Neural Image Compression:
Generalization, Robustness, and Spectral Biases

Kelsey Lieberman[1] **James Diffenderfer**[2] **Charles Godfrey**[3] **Bhavya Kailkhura**[2]

Abstract

Recent neural image compression (NIC) advances have produced models which are starting to outperform traditional codecs. While this has led to growing excitement about using NIC in real-world applications, the successful adoption of any machine learning system in the wild requires it to generalize (and be robust) to unseen distribution shifts at deployment. Unfortunately, current research lacks comprehensive datasets and informative tools to evaluate and understand NIC performance in real-world settings. To bridge this crucial gap, we provide a comprehensive benchmark suite to evaluate the out-of-distribution (OOD) performance of image compression methods and propose spectrally inspired inspection tools to gain deeper insight into errors introduced by image compression methods as well as their OOD performance. We then carry out a detailed performance comparison of a classical codec with NIC variants, revealing intriguing findings that challenge our current understanding of NIC.

1. Introduction

Consider the Mars Exploration Rover, whose scientific objective is to search for clues to past activity of water (and perhaps life) on Mars. To achieve this, the rover collects images of interesting rocks and soils to be analyzed by the scientists on Earth. Sending these images down the Earth-bound data stream in their original form is too slow and expensive due to limited bandwidth. Thus, it is well accepted that image compression could play a key role in producing scientific breakthroughs (Malin et al., 2017).

[1]Department of Computer Science, Duke University, Durham NC, USA [2]Lawrence Livermore National Laboratory, Livermore CA, USA [3]Pacific Northwest National Laboratory, Richland, WA USA. Correspondence to: Kelsey Lieberman <kelsey.lieberman@duke.edu>.

Employing image compression in such a setting is challenging for three main reasons: 1) a *high compression ratio* is desired due to low communication bandwidth, 2) given the battery-operated nature of these devices, the compression module has to be *lightweight* so it consumes less memory and power, and 3) *robustness and generalization* to environmental noises and domain shifts, respectively, is desired due to limited Mars-specific training data. These requirements are not specific only to the planetary exploration use case but arise in a wide range of scientific applications using image compression in the wild (Klasky et al., 2021).

Recently, neural image compression (NIC) has demonstrated remarkable performance in terms of rate-distortion and runtime overhead on in-distribution (IND) data (Ballé et al., 2018; Minnen et al., 2018)—satisfying requirements 1) and 2). However, there is limited work on understanding the out-of-distribution (OOD) robustness and generalization performance of image compression methods (requirement 3) (Liu et al., 2022). Our work is driven by several open fundamental empirical and theoretical questions around this crucial issue.

> *How can the expected OOD performance of image compression models be reliably assessed? Can we gain a deeper understanding of the modus operandi of different image compression methods? How do training data properties and biases impact data-driven compression?*

Main Contributions: This paper takes a critical view of the state of image compression and makes several contributions toward answering the aforementioned questions. ❶ First, we design *comprehensive benchmark datasets* for evaluating the OOD performance of image compression methods. Inspired by existing OOD benchmarks for classification and detection (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021; Sun et al., 2022b;a), we design CLIC-C by introducing 15 common shifts emulating train-deployment distribution mismatch to the popular CLIC dataset. ❷ Next, we focus on understanding the image compression performance. The de-facto approach is to use rate-distortion (RD) curves measured with perceptual quality metrics, such as PSNR. Such scalar metrics, although easy to compute, are known to be extremely limited in what they can capture and sometimes can even be misleading (Wang

| | Original | Snow | Glass Blur | Shot Noise |

Low	Snow, Frost, Fog, Brightness, Contrast
Med	Motion blur, Zoom blur, Defocus blur, Glass blur, Elastic transform, Pixelate, JPEG compression
High	Gaussian noise, Shot noise, Impulse noise

(b)

