under distribution shifts for other vision tasks such as object detection [21] and segmentation [18], with similar conclusions.

Vulnerability to adversarial perturbations. Adversarial perturbations [4, 28] are crafted noise signals designed to maximally confuse a model. These perturbations are categorized into white-box attacks [8, 20, 22, 23, 28], where the attacker has accessibility to model weights and gradients and black-box attacks [3, 6, 9], where the attacker can only query the model. Here, we focus on robustness to common corruptions, encountered even without an adversary.

Improving robustness. Hendryks *et al*. [16, 14] showed that pre-training on large datasets such as ImageNet-21k improves robustness. Xie *et al*. [34] trained large models on ImageNet and YFCC100M [30] in a semi-supervised manner to obtain improved i.i.d and OOD performance. Taori *et al*. [29] claimed that larger datasets improve performance on OOD data, but are far from closing the performance gap. An effective measure to improve OOD performance is data augmentation. Ford *et al*. [10] observed that augmentation techniques such as Gaussian or adversarial noise bias the model to be robust against certain corruption types, while degrading on others. Yin *et al*. [35] showed that these trade-offs can be better understood by looking at the Fourier statistics of the different corruption types. Geihos *et al*. [11] showed that using stylized images for training increases shape-bias and thus, improves robustness. Rusak *et al*. [26] studied noise corruptions and established a strong baseline on ImageNet-C. Rusak *et al*. [26] also evaluated the impact of feature denoising combined with adversarial training [33] on robustness to common corruptions. Hendryks *et al*. [17] showed that diverse data augmentation can obtain strong results on the ImageNet-C benchmark. Recently, Schneider *et al*. [27] showed that performance can be further improved by adapting batch-norm statistics at test-time.

3. Effect of data augmentation on robustness

3.1. Robustness trade-offs of data augmentation

High frequency robustness. It has been shown that models trained with Gaussian noise or adversarial training exhibit improved resilience to corruptions that affect the high frequencies of the signal [35]. Such corruptions include different noise corruptions like Gaussian or salt-and-pepper noise. Also corruptions that include blur affect the high-frequency components, as they diminish high-frequency image features such as edges. Data augmentation with operations that act on the high-frequencies make the trained model rely less on high-frequency features and have been shown to improve robustness to corruptions concentrated in the high-frequency spectrum considerably. However, as they remove high-frequency features from the model, they also reduce performance on clean images considerably.

Low frequency robustness. Achieving robustness to low frequency corruptions, such as fog, haze, contrast, is less obvious compared to high-frequency robustness. Natural images are inherently dominated by the low-frequency components. Yin *et al*. [35] showed that a data augmentation approach such as randomly perturbing Fourier components with magnitudes sampled from a low-frequency corruption type does not improve low-frequency robustness. The perturbation destroys natural image statistics and even degrades performance on corruptions such as fog. They claimed that no clear trade-off exists for low frequency corruptions. We investigate this further in Sec. 5.3.

3.2. Diverse data augmentation

A way to circumvent the above trade-offs is the application of diverse data augmentation transformations, which has been shown to improve robustness across the frequency spectrum [17, 35]. AugMix and DeepAugment are two such data augmentation methods.

AugMix. AugMix [17] composes image transformations from a variety of augmentation operations [7]. It involves sampling k random sequences of augmentation operations, resulting in k augmented images. These augmented images are then mixed element-wise with randomly sampled weighting factors. A final image is obtained by mixing the augmented image again with the clean version. AugMix models are trained with an additional consistency loss to enforce similar responses for the clean and augmented image embeddings. In particular, the Jensen-Shannon divergence (JSD) among the posterior distributions of the original sample and its augmented variants is minimized.

DeepAugment. DeepAugment [14] uses encoder-decoder networks trained for image superresolution and image compression to generate augmented images. Distorted images are generated by passing an image through these networks but with the weights being perturbed by random transformations. The distorted images are precomputed before using them for training.

4. RoHL: combining frequency biased models

Models trained with different robustness biases are likely to make different errors. We hypothesize that combining models with orthogonal low and high frequency biases should boost performance across the frequency spectrum. We propose RoHL based on this hypothesis and show that it is complementary to diverse data augmentation.

