Reconstruction Distortion of Learned Image Compression with Imperceptible Perturbations

Yang Sui [1]♥ Zhuohang Li [2]♡ Ding Ding [3] Xiang Pan [3] Xiaozhong Xu [3] Shan Liu [3] Zhenzhong Chen [4]♣

Abstract

Learned Image Compression (LIC) has recently become the trending technique for image transmission due to its notable performance. Despite its popularity, the robustness of LIC with respect to the quality of image reconstruction remains under-explored. In this paper, we introduce an imperceptible adversarial attack approach designed to effectively degrade the reconstruction quality of LIC, resulting in the reconstructed image being severely disrupted by noise where identifying any object in the reconstructed image is virtually impossible. More specifically, we generate adversarial examples by introducing a Frobenius norm-based loss function to maximize the discrepancy between original images and reconstructed images from adversarial examples. Further, leveraging the human vision's insensitivity to high-frequency components, we introduce Imperceptibility Constraint (IC) to ensure that the perturbations remain inconspicuous. Experiment results on the Kodak dataset with various LIC models demonstrate the effectiveness of our method. In addition, we provide several findings and suggestions for designing future defenses.

1. Introduction

Learned Image Compression (LIC) has recently achieved tremendous success in transmitting images under bit-rate limits due to its superior performance over traditional image compression. Specifically, LIC frameworks (Ballé et al., 2017; 2018; Minnen et al., 2018; Cheng et al., 2020) exhibit significant rate-distortion (R-D) performance and outperform standard image compression methods such as JPEG (Wallace, 1991), JPEG2000 (Taubman & Marcellin, 2002), and BPG (Bellard, 2016). Generally, the fundamental LIC structure leverages the auto-encoder framework, which adopts a Deep Neural Network (DNN)-based non-linear transformation with an encoder for image compression and a decoder for reconstruction.

Despite demonstrating a high recovery ability, the DNN-based encoder and decoder also bring concern about robustness. Adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015) are a type of security threat against DNN-based systems where the attacker injects specially crafted perturbations to images to construct adversarial examples, which are natural-looking images that can cause misclassification to DNN models. Originally discovered in image classification tasks (Szegedy et al., 2014; Goodfellow et al., 2015), adversarial examples have then attracted great research attention in many fields of computer vision, including object detection and semantic segmentation (Xie et al., 2017), facial recognition (Dong et al., 2019), and visual question answering (Li et al., 2021).

While the robustness of downstream tasks has been extensively investigated (Carlini & Wagner, 2017; Madry et al., 2018), the robustness of LIC, which is evaluated in terms of image reconstruction quality instead of classification accuracy, has received little attention (Tan et al., 2023). As depicted in Fig. 1, a typical LIC system can accurately reconstruct an original (unperturbed) image. However, the LIC is considered to be not robust if an attacker can introduce small perturbations into the original image to significantly disrupt the reconstructed image, resulting in the main object in the reconstructed image being unrecognizable.

In this paper, we propose to investigate the robustness of LIC by attacking the image reconstruction process. The main idea is to solve an optimization problem to minimize the adversarial perturbation while maximizing the ***Frobenius norm-based loss*** metric between the original and reconstructed images. However, the adversarial perturbations generated with unconstrained Frobenius norm-based loss are likely to be sensitive to human eyes. To improve the imperceptibility of the generated adversarial images, we

♥The work of Y. Sui was done during the internship at Tencent America.♡The work of Z. Li was done during the visit at Tencent America.♣The work of Z. Chen was done during the visit at Tencent. [1]Rutgers University [2]Vanderbilt University [3]Tencent America [4]Wuhan University. Correspondence to: Yang Sui <yang.sui@rutgers.edu>, Zhenzhong Chen <zzchen@whu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

draw insights from the observation that high-frequency components are less perceptible to human vision (Sharma et al., 2019), and consider generating perturbations from a frequency perspective by introducing a Discrete Cosine Transform (DCT)-based ***Imperceptibility Constraint (IC)*** into the adversarial loss function, rendering the perturbations more unnoticeable by human perception. Our contributions can be summarized as follows: ❶ We conduct a systematic investigation on the robustness of LIC by launching a series of attacks that disrupts the image reconstruction process by introducing a Frobenius norm-based loss with IC. ❷ Our experiments demonstrate that our proposed attack can disrupt LIC while maintaining the imperceptibility of the induced perturbations. ❸ Based on our experiments, we provide several intriguing findings and potential insights regarding designing robust LICs.