(a)
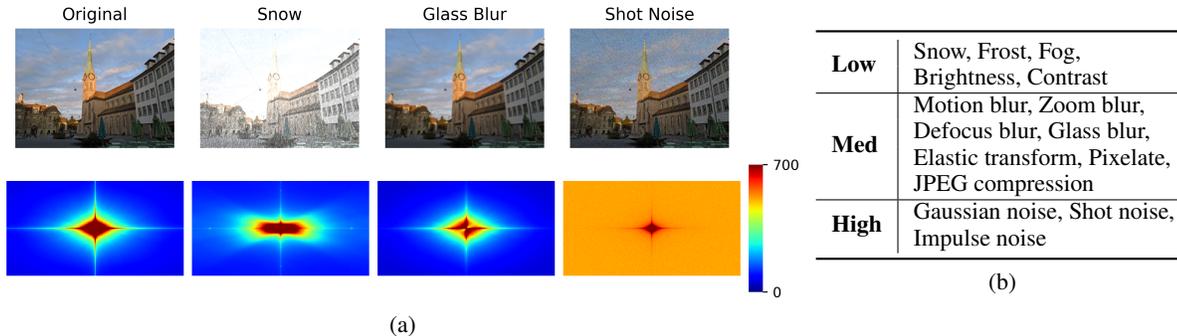
Figure 1: **(a) Top row:** An original CLIC image and the same image with 3 different corruptions in CLIC-C (severity 5). **Bottom left:** Average PSD of CLIC dataset, $\frac{1}{N}\sum_{k=1}^{N} PSD(X_k)$. **Bottom row, other figures:** Average PSD of the difference between the corrupted images and the clean images for each given CLIC-C corruption c, $\frac{1}{N}\sum_{k=1}^{N} PSD(c(X_k) - X_k)$. **(b)** CLIC-C corruptions categorized as low, medium, or high based on corruption average PSD.

et al., 2004; Wang & Bovik, 2009). To complement RD curves, we propose *spectrally-inspired inspection tools* that provide a more nuanced picture of (a) compression error, and (b) OOD performance of a given method. Specifically, we introduce a power spectral density (PSD) based approach to understand the reconstruction error. Our approach not only quantifies how much error was made but also highlights precisely where it was made (in the frequency domain). Similarly, to understand the OOD performance of a compression method in unseen deployment scenarios, we propose *Fourier sensitivity heatmaps*—a visualization tool for highlighting the sensitivity of the reconstruction performance of a compression method to different perturbations in the frequency domain.[1] ❸ Using our benchmark datasets and inspection tools, we carry out *a systematic empirical comparison* of a classical codec with two variants of NIC models.

2. Out-of-distribution image compression datasets

To evaluate NIC in the presence of environmental or digital distribution shifts, we generated variants of the CLIC dataset, which we refer to as CLIC-C. Following the techniques presented in (Hendrycks & Dietterich, 2019) for studying the performance of DNN classifiers encountering distributional shifts "in the wild", our -C datasets consist of images augmented by 15 common corruptions. For each image in the original dataset, the -C dataset contains a corrupted version of the image for each of the 15 common corruptions[2], and for each of five corruption severity levels, with 1 being the lowest severity and 5 being the highest. Figure 1a shows a sample of some corruptions in CLIC-C.

---

[1]Method and results for Fourier sensitivity heatmaps are presented in Appendix A.

[2]We used github.com/bethgelab/imagecorruptions to apply corruptions to CLIC images

While each -C dataset offers a broad sampling of environmental or digital image corruptions, it also provides a spectrally diverse collection of corruptions, in the sense that each corruption can be categorized as low, medium, or high frequency based on the frequency content used for perturbations. We will write $PSD(\cdot)$ to denote the function that converts the input image from the spatial to the frequency domain by computing the power spectral density of the input. Practically, computing $PSD(\cdot)$ is done by applying the fast Fourier transform (FFT) (Brigham & Morrow, 1967), followed by a shift operation to center the zero-frequency component, then taking the absolute value. Now suppose we have a set $\mathcal{X} = \{X_k\}_{k=1}^{N}$ of uncorrupted images and some corruption function $c(\cdot)$ (e.g., frost, gaussian noise, etc.). We analyze the spectrum each corruption $c(\cdot)$ by computing $\frac{1}{N}\sum_{i=1}^{N} PSD(X_i - c(X_i))$ (see Figure 1a). Identifying dominant frequencies in the Fourier spectrum for each corruption yields a rough categorization into low, medium, and high-frequency corruptions, provided in Table 1b.