4.1. Data augmentation targeted for high and low frequencies

To cover high-frequency corruptions, we use Gaussian noise and Gaussian blur as generic transformations for data augmentation when training the high-frequency (HF) expert of the ensemble. For added high-frequency robustness we further suggest a new regularization approach when training this expert; see Sec. 4.3.

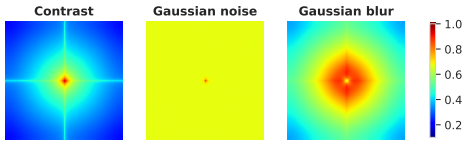


Figure 2: Fourier spectrum of three basic corruptions. Low-frequency components are near the center of the spectrum. Left: Contrast augmentation mostly affects the low-frequency components. Middle, Right: Gaussian noise and blur have relatively larger concentrations in high-frequency regions (away from the center). For visualization details; see supplementary (Sec. 1).

The second member of the ensemble is optimized for low-frequency (LF) corruptions. We do so by using contrast change as a simple generic augmentation operation that has dominant low-frequency components.

The Fourier spectrum of these simple data augmentation operations is visualized in Fig. 2. Both experts are trained by additionally using diverse data augmentation (we test AugMix and DeepAugment). Implementation details are discussed in the experimental section.

4.2. Combination of expert predictions

The derived expert models for HF and LF robustness are combined and tested on object classification and detection. We combine model predictions by simply averaging predictions of the two member models. We also explored more sophisticated learned merging models. The improvement in performance, however, did not justify the increased complexity over simple averaging (Occam's razor). We denote this combination as RoHL (HF, LF).

4.3. TV minimization on feature maps

We improve on the HF expert by introducing a new regularization operation on the early feature maps of the network. In classical image processing, TV minimization has been widely used for various signal restoration problems [2]. TV minimization is particularly useful for removing oscillations in the signal. Unlike conventional low-pass filtering, TV minimization is a nonlinear operation and is formulated as an optimization problem.

TV minimization could directly filter out noise in the test images, but this requires solving an optimization problem for each test image, which makes the approach slow. Moreover, denoising will also destroy important high-frequency signals and may introduce new artefacts which can contribute towards additional performance degradation [15].

We rather propose to use TV minimization at training

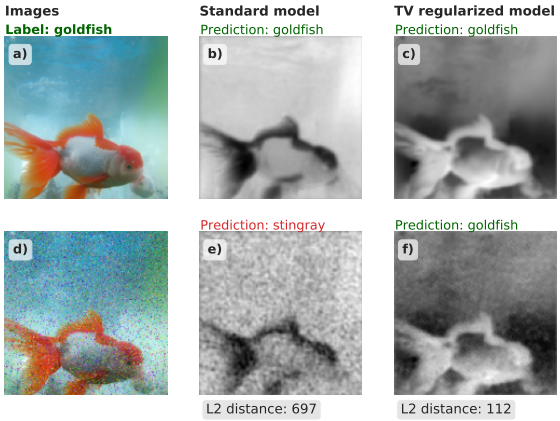


Figure 3: Effect of training with TV regularization. a) and d) show a clean and a noisy test image. We compare feature map visualizations of a standard and a TV regularized model. b) and e) show the most active feature map generated after forwarding of a clean and noisy image, respectively. c) and f) show the same for a TV regularized model. Larger activation values have a lighter shade. We also show the average L2 distance between feature maps from the clean and the noisy *test* images. For a more robust model, the activation statistics should fluctuate less under the influence of noise. The TV regularized model learns to suppress noise that was unseen during training: f) is smoother compared to e) and is closer to c).