2. Preliminary

2.1. Learned Image Compression

Given the non-linear encoder $g_a(\cdot)$ and decoder $g_s(\cdot)$, let $\boldsymbol{\mathcal{X}}$ and $\hat{\boldsymbol{\mathcal{X}}}$ denote the original input and reconstructed images, and $\boldsymbol{\mathcal{Y}}$ and $\hat{\boldsymbol{\mathcal{Y}}}$ denote the pre-quantized and quantized latent representation, respectively. The image compression process is formulated as follows:

$$\boldsymbol{\mathcal{Y}} = g_a(\boldsymbol{\mathcal{X}}), \;\; \hat{\boldsymbol{\mathcal{Y}}} = \mathrm{AD}(\mathrm{AE}(Q(\boldsymbol{\mathcal{Y}}))), \;\; \hat{\boldsymbol{\mathcal{X}}} = g_s(\hat{\boldsymbol{\mathcal{Y}}}), \tag{1}$$

where $Q(\cdot)$ is the quantization operation, and AE and AD represent the arithmetic encoding and decoding processes, respectively. The reconstructed image $\hat{\boldsymbol{\mathcal{X}}}$ is the output of the corresponding (inverse) transform. In addition, a hyperprior is used as side information to reduce the bit-rate.

2.2. Adversarial Attack

Given a natural image $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{H\times W\times C}$, the corresponding label k, and a classification model $f(\cdot)_i$ which predicts the probability of the image belonging to class i, the goal of adversarial attacks is to craft an adversarial perturbation $\boldsymbol{\delta} \in \mathbb{R}^{H\times W\times C}$ to be added onto $\boldsymbol{\mathcal{X}}$ so that it is misclassified by $f(\cdot)$, which can be formulated as:

$$\arg\max_{i} f(\boldsymbol{\mathcal{X}}+\boldsymbol{\delta})_i \neq k, \quad \|\boldsymbol{\delta}\|_p \le \epsilon, \tag{2}$$

where ϵ controls the perturbation budget.

3. Distortion with Imperceptible Perturbation

3.1. Reconstruction Distortion Attack

Unlike the traditional adversarial attack, which aims to mislead the classification model into predicting a wrong label, the adversarial attack on LIC aims to introduce small noise to the original image so that the reconstructed image is severely corrupted. This objective can be formulated as:

$$\arg\max_{\boldsymbol{\delta}} \;\; \mathtt{dis}(\boldsymbol{\mathcal{X}}, g_s(Q(g_a(\boldsymbol{\mathcal{X}}+\boldsymbol{\delta})))), \|\boldsymbol{\delta}\|_p \le \epsilon, \tag{3}$$

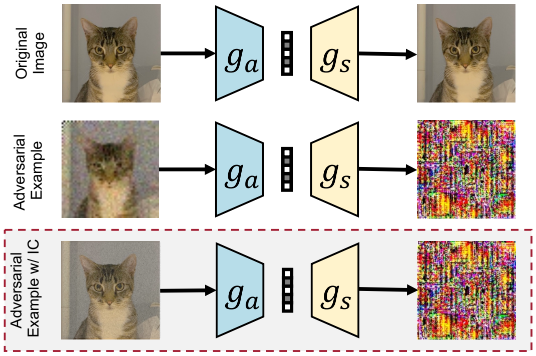

Figure 1. Illustration of proposed adversarial attack with IC on LIC to disrupt the reconstruction image. Top: The standard LIC process. Middle: The proposed adversarial attack against LIC. Bottom: The proposed adversarial attack with IC against LIC while ensuring the perturbation on the original image remains more imperceptible.

where $\mathtt{dis}(\cdot,\cdot)$ denotes a distance function that computes the distance between two tensors. In addition, we utilize differential approximation quantization (Shin & Song, 2017) to achieve gradient-based attacks. We note that, in practice, the underlying technology for image compression is often standardized or industrially recommended (such as JPEG compression) to ensure compatibility across all scenarios. Therefore, we assume the attacker has complete knowledge of the LIC and can launch the attack in a white-box manner.

3.2. Imperceptible Perturbations

As illustrated in Fig. 1 (middle), although the image can be heavily distorted by solving the Eq. 3, in order to reduce the reconstructed image quality, a significant amount of perturbation needs to be injected, which makes the primary subject of the image barely recognizable and the attack easily detectable by human inspectors.

Previous research (Sharma et al., 2019) has demonstrated that high-frequency perturbations are less noticeable. Typically, performing DCT on an image reveals that the low-frequency components carry the major semantic information, while the high-frequency components include the edge structural information. Human beings tend to focus on semantic information, such as the main object in an image, but ignore detailed information, particularly beneath edge structures. Hence, human visual perception is typically not sensitive to these high-frequency perturbations.