3. Spectral inspection tools

While existing scalar metrics, such as PSNR, are able to summarize the visual similarity of reconstructed images to the original, we will demonstrate that such metrics can provide an incomplete (and sometimes misleading) picture when measuring the impact of compression in OOD settings. Notably, existing tools do not consider the impact of compression on different frequency ranges of images within a dataset. To more thoroughly analyze the effects of image compression, we propose to measure and visualize the effect of image compression in the spectral domain. Given an image compression model $\mathcal{C}$ that returns reconstructed images, we introduce tools for analyzing compression error in the Fourier domain to better understand (i) which *spectral frequencies* are distorted by $\mathcal{C}$, (ii) the *OOD generalization* error, and (iii) the *robustness* error in the presence
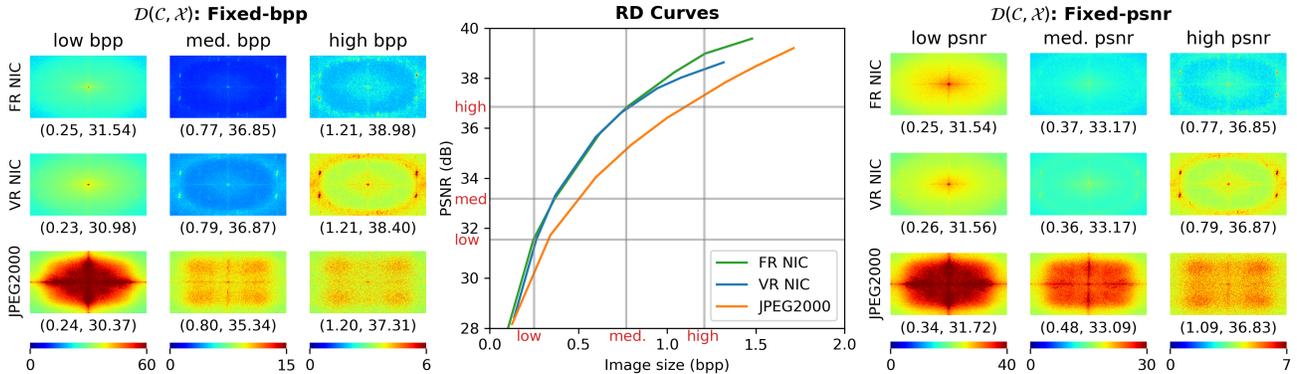
Figure 2: **Visualizing distortion via CLIC test set evaluation. Left:** spectral measure of in-distribution reconstruction error $\mathcal{D}$ under the fixed-bpp constraint at three rates. **Center:** Rate-distortion curves with vertical lines indicating fixed-bpp values and horizontal lines indicating fixed-PSNR values. **Right:** $\mathcal{D}$ under fixed-PSNR constraint. Each $\mathcal{D}$ plot is labeled with a tuple of that model's (bpp, PSNR) on CLIC. Hotter colors (red) indicate more error in that frequency range.

of distributional shifts.

Definition 3.1 (Spectral Measure of Distortion Error). To analyze (i), we evaluate the image compression model $\mathcal{C}$'s ability to reconstruct components of an image across a range of frequencies. To quantify this, we compute the average PSD of the difference between each image X_k in a dataset $\mathcal{X}$ and the reconstructed version $\mathcal{C}(X_k)$ of X_k: $\mathcal{D}(\mathcal{C}, \mathcal{X}) := \frac{1}{N} \sum_{k=1}^{N} PSD(X_k - \mathcal{C}(X_k))$.