time. Instead of applying it to the input images, we apply it to the feature maps of the first conv layer, which processes the input image. As we have discussed, standard CNN models are biased towards using high-frequency information, such as textures. Such a biased model contains filters that fire erratically whenever high-frequency information is present in the input image, resulting in large, noisy activations. This causes downstream layers — which rely on the first convolutional feature maps — to behave in unpredictable ways. We hypothesize that removing spatial outliers (oscillations) in the first conv feature maps will yield more stable representations and, thus, improves robustness to high-frequency corruptions. Since high-frequency signals are picked up best by the first network layer, this is the best placement of the regularizer. We verified this also empirically; see supplementary (Sec. 3). For continuous functions $f : \mathbb{R}^{H\times W} \supset \Omega \rightarrow \mathbb{R}$, the TV norm of f is defined as:

$$\mathcal{L}_{TV}(f) = \int_{\Omega} |\nabla f|.$$

The feature maps $\mathbf{x} \in \mathbb{R}^{H\times W}$ are on a discrete grid. The finite difference approximation reads:

$$\mathcal{L}_{TV}(\mathbf{x}) = \sum_{i,j} |x_{i,j+1} - x_{i,j}| + |x_{i+1,j} - x_{i,j}|.$$

This loss can be combined with the standard cross entropy loss ($\mathcal{L}_{CE}$) for image classification:

$$\mathcal{L}(\bar{\mathbf{y}}, \mathbf{y}, \mathbf{F}) = \mathcal{L}_{CE}(\bar{\mathbf{y}}, \mathbf{y}) + \lambda \sum_{c} \mathcal{L}_{TV}(\mathbf{F}_c)$$

where $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denotes conv feature maps with C channels. $\bar{\mathbf{y}}$ and $\mathbf{y}$ denote the predictions and targets respectively. The factor λ controls the regularization strength (larger values will result in smoother feature maps). The effect of training models with TV regularization is shown in Fig. 3. Models trained with TV regularization yield more consistent feature maps for clean and noisy images. We note that this application of TV regularization is different from standard TV-based image denoising as the reconstruction loss (the data term) is replaced by cross entropy loss.

5. Experiments

5.1. Experimental setup

5.1.1 Datasets

ImageNet & ImageNet-C. The ImageNet dataset consists of approximately 1.2 million images categorized into 1000 classes. To evaluate i.i.d performance we used the standard clean test set. To evaluate performance under distribution shift we used the ImageNet-C dataset [15], a corrupted version of ImageNet's clean test set. ImageNet-C consists of images distorted with 15 different synthetic corruption types (grouped into noise, blur, weather, and digital corruption). Each corrupted subset has 5 severity levels.

ImageNet-100 & ImageNet-C-100. For quicker experimentation, we ran ablations on a smaller subset of the ImageNet dataset consisting of 100 classes. We refer to this dataset as ImageNet-100. The corrupted version of this dataset is denoted as ImageNet-C-100.

Datasets with natural corruptions. To evaluate on natural corruptions we used BDD100k [36] and DAWN [19]. BDD100k consists of driving scenes recorded in varying weather conditions and different times of the day. It is an object detection dataset. We follow [21] to create test splits for different weather conditions: clear, rainy and snowy. DAWN contains a collection of 1000 images taken from road traffic environments with severe weather corruptions. The samples are divided into four weather conditions: fog, rain, snow, and sandstorm. DAWN is used for testing only.

Datasets with other distribution shifts. For non-corruption based shifts we used ImageNet-R [14] and ObjectNet [1]. ImageNet-R contains images of styles, such as abstract or artistic renditions of object classes. ImageNet-R contains 30k image renditions for 200 ImageNet classes. ObjectNet contains 50k images with 313 object classes with 109 classes overlapping with ImageNet. Images contain varying pose and background.

5.1.2 Implementation details

Evaluation. Classification models are usually compared using the error computed on the clean test set (i.i.d). The error metric measures the percentage of misclassification and is computed as: $(100 - \text{Top-1-Accuracy})\%$. Besides the clean error, for corruption datasets, we report the *mean corruption error* (mCE). This involves first computing the unnormalized corruption error (uCE_c) of a given corruption type (c) by averaging across the 5 severity levels. Then, for ImageNet-C-100, we average uCE_c for all 15 corruption types to compute mCE. For ImageNet-C, we follow the convention [15] of normalizing (uCE_c) with AlexNet's corruption error, before averaging over all corruption types. To evaluate classification performance on natural corruptions, we report errors on different corruption types and their mean. For object detection performance, we use the COCO Average Precision (AP) metric, which averages over IoUs between 50% and 90%. On corrupted data we also report mean AP over corruption types and denote it as mAPc.

