Machine Perceptual Quality: Evaluating the Impact of Severe Lossy Compression on Audio and Image Models

Dan Jacobellis*, Daniel Cummings†, and Neeraja J. Yadwadkar*

*University of Texas at Austin
Austin, TX, 78712, USA
danjacobellis@utexas.edu
neeraja@austin.utexas.edu

†Intel Labs, Intel Corporation
Austin, TX, 78746, USA
daniel.cummings@intel.com

Abstract

In the field of neural data compression, the prevailing focus has been on optimizing algorithms for either classical distortion metrics, such as PSNR or SSIM, or human perceptual quality. With increasing amounts of data consumed by machines rather than humans, a new paradigm of machine-oriented compression—which prioritizes the retention of features salient for machine perception over traditional human-centric criteria—has emerged, creating several new challenges to the development, evaluation, and deployment of systems utilizing lossy compression. In particular, it is unclear how different approaches to lossy compression will affect the performance of downstream machine perception tasks. To address this under-explored area, we evaluate various perception models—including image classification, image segmentation, speech recognition, and music source separation—under severe lossy compression. We utilize several popular codecs spanning conventional, neural, and generative compression architectures. Our results indicate three key findings: (1) using generative compression, it is feasible to leverage highly compressed data while incurring a negligible impact on machine perceptual quality; (2) machine perceptual quality correlates strongly with deep similarity metrics, indicating a crucial role of these metrics in the development of machine-oriented codecs; and (3) using lossy compressed datasets, (e.g. ImageNet) for pre-training can lead to counter-intuitive scenarios where lossy compression increases machine perceptual quality rather than degrading it. To encourage engagement on this growing area of research, our code and experiments are available at: https://github.com/danjacobellis/MPQ.

Introduction

In contemporary machine perception pipelines, lossy compression techniques are often employed, but using legacy codecs at near-lossless quality levels, thus limiting potential savings in data rate [1]. For instance, the ImageNet dataset, a cornerstone for image classification tasks, utilizes JPEG compression with an average compression ratio of roughly 5:1. As a result, the full ImageNet-21k is over 1.3 TB in size, and common practice is to discard most of this information using a 224x224 reduced resolution version [2]. Additionally, many types of sensors necessitate extremely high compression ratios, sometimes exceeding 1000:1 [3], resulting from high resolution measurements combined with limited communication bandwidth. As a result, vast amounts of rich, high-fidelity data captured by modern sensors are underutilized or even discarded entirely.

In the decades since the introduction of the ubiquitous JPEG and MPEG standards for images and audio, advancements in lossy compression technologies have
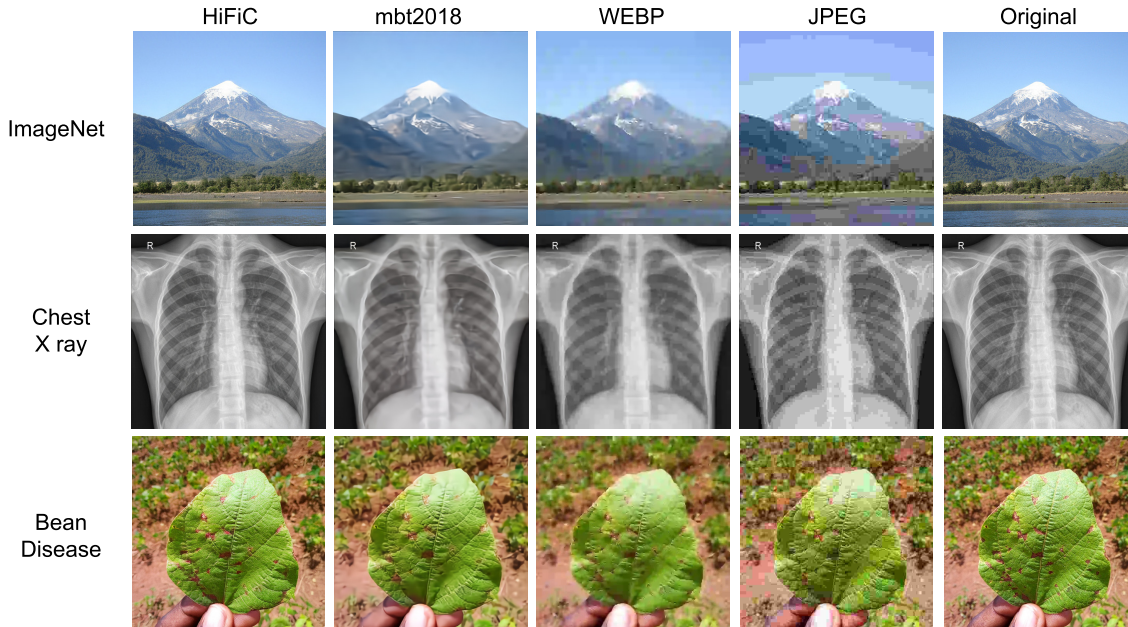
Figure 1: Visual comparison of image compression methods. The original ImageNet image is JPEG compressed at near-lossless quality level of 96 (5.1 BPP), while the Chest X-ray and bean disease original images are lossless.