Definition 3.2 (Spectral Measure of OOD Generalization Error). For (ii), we evaluate $\mathcal{C}$'s ability to faithfully reconstruct OOD images. To quantify this, we extend the metric $\mathcal{D}(\mathcal{C}, \mathcal{X})$ to account for a corrupted version $c(\mathcal{X})$ of $\mathcal{X}$ as follows: $\mathcal{G}(\mathcal{C}, \mathcal{X}, c) := \frac{1}{N} \sum_{k=1}^{N} PSD(c(X_k) - \mathcal{C}(c(X_k)))$.

Definition 3.3 (Spectral Measure of OOD Robustness Error). For (iii), we evaluate $\mathcal{C}$'s denoising ability. To quantify this, we compute the average PSD of the difference between each uncorrupted image X_k and the reconstructed version $\mathcal{C}(c(X_k))$ of the corresponding corrupted image $c(X_k)$: $\mathcal{R}(\mathcal{C}, \mathcal{X}, c) := \frac{1}{N} \sum_{k=1}^{N} PSD(X_k - \mathcal{C}(c(X_k)))$. [3]

Note that $\mathcal{G}$ provides insight into $\mathcal{C}$'s ability to generalize to a distribution shift c while $\mathcal{R}$ visualizes the denoising effect (or lack thereof) of $\mathcal{C}$ across the frequency domain.

4. Experiments and findings

We analyze the performance of the following image compression methods. ❶ **Classical Codec:** We apply the JPEG2000 algorithm over several compression rates q. ❷ **Neural Image Compressors (NIC):** NIC model optimization uses a hyperparameter λ to control the relative weight of distortion (quality of reconstruction) and rate (level of compression) terms in the objective function. Our experi-

ments include eight **Fixed-Rate (FR)** models, each trained on a single λ value, and one **Variable-Rate (VR)** model, trained over a continuous range of λ values using loss conditional training (Dosovitskiy & Djolonga, 2020). For both FR and VR NICs, the models were optimized for PSNR and the base architectures were the scale hyperprior model of (Ballé et al., 2018). All NIC models were optimized on the train split of the 2020 CLIC dataset (Toderici et al., 2020).

Evaluation setup. We compare results under three constraints: (a) **no constraint**, (b) **fixed-bpp**, and (c) **fixed-PSNR**. In (a), we compare methods over their full range of rate-distortion tradeoffs by generating rate-distortion curves. In (b), we compare models with hyper-parameters which give a very similar bpp result on a particular dataset. For example, we find that on the CLIC dataset, FR NIC with $\lambda = 0.15$, VR NIC with $\lambda = 0.21$, and JPEG2000 with $q = 10$, all give a bpp very close to 1.21. Thus, comparing these three models with those hyper-parameters on CLIC under a fixed bpp constraint, *emulates a setting in which a fixed budget is available to store images.* Analogously, in (c) we compare models with hyper-parameters yielding a fixed PSNR. This *emulates a setting with a requirement on minimum allowable reconstructed image quality.* Scenarios (b) and (c) are used when evaluating $\mathcal{D}, \mathcal{G}, \mathcal{R}$ and Fourier heatmaps (Appendix A).

Test data. All models are tested on (a) in-distribution (IND) and (b) corrupted (or OOD) datasets. We use CLIC and CLIC-C for (a) and (b) respectively.

4.1. Evaluating spectral distortion on IND data

On IND data, the existing RD curve metrics in the center of Figure 2 verify the established trend that NIC models outperform the JPEG2000 model over the compression rates that the NIC model is trained on (bpp $\in (0.1, 1.5)$) (Ballé