Architectures. Our experiments use ResNet50. For ablation experiments on ImageNet-100, we moved to the smaller ResNet18 architecture. The object detection experiments use FasterRCNN [25] with ResNet50 as backbone.

Training. We employ AugMix data augmentation together with the JSD consistency loss and the default hyperparameters [17]. For DeepAugment, we use augmented images pre-computed by Hendrycks *et al*. [14]. To train with TV regularization, we use a regularization factor $\lambda = 1e^{-5}$ for all experiments. The cross-entropy loss is very small compared to the total variation of feature maps, thus, a small λ is needed to balance the two losses (a sensitivity analysis for λ is included in the supplementary, Sec. 3). We finetune models to induce HF and LF robustness biases with data augmentation operations. For object detection with FasterRCNN, we used MMDetection framework's implementation [5]. For more detailed training settings see supplementary (Sec. 2).

5.2. Effect of training with TV regularization

We considered the following settings: **a)** standard baseline model trained on natural images, **b)** trained with AugMix data augmentation (denoted as AM), **c)** trained with AugMix data augmentation and TV regularization (denoted as AM_{TV}). Fig. 4 shows that **the TV regularized model consistently improves over the standard and the AugMix model on all corruptions that affect high frequencies**. On low-frequency corruptions (Eg: brightness, contrast, fog), TV regularization has a negative effect. Moreover, Tab. 1 shows that it increases the clean error. This shows that TV regularization induces a *high-frequency robustness* bias, which can be exploited by the proposed high-frequency expert from Sec. 4.2.

Table 1: Classification error of the TV regularized model compared to regular training and training with AugMix (ImageNet-100). Standard: baseline model trained on natural images. TV regularization considerably improves on the corrupted test set, but increases the error on clean images.

Model	Clean err.	mCE
Standard	12.2	49.9
AM	**11.8**	40.9
AM_{TV}	14.8	**35.9**

We also investigated layer-wise application of TV regularization and its impact on the high-frequency robustness. Applying TV regularization on early conv feature maps is crucial for achieving strong high-frequency robustness. Also we evaluated applicability to architectures that do not belong to the ResNet family, namely, DenseNet and MNasNet. Performance gains were similar to ResNet18 with no hyperparameter changes. We also experimented with the generic L_p norm formulation for TV and tried different values for p. We found that the value of p does not significantly impact performance. These additional results are included in the supplement (Sec. 3).

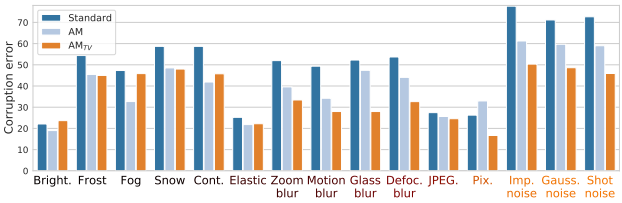
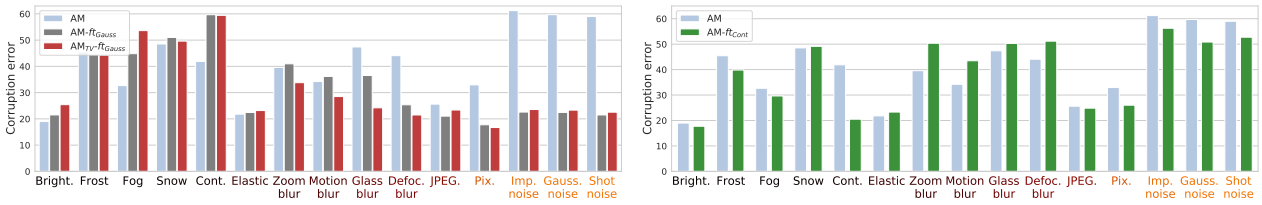


Figure 4: Classification error of an AM_{TV} model on different corruption types (ImageNet-C-100). Y-axis: mean error for a given corruption type over all severities. X-axis: corruption types ordered from low to high frequency (indicated by the colour gradient). Ordering is based on the amount of high-frequency content in corruption types; see supplementary (Sec. 1). Standard denotes a baseline model trained on natural images. Models trained with AugMix are generally more robust, and TV regularization complements this with consistently better performance on all high-frequency corruptions, making it an excellent high-frequency expert.