demonstrated the capability to achieve high compression ratios with minimal degradation in quality. For example, it has been shown that storing the ImageNet-1k dataset using the tokens produced by a ViT-VQGAN neural compression model saves a factor of 100:1 in storage and leads to faster and simplified training [4][5]. While the advantages of employing more potent lossy compression techniques are evident, uncertainty surrounding their impact on downstream machine perception tasks remains a significant barrier. For example, Ilyas et al., [6] demonstrate the existence of signal components, called non-robust features, which are highly predictive yet imperceptible to humans. Lossy compression during training is likely to eliminate these features and thereby lead to sub-optimal models [7]. Additionally, failure to match the exact lossy compression method and settings during training and inference could lead to distribution shift and unpredictable model behavior.

Our work aims to systematically evaluate the impact of various types of lossy compression—both conventional and neural—on both audio and visual machine learning tasks. By understanding these effects, we aim to bridge the gap between the promising capabilities of advanced lossy compression techniques and their practical implementation in machine learning pipelines.

Background

Conventional media compression standards (called codecs) rely on simple but effective linear transforms that exploit the redundancies of natural signals. For example, the discrete cosine transform (DCT) used in JPEG and MP3 compresses signal energy

into fewer coefficients. Carefully designed quantization matrices then assign more bits to perceptually important temporal or spatial sub-bands based on models of human sensitivity. These compression techniques have remained popular for decades since they offer a decent compression rate without excessively compromising signal quality.

Two key developments led to a greater focus on neural network based compression. Ballé et al. [8] showed that autoencoders optimized end-to-end for both rate and distortion (a.k.a. rate distortion autoencoders or RDAEs) compress images more effectively than traditional codecs. In parallel, Van den Oord [9] introduced the vector quantized variational autoencoder (VQ-VAE) as a method of representation learning. Variants of these architectures emerged specializing them for better human perceptual quality, both for audio [10] [11] and images [12]. The advent of generative compression methods [13] led to observation of a rate-distortion-perception trade-off [14].

In addition to better rate-distortion performance, ongoing codec development efforts also aim to optimize for machine perception, a paradigm referred to as "compression for machines [15]" or "machine-oriented compression [16]." Most notably, The JPEG AI standard [17] proposes a single stream image encoder supporting multiple decoders for both human and machine perception. Harell et al., [18] proposed a taxonomy of three different machine-oriented compression approaches. Notable to our work is the method of full-input machine-oriented compression, where the signal is fully decoded before performing downstream tasks; this can either be achieved using an existing codec or by optimizing the compression system for the downstream task.

While the evaluation of human perceptual quality has been extensively studied, it is less clear how different types lossy compression affect machine perception. Hendrycks et al., [19] study the impact of various corruptions, including JPEG compression, on image classification performance. Matsuraba et al., [20] study the impact of various image compression methods on classification and segmentation. Despite these contributions, a dedicated analysis of severe lossy compression effects across a variety of applications, including generative compression methods and audio models, remains unexplored and is the focus of our investigation.