---

[3]For simplicity, when $(\mathcal{C}, \mathcal{X}, c)$ is clear from the context, we will just write $\mathcal{D}, \mathcal{G}$, or $\mathcal{R}$.
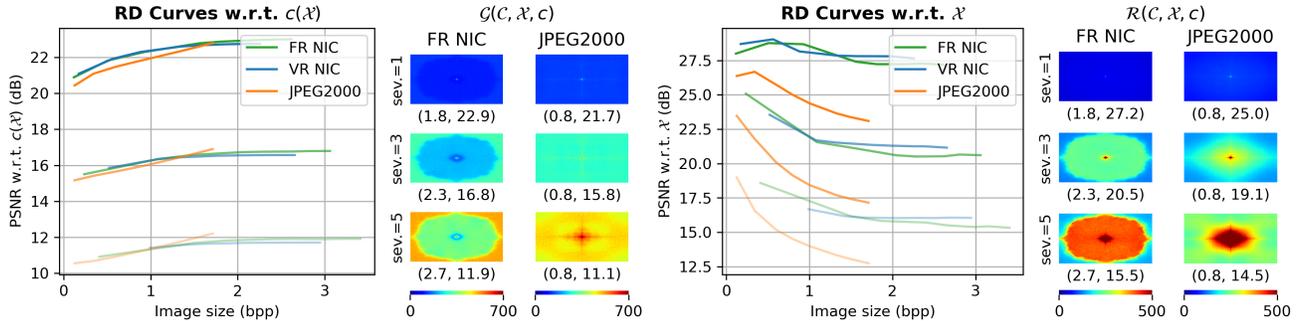
Figure 3: **Generalization errors and denoising errors for a high-frequency shift (shot noise) on CLIC-C. Left:** Generalization. RD curves of of $\mathcal{C}(c(\mathcal{X}))$ w.r.t. $c(\mathcal{X})$ (*i.e.*, PSNR of the reconstructed shifted images w.r.t. the original shifted images) and spectral plots $\mathcal{G}$. **Right:** Denoising. RD curves of $\mathcal{C}(c(\mathcal{X}))$ w.r.t. $\mathcal{X}$ (*i.e.*, PSNR of the reconstructed shifted images w.r.t. the original clean images) and spectral plots $\mathcal{R}$. RD curve plots have three curves for each model: severity=1 (least transparent), severity=3, and severity=5 (most transparent); $\mathcal{G}$ and $\mathcal{R}$ are also tested at severities 1, 3, 5. The plots of $\mathcal{G}$ and $\mathcal{R}$ are labeled with tuples of the model's (bpp, PSNR w.r.t. $c(\mathcal{X})$) and (bpp, PSNR w.r.t. $\mathcal{X}$) respectively.

et al., 2018). VR NIC obtains the same performance as FR NIC for low and moderate bpps; however, FR NIC outperforms the VR NIC at high bpps, despite the fact that both models were trained on the same range of λ. This result follows (Dosovitskiy & Djolonga, 2020) and suggests that the VR NIC may not be expressive enough to learn the high PSNR regime.

Next, we use our spectral inspection tool $\mathcal{D}$ to better understand the effects of different compression methods. Specifically, Figure 2 shows plots of $\mathcal{D}$ under three fixed-bpp and three fixed-PSNR scenarios on the CLIC dataset (*e.g.*, models on the "high psnr" column all have hyper-parameters which result in a PSNR $\approx$ 36.85 on the CLIC test set). Despite having comparable PSNRs, the plots of $\mathcal{D}$ vary greatly between the NIC models and JPEG2000. In particular, NIC models distort high frequencies significantly more than medium frequencies (notice the warmer-colored rings around the edges of the $\mathcal{D}$ plots with cooler-colored centers) while JPEG2000 distorts low and medium frequencies more than high frequencies (notice the large rectangles of warmer colors). This suggests that *NIC models produce inherently different spectral artifacts than classical codecs.*

4.2. Evaluating generalization and robustness of image compression on OOD data

Using our CLIC-C dataset and spectral tools, we study the OOD performance of different compression methods and show results for one example corruption in Figure 3.[4]