5.3. Inducing targeted robustness biases

5.3.1 High frequency robustness

We have seen previously that TV regularization reduces error on high-frequency corruptions at the cost of a higher error on clean images and low-frequency corruptions. In particular, we observed improved robustness for noise and



(a) High frequency robustness bias

(b) Low frequency robustness bias

Figure 5: Robustness bias and its impact on performance across corruption types. Figures 5a & 5b show corruption errors for models exhibiting high and low frequency robustness biases, respectively. Y-axis: corruption error for different corruption types (averaged over severity levels). X-axis: corruption types ordered from low to high-frequency. In Fig. 5a, both AM$_{\text{TV}}$ and AM$_{\text{TV}}$-ft_{Gauss} are robust to high frequency corruptions. AM$_{\text{TV}}$-ft_{Gauss} shows larger improvements on blur corruptions. Fig. 5b shows that AM-ft_{Cont} improves on low-frequency corruption types. Surprisingly, it also improves performance on some noise corruptions. Comparing figures 5a & 5b, we see that these models have very different biases.

Table 2: Robustness bias due to data augmentation (results on ImageNet-100). Finetuning with Gaussian noise and Gaussian blur induces a high-frequency robustness bias, whereas using contrast augmentation induces a low-frequency robustness bias.

Model	Rob. bias	Clean err.	mCE
AM	-	**11.8**	40.9
AM-ft_{Cont}	LF	**11.8**	39.1
AM-ft_{Gauss}	HF	13.2	32.5
AM$_{\text{TV}}$	HF	14.8	35.9
AM$_{\text{TV}}$-ft_{Gauss}	HF	16.0	**31.5**

blur corruptions. We tested to what degree this effect can be achieved by finetuning the AugMix models with Gaussian noise and Gaussian blur augmentation applied to the images. We used additive Gaussian noise sampled from $\mathcal{N}(0, 0.08)$. For Gaussian blur, we used a kernel size of 3. We finetuned both AM and AM$_{\text{TV}}$ models with these HF augmentation operations. We denote these models as AM-ft_{Gauss} and AM$_{\text{TV}}$-ft_{Gauss}.

Tab. 2 shows that TV regularization combined with HF augmentation operations obtains the best mCE. Although the gap compared to AM-ft_{Gauss} seems small, these gains are more pronounced for blur corruptions (see Fig. 5a). Thus, **TV regularization has a complementary effect to Gaussian noise and blur augmentation**. As we add more high-frequency robustness bias, performance on clean images and low-frequency corruptions deteriorated.

5.3.2 Low frequency robustness

To induce robustness on low-frequency distortions, we finetune with contrast augmentation, which is a simple generic transformation that mainly affects the low-frequency components (see Fig. 2).

Yin *et al*. [35] evaluated a data augmentation scheme by explicitly adding noise to Fourier components with magnitudes sampled from the *fog* corruption, and found that such an approach degrades performance on low-frequency corruption types (even on *fog*) — suggesting that a clear trade-off does not exist. On the contrary, we observe that **finetuning models with a low-frequency perturbation such as contrast augmentation *does* improve performance on other low-frequency corruptions (fog, frost, brightness). Also it does not degrade the clean error**, as shown in Tab. 2. Fig. 5b shows that it also improves performance for certain high-frequency corruptions like noise while degrading it on blur. This suggests that trade-offs are more nuanced compared to high-frequency augmentation operations. We also tried a LF data augmentation approach which randomly perturbs patches centered at the 0-frequency component. However, this approach performed worse compared to contrast (see supplementary, Sec 4.).

5.4. Combining frequency biased models

Table 3: Performance comparison to a standard ensemble (ImageNet-100). Model$_1$ and Model$_2$ denote the two members. For a standard ensemble, the two models are independently trained but with similar biases (first two rows). Our results (third and fourth row) show improved performance on corruptions while preserving clean performance.