Motivated by the frequency perspective, we propose a DCT-based IC to generate imperceptible high-frequency perturbations. In particular, IC encourages the perturbation to mainly modify the high-frequency components of the original images, while constraining the low-frequency components of the adversarial images to remain consistent and close to those in the original images, which can be formulated as:

$$\mathcal{I}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}+\boldsymbol{\delta}) = \|\mathcal{T}(\boldsymbol{\mathcal{X}}) - \mathcal{T}(\boldsymbol{\mathcal{X}}+\boldsymbol{\delta})\|_F, \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the distance.

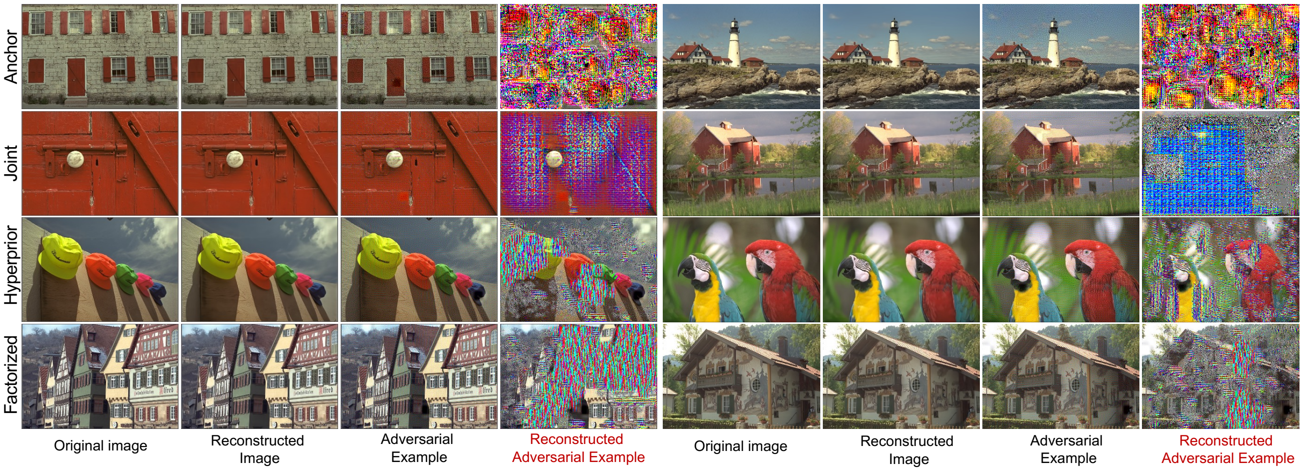

Figure 2. Visual results of our proposed reconstruction quality attack with IC. The attack is evaluated on the Kodak dataset with four various LIC models as victim models.

Model-ϵ	Reconstructed Original Image		Adversarial Example (AE)		Reconstructed AE		Δ of AE vs. Reconstructed AE	
	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	$\downarrow$ PSNR	$\downarrow$ MS-SSIM
Anchor-(32/255)	32.99	0.9880	22.47	0.9112	7.500	0.2188	**14.97**	**0.6923**
Anchor-(64/255)	32.99	0.9880	17.63	0.8281	6.617	0.1816	**11.01**	**0.6464**
Hyperprior-(32/255)	36.48	0.9943	20.30	0.8919	12.05	0.6617	**8.246**	**0.2301**
Hyperprior-(64/255)	36.48	0.9943	15.15	0.8135	7.118	0.3608	**8.031**	**0.4527**
Factorized-(32/255)	33.29	0.9901	19.90	0.9020	7.614	0.4443	**12.28**	**0.4570**
Factorized-(64/255)	33.29	0.9901	14.99	0.8221	5.946	0.2868	**9.050**	**0.5353**
Joint-(32/255)	35.21	0.9911	21.38	0.9205	5.921	0.1633	**15.46**	**0.7572**
Joint-(64/255)	35.21	0.9911	17.52	0.8662	5.733	0.1335	**11.78**	**0.7327**

Table 1. Average PSNR and MS-SSIM of the reconstructed original images, adversarial examples (AE), and reconstructed adversarial examples across 24 images from the Kodak dataset. The bold number denotes the average degradation of PSNR and MS-SSIM between the adversarial examples and their reconstructed counterparts by our proposed attack method.

$\mathcal{T}(\cdot)$ denotes the function truncates the low-frequency band based on DCT, which can be formulated as follows:

$$\mathcal{T}(\boldsymbol{\mathcal{X}}) = \texttt{IDCT}(\boldsymbol{M} \odot \texttt{DCT}(\boldsymbol{\mathcal{X}})), \tag{5}$$

where $\odot$ denotes the Hadamard product (element-wise product). $\boldsymbol{M} \in \mathbb{R}^{H \times W}$ is a binary mask applied to the frequency domain of the tensor $\boldsymbol{\mathcal{X}}$ after DCT to constrain its frequency components. To make the perturbation imperceptible, we mask out half of the components with lower frequencies and only preserve the higher half of the frequency components.