The generalization metrics (left side of Figure 3) show that *all the compression methods fail to generalize to high-frequency shifts.* Specifically, on this corruption, all three models have RD curves with low PSNRs w.r.t. $c(\mathcal{X})$ (PSNRs range from 10-24 vs. 28-40 for IND data in Fig-

ure 2) and leave much larger errors in $\mathcal{G}$ (notice the color bar ranges from 0-700 in Figure 3 vs. 0-60 in Figure 2). From the RD curves, we see that the main factor determining the PSNR is the severity of the corruption and not the model type or bpp. *These results might lead us to expect that these models make similar reconstruction mistakes, but our spectral inspection tools indicate that this is not the case at all.* Our plots of $\mathcal{G}$ indicate that NIC models distort the higher frequencies significantly more than the low and medium frequencies while JPEG2000 distorts medium and low frequencies more than high frequencies. This nuance, which is equivalent to one we previously observed with IND data in Figure 2, becomes more apparent from the plots of $\mathcal{G}$ on a high-frequency corruption because high-frequency signals are much more prevalent on this data than on the IND data.

The robustness (*i.e.*, denoising), metrics (right side of Figure 3) reveal more differences between the models. In particular, they show that *NIC models are better at denoising high-frequency corruptions than JPEG2000.* The evidence for this is (a) at each severity, NIC models achieve significantly better PSNR w.r.t. $\mathcal{X}$ than JPEG2000 and (b) the edges of $\mathcal{R}$ plots (representing the high-frequency signals) are comprised of cooler colors (less error) with FR NIC than JPEG2000. The implication of this finding extends into adversarial example denoising, where JPEG and JPEG2000 were previously used (Aydemir et al., 2018).

5. Conclusion

We proposed benchmark datasets and spectral inspection tools to gain a deeper understanding of the robustness and the generalization behavior of image compression models. Using our tools, we revealed similarities and differences of compression methods via a systematic OOD evaluation.

---

[4]Results for additional corruptions are in Appendix B

References

Aydemir, A. E., Temizel, A., and Temizel, T. T. The effects of jpeg and jpeg2000 compression on attacks using adversarial examples. *arXiv preprint arXiv:1803.10418*, 2018.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkcQFMZRb.

Brigham, E. O. and Morrow, R. E. The fast fourier transform. *IEEE Spectrum*, 4(12):63–70, 1967. doi: 10.1109/MSPEC.1967.5217220.

Dosovitskiy, A. and Djolonga, J. You only train once: Loss-conditional training of deep networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HyxY6JHKwr.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, 2021. doi: 10.1109/ICCV48922.2021.00823.

Klasky, S., Thayer, J., and Najm, H. Data reduction for science: Brochure from the advanced scientific computing research workshop. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States); SLAC National Lab., 2021.

Liu, K., Wu, D., Wang, Y., Feng, D., Tan, B., and Garg, S. Malice: Manipulation attacks on learned image compression, 2022.

Malin, M. C., Ravine, M. A., Caplinger, M. A., Tony Ghaemi, F., Schaffner, J. A., Maki, J. N., Bell III, J. F., Cameron, J. F., Dietrich, W. E., Edgett, K. S., et al. The mars science laboratory (msl) mast cameras and descent imager: Investigation and instrument descriptions. *Earth and Space Science*, 4(8):506–539, 2017.

Minnen, D., Ballé, J., and Toderici, G. D. Joint autoregressive and hierarchical priors for learned image compression. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/53edebc543333dfbf7c5933af792c9c4-Paper.pdf.

Sun, J., Mehra, A., Kailkhura, B., Chen, P.-Y., Hendrycks, D., Hamm, J., and Mao, Z. M. A spectral view of randomized smoothing under common corruptions: Benchmarking and improving certified robustness. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pp. 654–671. Springer, 2022a.

Sun, J., Zhang, Q., Kailkhura, B., Yu, Z., Xiao, C., and Mao, Z. M. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*, 2022b.

Toderici, G., Shi, W., Timofte, R., Theis, L., Balle, J., Agustsson, E., Johnston, N., and Mentzer, F. Workshop and challenge on learned image compression (clic2020), 2020. URL http://www.compression.cc.

Wang, Z. and Bovik, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. *A Fourier Perspective on Model Robustness in Computer Vision*. Curran Associates Inc., Red Hook, NY, USA, 2019.