Model$_1$	Model$_2$	Clean err.	mCE
AM	AM	**10.9**	39.1
AM $_{Gauss,\ Cont}$	AM $_{Gauss,\ Cont}$	11.0	29.0
AM-ft_{Gauss}	AM-ft_{Cont}	11.4	28.4
AM$_{\text{TV}}$-ft_{Gauss}	AM-ft_{Cont}	11.7	**25.9**

Can we improve on corruption without degrading the clean error? Tab. 2 shows that biasing models for high-frequency robustness improves the corruption error but degrades the clean error. AM-ft_{Cont} models retain performance on the clean dataset while improving performance on some

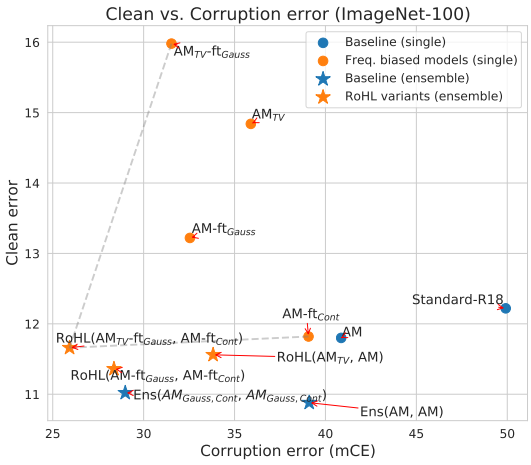


Figure 6: Clean vs corruption error on ImageNet-100. Each point represents a model with a certain corruption error (x-axis) and clean error (y-axis). Points closer to the origin indicate a better trade-off between clean and corruption error. Blue: baselines. Orange: variants of RoHL. Dots: single models. Stars: ensembles of two models.

corruptions, mostly the low-frequency ones. Since these two models have different frequency biases, it is natural to ask — can we improve performance by combining them?

Since ensembles generally have a positive effect on classification accuracy, we set up standard ensemble baselines to compare the proposed expert ensemble. The first baseline consists of two AM models. As we have seen that additional augmentation operations improve mCE, we consider a second ensemble, where each AM model is finetuned with all the used augmentation operations (Gaussian noise, blur, and contrast in addition to the default AugMix operations). We denote members of the second ensemble as AM $_{Gauss, Cont}$. In these baseline ensembles, the member models have the same biases, as they use the same training pipeline.

Tab. 3 shows that the expert combination (AM$_{\text{TV}}$-ft_{Gauss}, AM-ft_{Cont}) provides the best clean and corruption error trade-off. These two models constitute the HF and LF experts for our RoHL approach. It improves the corruption errors by 13.2% points compared to the AM ensemble baseline, while degrading the clean error by only 0.8% points. The trade-off between a low clean error and high robustness to corruptions is best visualized in Fig. 6, where we plot the clean vs corruption error for various models. **Combining models with different biases offers a better trade-off than combining models with the same bias.**

5.5. Scaling to ImageNet

In the previous experiments, we progressively showed training schemes for the HF and LF expert models constituting RoHL. In this section, we verify that the concept carries over to the larger ResNet50 architecture and the full ImageNet dataset. Additionally, we did not just use AugMix for diverse data augmentation, but a combination of AugMix with DeepAugment, a model that was recently suggested by Hendrycks *et al*. [14].

Table 4: Results on ImageNet and ImageNet-C. We compare RoHL to other state-of-the-art approaches using a ResNet50 architecture and an ensemble of two AMDA models with already improves the state-of-the-art. RoHL shows the best trade-off between clean error and mCE.

	Model	Clean err.	mCE
	Standard [13]	23.9	76.7
SOTA approaches	IN-21K-Pretrained [14]	22.4	65.8
	SE (Self-Attention) [14]	22.4	68.2
	CBAM (Self-Attention) [14]	22.4	70.0
	AdversarialTraining [32]	46.2	94.0
	SpeckleNoise [26]	24.2	68.3
	StyleTransfer [12]	25.4	69.3
	AugMix (AM) [17]	22.5	65.3
	DeeAugmet (DA) [14]	23.3	60.4
	AugMix+DeepAugment (AMDA) [14]	24.2	53.6
	Baseline Ensemble (AMDA, AMDA)	24.0	51.9
Ours	RoHL (AM$_{\text{TV}}$, AM)	**22.2**	61.1
	RoHL (AMDA$_{\text{TV}}$, AMDA)	23.6	49.7
	RoHL (AMDA$_{\text{TV}}$-ft_{Gauss}, AMDA-ft_{Cont})	22.7	**47.9**