Methodology

We investigate the impact of different audio and image compression techniques on machine perceptual quality under severe lossy compression—which we define as ratios compression ratios between 20:1 and 1000:1—and compare against a baseline that does not have additional compression. We employ six datasets, seven different lossy compression methods, and use popular pre-trained models for various discriminative tasks as summarized in Tables 1 and 2. We use the performance on the validation split of each dataset as a measure of machine perceptual quality. We evaluate the compression performance based on bitrate, conventional distortion metrics, and deep similarity metrics.

Models and datasets. We employ the ImageNet-1k dataset for image classification, using a vision transformer (ViT) pre-trained on ImageNet-21k [21]. The NIH ChestX-ray8 dataset [22] for pneumonia classification and the bean disease dataset

Table 1: Summary of Datasets and Models

Dataset	Task Type	Model	Metric
ImageNet-1k	Image classification	ViT	Top-1 Accuracy
ChestX-ray8	Image classification	ViT	Top-1 Accuracy
Bean Disease	Image classification	ViT	Top-1 Accuracy
ADE20k	Image segmentation	SegFormer	Mean intersection over union
Common Voice	Speech recognition	Whisper	Word recognition accuracy
MUSDB-HQ	Music separation	Demucs v3	Signal-to-distortion ratio

Table 2: Summary of Compression Methods

Method	Description	Setting
JPEG	Legacy transform-coding based image codec	Quality: 5
WEBP	Modern transform-coding based image codec	Quality: 0
MBT2018	Neural image codec with MSE objective	Quality: 1
HiFiC	Neural image codec with adversarial objective	Quality: Low
MP3	Legacy transform-coding based audio codec	Bitrate: 8 kbps
Opus	Modern transform-coding based audio codec	Bitrate: 6 kbps
EnCodec	Neural audio codec with adversarial objective	Bitrate: 6 kbps

[23] are also used in conjunction with an ImageNet-21k pre-trained ViT. Semantic segmentation is performed on the ADE20k dataset [24] using the SegFormer model [25]. The Common Voice 11.0 dataset [26] and the Whisper model [27] are used for speech recognition. Finally, the MUSDB-HQ dataset, an uncompressed version of MUSDB18 [28], and the Demucs v3 model [29] are used for music source separation. These datasets and corresponding models are summarized in Table 1.

Compression methods The image compression methods in our study include JPEG, WEBP, the distortion-optimized neural compression approach from Minnen et al. [30] , and the generative compression method HiFiC [13]. For audio compression, we use MPEG Layer III (MP3), Opus, and the neural audio model EnCodec [11]. Table 2 summarizes these methods. Additional implementation details and a listing of specific model variants are available in our code repository [1].

Evaluation metrics. We use conventional rate-distortion metrics as well as deep similarity metrics—quality metrics derived from deep neural networks and trained to predict human judgments of quality. Each metric is calculated on a per-sample basis.

- *Bits Per Pixel (BPP)* and *Bits Per Sample (BPS)* are used to measure the rate of images and audio signals respectively. For the EnCodec model, which supports a default mode where the VQVAE codes are directly stored and a secondary mode that uses additional entropy coding, we use the default mode without
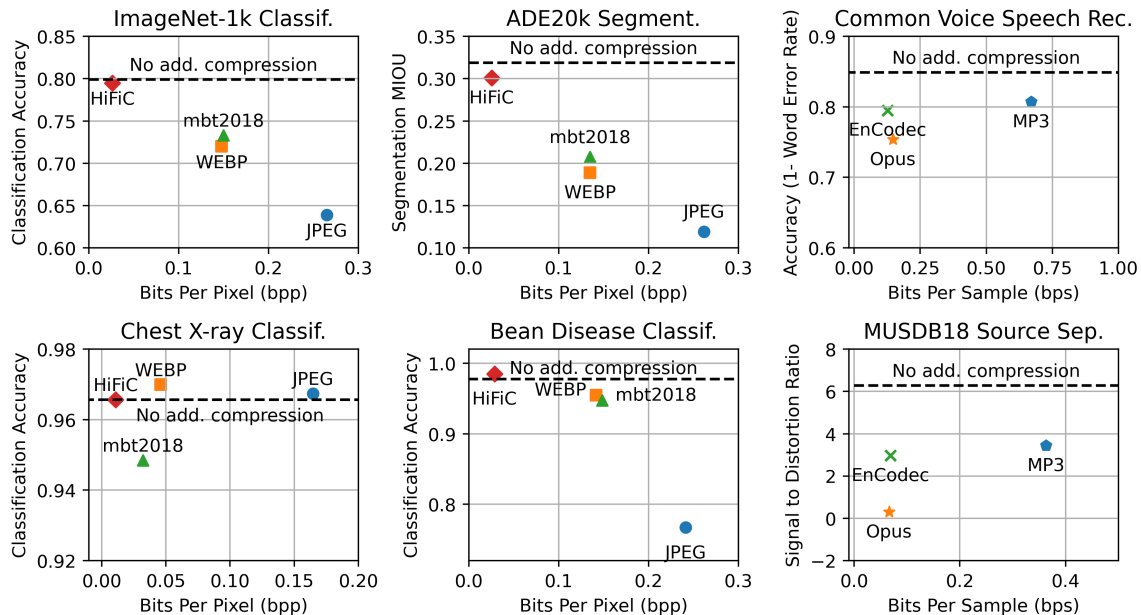
---