A. Fourier Sensitivity Heatmap Tool

While our spectral tools allow us to measure different capabilities of compression models, they require the availability of OOD, or corrupted, data. To support a setting in which such OOD data is unavailable, we propose adopting another tool: the Fourier sensitivity heatmap. This tool evaluates the PSNR of a compression model on data perturbed with Fourier basis elements (Yin et al., 2019). The resulting visualization is a heatmap where the value at coordinate (i, j) is the PSNR of the compression method on perturbed data $\{X_k + r_k \varepsilon U_{i,j}\}_{k=1}^{N}$, where each r_k is selected uniformly at random from $\{-1, 1\}$, ε is the norm of the perturbation, and $U_{i,j}$ is the (i, j)th Fourier basis matrix. From (Yin et al., 2019), $U_{i,j} \in \mathbb{R}^{n_1 \times n_2}$ and satisfies (i) $\|U_{i,j}\| = 1$ and (ii) $\mathcal{F}(U_{i,j})$ has at most two non-zero coordinates specifically at (i, j) and the coordinate symmetric to (i, j) about the matrix center.

We consider two versions of Fourier heatmaps: (a) with respect to perturbed data and (b) with respect to original data. To analyze (a) and (b), the PSNR at each coordinate (i, j) is computed with respect to the perturbed dataset, $\{X_k + r_k \varepsilon U_{i,j}\}_{k=1}^{N}$, and the unperturbed dataset, $\{X_k\}_{k=1}^{N}$, respectively. (a) and (b) can be seen as measures of generalization and robustness respectively.

We computed Fourier heatmaps over the CLIC dataset in Figure 4. We selected hyperparameters for these models using a fixed-bpp constraint on the clean data. Note: in these Fourier heatmap plots, warmer colors represent *higher PSNR*, which is in contradiction to the plots of $\mathcal{D}$, $\mathcal{G}$, and $\mathcal{R}$ where warmer colors represented *more error*.

On the left side of Figure 4, we analyze the generalization of the image compression models to various frequency shifts. We find that all image compression models generalize to low- and medium-frequency signals better

Figure 4: Fourier heatmaps of methods under the fixed-bpp setting. **Left:** PSNRs of $\mathcal{C}(c(\mathcal{X}))$ with respect to $c(\mathcal{X})$. **Right:** PSNR of $\mathcal{C}(c(\mathcal{X}))$ with respect to $\mathcal{X}$. Warmer colors (red) indicate higher PSNR.

than high-frequency signals. Specifically, observe how the centers of the plots–corresponding to low-frequency signals–have hotter colors than the edges of the plots. This corroborates our findings from both the RD curves and $\mathcal{G}$ metrics on the CLIC-C dataset.

On the right side of Figure 4, we analyze the robustness, or denoising, capabilities of the models. These plots show that NIC models are better at denoising high-frequency corruptions than low-frequency corruptions (notice the cooler-colored diamonds in the center of the NIC plots). Meanwhile JPEG2000 has very consistent and poor denoising properties across the entire range of corruption frequencies (the plots are blue with a narrow gradient of color compared to the NIC plots).

B. Additional Results - Other CLIC-C Corruptions

Figures 5 and 6 show the results of 12 other corruptions from CLIC-C which were not included in the main body.