We first trained a model with TV regularization and AugMix. To train with DeepAugment, we followed Hendrycks *et al*. [14] and finetuned this model with AugMix and DeepAugment (denoted as AMDA$_{\text{TV}}$). The high-frequency expert model (denoted as AMDA$_{\text{TV}}$-ft_{Gauss}) was obtained by finetuning the AMDA$_{\text{TV}}$ model with Gaussian noise and blur augmentation. The low-frequency expert was obtained by finetuning the publicly available AMDA model with contrast augmentation. We denote this model as AMDA-ft_{Cont}. Tab. 4 and Fig. 1 compare our RoHL approach to the state of the art for a ResNet50 model. The standard baseline is a model trained on clean images with random cropping and horizontal flipping. Ensemble (AMDA, AMDA) is a two-member ensemble of the state-of-the-art AMDA model trained with AugMix and DeepAugment. **RoHL improves on both the clean and the corrupted error over the previous state-of-the-art (AMDA) and its ensemble version**.

5.6. Results on real image corruptions

5.6.1 Object classification

BDD100k and DAWN are object detection datasets containing multiple object instances per image and hence cannot be directly used in the classification setting. We extracted object images for each class using 2D bounding box annotations to first transform these datasets to the standard classification setting. The transformed variants are denoted as BDD100k-cls and DAWN-cls.

Table 5: Object classification performance on natural corruptions. We show errors on various weather corruptions in the DAWN-cls test set. DAWN does not have a uncorrupted test set, hence we show results on the "Clear" test split of BDD100k-cls.

Model	Clear error	mCE	Fog	Rain	Sand	Snow
			errors			
Standard data augmentation	5.3	23.5	26.3	16.1	30.3	21.5
AMDA	4.9	16.4	19.4	10.9	21.6	13.6
Ensemble(AMDA, AMDA)	4.9	16.2	19.0	10.8	21.4	13.5
RoHL (AMDA_{TV}-ft_{Gauss},AMDA-ft_{Cont})	**4.7**	**14.5**	**17.7**	**10.6**	**19.0**	**10.6**

We finetuned our ResNet50 models (pre-trained on ImageNet) on the "clear" split of BDD100k-cls. For RoHL, we finetune with the HF and LF biases. We evaluated on corrupted test sets of BDD100k-cls and DAWN-cls.

We observed that weather distortions present in BDD100k are rather benign [19, 21]. Thus the corrupted test sets do not impact performance of models trained even with standard data augmentation ($\sim 2\%$ gap between i.i.d and OOD; see supplementary, Sec. 6). DAWN contains more severe distortions and thus, is more challenging (for examples see supplementary, Sec. 8). Tab. 5 compares performance of RoHL. **Compared to the baselines, RoHL performs better on all real corruptions.**

5.6.2 Object detection

Table 6: Object detection performance with different ResNet50 backbones used in FasterRCNN. We report AP scores on the "Clear" split of BDD100k and corrupted test sets in DAWN. Higher AP scores are better. mAPc denotes the mean AP over corruption types.

Pretrained Backbone	Clear AP	mAPc	Fog	Rain	Sand	Snow
			AP			
Standard data augmentation	31.3	24.9	21.5	25.1	24.8	21.7
AMDA	32.4	27.2	24.9	26.2	27.6	24.8
Ensemble(AMDA, AMDA)	32.4	27.2	**25.4**	**26.2**	27.6	24.2
RoHL (AMDA_{TV}-ft_{Gauss},AMDA-ft_{Cont})	**32.6**	**28.8**	24.9	24.9	**28.1**	**33.4**

To evaluate on object detection, we used the models finetuned on BDD-100k-cls as backbone in the FasterRCNN architecture. To combine predictions for the baseline ensemble and RoHL, we averaged bounding box predictions and class probabilities (both at the RPN and Fast-RCNN stages [25]). For implementation details, see the supplementary (Sec. 2). **Tab. 6 shows that RoHL improves over the baselines also in the scope of object detection.**

5.7. Results on other domain shifts

To measure performance on distribution shifts other than image corruptions, we evaluated RoHL on ImageNet-R and ObjectNet. Similar to the previous sections, we compare to the two-member ensemble of AMDA models. On ImageNet-R, RoHL improves the error by 0.7% points. On ObjectNet, we obtain an improvement of 1.5% points. Gains for these distribution shifts are marginal. This is to be expected, as object pose changes, for example, are high-level modifications not covered by our approach. See supplementary (Sec. 7) for detailed results.