[1]Github: danjacobellis/MPQ

Figure 2: Performance on various machine perception tasks when using different types of lossy compression.

entropy coding and calculate BPS using the the product of the codebook size and the number codes. For all other codecs, we directly measure the rate based on the size of the encoded file.

- *Peak Signal-to-Noise Ratio (PSNR)* is used as a conventional distortion metric for both images and audio. We represent image signals using the range $[0, 255]$ and represent audio signals using the range $[-1, 1]$, so $\text{PSNR} = 20 \log_{10}(255) - 10 \log_{10}(\text{MSE})$ for images and $\text{PSNR} = -10 \log_{10}(\text{MSE})$ for audio.

- *Learned Perceptual Image Patch Similarity (LPIPS)*[31] is a deep similarity metric specifically designed for images. It captures complex perceptual differences that simpler metrics like PSNR or SSIM are insensitive to. In the table, we report $-10 \log_{10}(\text{LPIPS similarity})$ to align it with the other quality metrics.

- *Contrastive Deep Perceptual Audio Similarity Metric (CDPAM)* [32] is a deep similarity metric is designed for audio. Like LPIPS for images, it captures perceptual differences more effectively than PSNR. Similar to LPIPS, we report $-10 \log_{10}(\text{CDPAM similarity})$.

Results

Our evaluation across multiple datasets and machine perception tasks reveals key insights into the impact of lossy compression on machine perceptual quality. The results are shown in Figure 2 and are summarized in Tables 3 and 4. For image-based tasks, LPIPS is a better predictor of downstream performance than PSNR,

Table 3: Summary of image results.

Metric	Dataset	Baseline	JPEG	WEBP	MBT2018	HiFiC
PSNR	ImageNet		23.18	24.76	**26.67**	26.25
	ADE20k		23.85	25.59	**28.05**	27.70
	Bean Disease		20.76	21.96	**22.92**	21.82
	Chest X-ray		30.00	32.69	34.70	**36.44**
LPIPS	ImageNet		6.109	7.017	7.945	**10.834**
	ADE20k		7.134	7.959	8.992	**11.81**
	Bean Disease		5.716	6.749	6.899	**9.779**
	Chest X-ray		6.851	7.584	7.799	**13.24**
BPP	ImageNet		0.2647	0.1478	0.1499	**0.0263**
	ADE20k		0.2616	0.1347	0.1347	**0.0254**
	Bean Disease		0.2413	0.1415	0.1484	**0.0286**
	Chest X-ray		0.1646	0.0459	0.0323	**0.0108**
Classification	ImageNet	**0.799**	0.639	0.720	0.733	0.795
Accuracy	Bean Disease	0.9774	0.7669	0.9548	0.9473	**0.9849**
	Chest X-ray	0.9656	0.9673	**0.9699**	0.9484	0.9656
Segment. MIOU	ADE20k	**0.3189**	0.1191	0.1886	0.2075	0.3008

and generative compression (HiFiC) performs the best at all but one of the tasks (Chest X-ray) despite having the lowest average bitrate, and consistently achieves results close to the uncompressed baseline. In the audio domain, similar trends are observed; the audio quality measured by CDPAM is better predictor of downstream performance than PSNR, and, among the methods tested, EnCodec provides the best trade-off between rate and downstream performance for both datasets.