Figure 5: RD Curves for corruptions not seen in the main body. Rows 1, 3, and 5 show PSNR w.r.t. $c(\mathcal{X})$. Rows 2, 4, and 6 show PSNR w.r.t. $\mathcal{X}$.
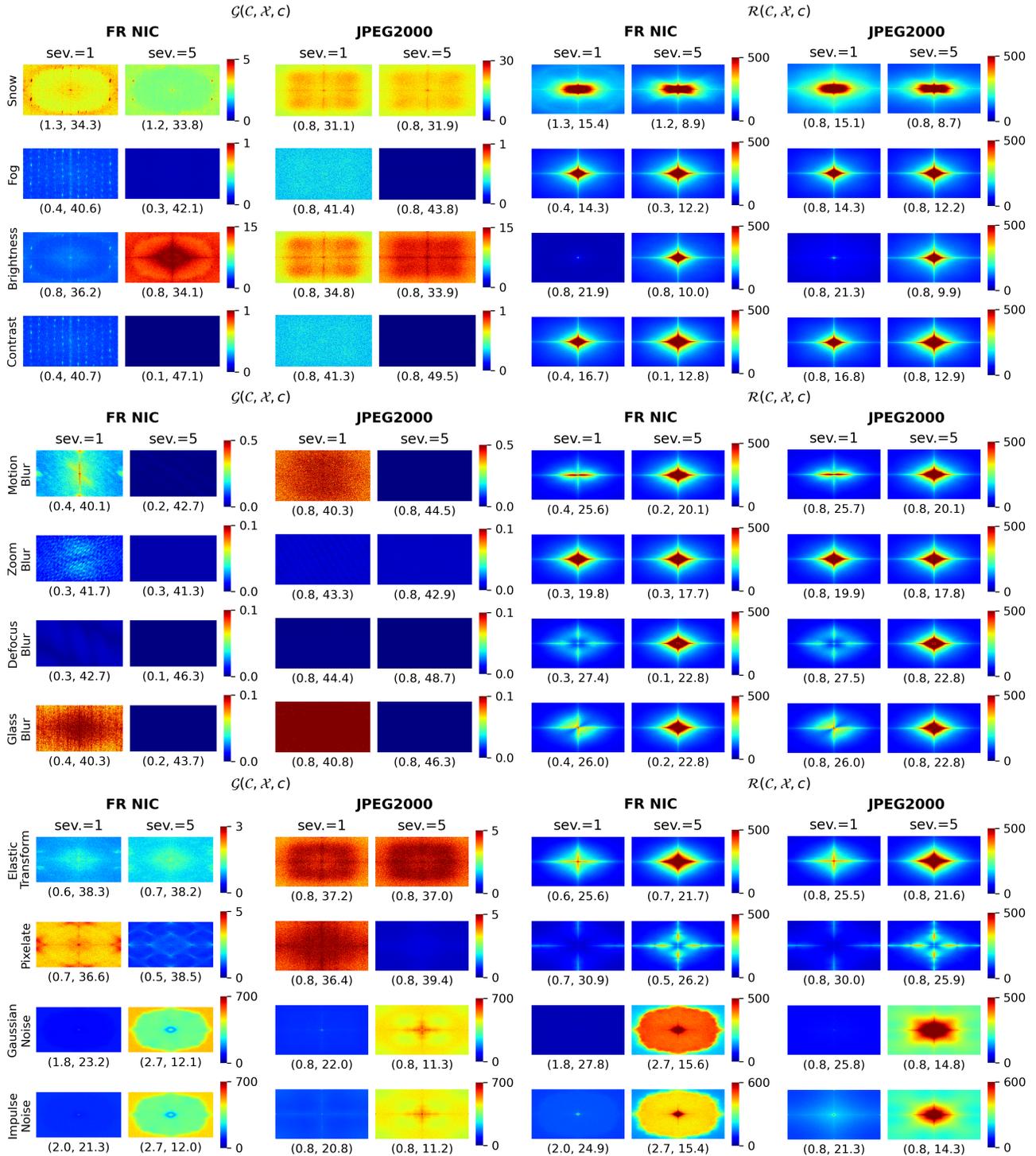
Figure 6: **Generalization error $\mathcal{G}$ and denoising error $\mathcal{R}$ for FR NIC and JPEG2000 on other corruptions of CLIC-C.** We plot both spectral metrics for one low, medium, and high-frequency corruption at severities 1 and 5. The plots of $\mathcal{G}$ and $\mathcal{R}$ are labeled with tuples of the model's (bpp, PSNR w.r.t. $c(\mathcal{X})$) and (bpp, PSNR w.r.t. $\mathcal{X}$) respectively.