Table 7: Results after adaptating BN statistics. Errors with & without adaption are shown in columns adapt and base.

Model	ImageNet-C mCE		DAWN-cls mCE	
	base	adapt	base	adapt
Standard	76.7	62.2	23.5	16.8
AMDA	53.6	45.4	16.4	13.6
Ensemble(AMDA, AMDA)	51.9	44.7	16.2	13.5
RoHL (AMDA_{TV}-ft_{Gauss},AMDA-ft_{Cont})	**47.9**	**41.2**	**14.5**	**12.4**

5.8. Unsupervised domain adaptation

We evaluated performance of our models after adaptation using Schneider *et al.*'s approach of updating batch-norm statistics at test time [27]. Tab. 7 shows results on ImageNet-C and DAWN-cls. **RoHL's improvements are preserved even after adaptation.**

6. Conclusions

We demonstrated that a mixture of two expert models – one specializing on corruptions in the high-frequency spectrum of the image and one specializing on the low-frequency ones – consistently improves the trade-off between a low error on corrupted samples and a low error on regular clean samples. We also showed that this approach adds to the benefits of a regular ensemble of the same size. Moreover, we introduced TV minimization on the first feature map as a new regularization technique, which consistently improves on high-frequency corruptions and is complementary to other measures in this realm. The principle is flexible with regard to the used base model and dataset size. We showed that the gains transfer to real-world corruptions and also apply to object detection.

Acknowledgements Experiments were mainly run on the Deep Learning Cluster funded by the German Research Foundation (INST 39/1108-1). We also thank Google for donating GCP credits. The research was funded by the German Federal Ministry for Science and Education within the project "DeToL – Deep Topology Learning", and by the German Federal Ministry for Economic Affairs and Energy within the project "KI Delta Learning – Development of methods and tools for the efficient expansion and transformation of existing AI modules of autonomous vehicles to new domains". It was also funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements davenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

[1] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019. 4

[2] Julien Bect, Laure Blanc-Féraud, Gilles Aubert, and Antonin Chambolle. A l1-unified variational framework for image restoration. In Tomás Pajdla and Jiří Matas, editors, *ECCV*, 2004. 2, 3

[3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *ICLR*, 2018. 2

[4] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *AISec Workshop*, 2017. 2

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv*, 2019. 5

[6] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec Workshop*, 2017. 2

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *CVPR*, 2019. 1, 3

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 2

[9] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, 2019. 2

[10] Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise. *ICML*, 2019. 1, 2

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 1, 2

[12] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *NeurIPS*, 2018. 1, 2, 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv*, 2020. 1, 2, 3, 4, 5, 7

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019. 1, 2, 3, 4, 5

[16] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *ICML*, 2019. 2

[17] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *ICLR*, 2020. 1, 2, 3, 5, 7

[18] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *CVPR*, 2020. 2

[19] A. Kenk and M. Hassaballah. Dawn: Vehicle detection in adverse weather nature dataset. *IEEE Trans. Intelligent Transportation Systems*, 2020. 4, 8

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv*, 2017. 2

[21] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *NeurIPS Workshop*, 2019. 2, 4, 8

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2

[23] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*. IEEE, 2016. 2

[24] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019. 1, 2

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5, 8

[26] Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. *ECCV*, 2020. 1, 2, 7

[27] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *NeurIPS*, 2020. 2, 8

[28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 2

[29] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020. 1, 2

[30] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and

Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 2

[31] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv*. 2

[32] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *ICLR*, 2020. 1, 7

[33] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, 2019. 2

[34] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 2

[35] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019. 1, 2, 3, 6

[36] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 4