Discussion

Generative compression preserves machine perceptual quality. One area of concern is that generative compression methods like HiFiC and EnCodec, whose adversarial training objectives allow them to discard details at the encoder and re-synthesise them at the decoder, are ill-suited for use within machine perception pipelines. However, our results indicate the contrary; despite having the highest compression ratios among the methods tested, these methods performed well across all tasks, often outperforming methods with significantly higher bitrate. Unfortunately, current generative compression methods are far from being production-ready, and rely on architectures which are difficult to train, adapt, and deploy. However, recent advancements have shown remarkable inference speedup in score-based generative models [33] and vastly simplified training procedures for VQVAEs [34]. By incorporating such advancements and making these methods more accessible, generative compression could enable new applications, such as satellite, maritime and aerial remote sensing systems that require very high compression ratios.

Table 4: Summary of audio results.

Metric	Dataset	Baseline	MP3	OPUS	Encodec
PSNR	MUSDB18		**29.17**	22.17	24.95
	CV		**33.89**	26.70	29.04
CDPAM	MUSDB18		38.43	36.46	**45.33**
	CV		37.89	38.27	**46.34**
BPS	MUSDB18		0.3628	**0.06615**	0.06871
	CV		0.6696	0.1439	**0.1262**
SDR	MUSDB18	**6.286**	3.440	0.2986	2.968
WRA	CV	**0.8488**	0.8072	0.7535	0.7950

Correlation of machine perceptual quality with deep similarity metrics. Deep similarity metrics like LPIPS and CDPAM are known to be highly effective at predicting human perceptual quality as measured by mean opinion score (MOS). Across the six datasets tested, our results indicate that such metrics are also strongly correlated with machine perceptual quality, despite only being trained in a supervised fashion to predict human judgments of signal distortion pairs. A promising avenue for future research would be to extend the training objectives for these metrics to include machine judgments of distortion pairs, making them even more robust.

Pretraining on lossy datasets. Our experiments reveal a surprising phenomenon: for models pre-trained on lossy datasets like ImageNet, additional lossy compression at test time may have negligible impact on performance, and can sometimes behave as an enhancement. For example, the top-1 classification accuracy on the bean disease dataset is higher when compressed using HiFiC (compression ratio of 839:1) than when using the original lossless images. Even more surprising is that severe JPEG compression (see Figure 1) results in an increase in pneumonia classification performance on the Chest X-ray dataset, despite having the lowest quality measured by PSNR or LPIPS. Viewing lossy compression as a type of distribution shift provides one possible explanation for this phenomenon; subtle high-frequency details that only exist in lossless images never occur in pre-training datasets like the JPEG-compressed ImageNet. Exploring pre-training with lossless data may be feasible considering the moderate compression ratios (5:1) used such datasets. The development of lossless datasets at ImageNet or larger scale could be valuable for the development of neural compression systems—for both human and machine applications.

Limitations and Future Directions. By describing the limitations of this work we hope to highlight subtopics for future research. One key limitation of this study is the exclusive use of pre-trained models under the framework of full-input machine-oriented compression. While this approach offers a practical perspective on how existing models may perform using available compression methods, it does not capture

the potential advantages of models that are tailored to compressed data. We do not explore other types of machine-oriented compression, such as model-splitting [18]. Although most of the codecs tested allow different quality settings, we only tested settings on the low end that result in severe loss.

Conclusion

We observe that lossy compression is underutilized in common machine learning pipelines. Our study reveals a surprising and promising outcome: significantly high compression rates can be achieved without excessively compromising machine perceptual quality. Thus, more potent lossy compression can be integrated into learning pipelines, by extending current similarity metrics and optimizing generative compression for production scenarios. This leads to two key advantages, (1) greater accessibility of large-scale pre-training due to reduced storage requirements and (2) better utilization of high-resolution sensor data in bandwidth-restricted systems. Future research should expand the diversity of perception tasks and compression scenarios, and consider the creation of lossless datasets to explore the effect of lossy compression during pre-training in greater depth.

References

[1] Max Ehrlich, "The first principles of deep learning and compression," *arXiv preprint arXiv:2204.01782*, 2022.

[2] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor, "Imagenet-21k pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.

[3] Eric Cocker, Julie A Bert, Francisco Torres, Matthew Shreve, Jamie Kalb, Joseph Lee, Michael Poimboeuf, Paloma Fautley, Samuel Adams, Joanne Lee, et al., "Low-cost, intelligent drifter fleet for large-scale, distributed ocean observation," in *OCEANS 2022, Hampton Roads*. IEEE, 2022, pp. 1–8.

[4] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu, "Vector-quantized image modeling with improved vqgan," *arXiv preprint arXiv:2110.04627*, 2021.

[5] Song Park, Sanghyuk Chun, Byeongho Heo, Wonjae Kim, and Sangdoo Yun, "Seit: Storage-efficient vision training with tokens using 1% of pixel storage," *arXiv preprint arXiv:2303.11114*, 2023.

[6] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[7] Ayse Elvan Aydemir, Alptekin Temizel, and Tugba Taskaya Temizel, "The effects of jpeg and jpeg2000 compression on attacks using adversarial examples," *arXiv preprint arXiv:1803.10418*, 2018.

[8] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[9] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[10] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[11] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[12] Dailan He, Ziming Yang, Hongjiu Yu, Tongda Xu, Jixiang Luo, Yuan Chen, Chenjian Gao, Xinjie Shi, Hongwei Qin, and Yan Wang, "Po-elic: Perception-oriented efficient learned image coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1764–1769.

[13] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11913–11924, 2020.

[14] Aaron B Wagner, "The rate-distortion-perception tradeoff: The role of common randomness," *arXiv preprint arXiv:2202.04147*, 2022.

[15] Lahiru D Chamain, Fabien Racapé, Jean Bégaint, Akshay Pushparaja, and Simon Feltman, "End-to-end optimized image compression for machines, a study," in *2021 Data Compression Conference (DCC)*. IEEE, 2021, pp. 163–172.

[16] Jung-Heum Kang, Muhammad Salman Ali, Hye-Won Jeong, Chang-Kyun Choi, Younhee Kim, Se Yoon Jeong, Sung-Ho Bae, and Hui Yong Kim, "A super-resolution-based feature map compression for machine-oriented video coding," *IEEE Access*, vol. 11, pp. 34198–34209, 2023.

[17] João Ascenso, Elena Alshina, and Touradj Ebrahimi, "The jpeg ai standard: Providing efficient human and machine visual data consumption," *Ieee Multimedia*, vol. 30, no. 1, pp. 100–111, 2023.

[18] Alon Harell, Yalda Foroutan, Nilesh Ahuja, Parual Datta, Bhavya Kanzariya, V Srinivasa Somayaulu, Omesh Tickoo, Anderson de Andrade, and Ivan V Bajic, "Rate-distortion theory in coding for machines and its application," *arXiv preprint arXiv:2305.17295*, 2023.

[19] Dan Hendrycks and Thomas Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[20] Yoshitomo Matsubara, Ruihan Yang, Marco Levorato, and Stephan Mandt, "Sc2 benchmark: Supervised compression for split computing," *Transactions on Machine Learning Research*, 2023.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[23] Vimal Singh, Anuradha Chug, and Amit Prakash Singh, "Classification of beans leaf diseases using fine tuned cnn model," *Procedia Computer Science*, vol. 218, pp. 348–356, 2023.

[24] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[25] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[26] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[27] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[28] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "Musdb18-a corpus for music separation," 2017.

[29] Alexandre Défossez, "Hybrid spectrogram and waveform source separation," *arXiv preprint arXiv:2111.03600*, 2021.

[30] David Minnen, Johannes Ballé, and George D Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[32] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 196–200.

[33] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever, "Consistency models," *arXiv preprint arXiv:2303.01469*, 2023.

[34] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen, "Finite scalar quantization: Vq-vae made simple," *arXiv preprint arXiv:2309.15505*, 2023.