By combining Eq. 3 and Eq. 4, the overall optimization objective is as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mathcal{X}}, \boldsymbol{\delta}) = & -\|\boldsymbol{\mathcal{X}} - g_s(Q(g_a((\boldsymbol{\mathcal{X}} + \boldsymbol{\delta}))))\|_F^2 \\ & + \eta \cdot \mathcal{I}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} + \boldsymbol{\delta}), \quad \text{s.t.} \|\boldsymbol{\delta}\|_\infty \leq \epsilon, \end{aligned} \tag{6}$$

where η controls the influence of IC.

4. Experiments

Setting. For LIC models, we train the `Anchor` (Cheng et al., 2020) , `Hyperprior` (Ballé et al., 2018), `Factorized` (Ballé et al., 2018), and `Joint` (Minnen et al., 2018) LIC models following the CompressAI (Bégaint et al., 2020), to evaluate the distortion. We use the differential approximation quantization from (Shin & Song, 2017) to execute the gradient-based attack. We set $\eta = 10^2$ for solving Eq. 6. All experiments are conducted on the Kodak dataset using a single NVIDIA A100.

Metric. The effectiveness of the proposed attack is evaluated by Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity Index (MS-SSIM). A higher value of PSNR and MS-SSIM indicates better reconstruction quality. "↓ PSNR" and "↓ MS-SSIM" denote the decrease in PSNR and MS-SSIM, measured between adversarial examples and their reconstructed counterparts, where a higher value indicates a more effective attack.

Results for Reconstruction Distortion. We generate adversarial examples on the four LIC models with perturbation budget $\epsilon = \{32/255, 64/255\}$ and visual results are presented in Fig. 2, which includes 8 images from the Kodak dataset. As observed, the LIC successfully reconstructs the original image. However, for the adversarial image, despite the perturbation being almost imperceptible and unnoticeable, the reconstructed adversarial examples are significantly disrupted. Table 1 presents the average PSNR and MS-SSIM of the reconstructed original images, adversarial examples, and reconstructed adversarial examples across 24 Kodak images. As shown in the Table, the average PSNR and MS-SSIM of reconstructed adversarial examples, denoted by the number with the underline, are extremely low,

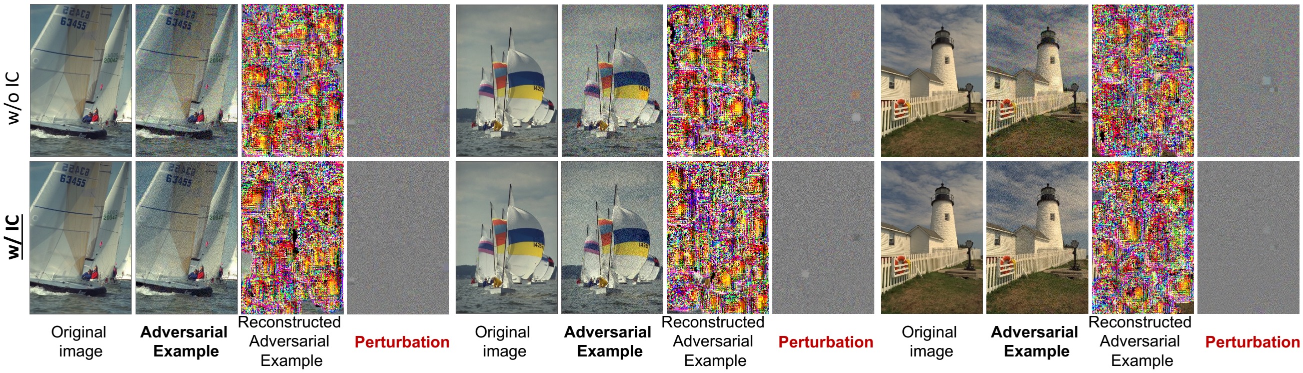

Figure 3. Impact of proposed IC. The visualization depicts perturbations generated from attacks with or without IC. As shown in the figure, the adversarial example generated with IC exhibits a more natural behavior compared to those without IC.

	ϵ	Low	Middle	High	Average
w/o IC	32/255	0.8268	0.8359	0.8374	0.8333
w/ IC	32/255	**0.9315**	**0.9015**	**0.9006**	**0.9112**
w/o IC	64/255	0.6820	0.6781	0.6873	0.6824
w/ IC	64/255	**0.8570**	**0.8147**	**0.8126**	**0.8281**

Table 2. The MS-SSIM of adversarial examples compared to original images on the low, middle, and high quality `Anchor` models with or without high-frequency constraints. A higher MS-SSIM indicates that the adversarial perturbations are more imperceptible.

Reconstruction Quality	Reconstructed Original Image		AE		Reconstructed AE		Δ of AE vs. Reconstructed AE	
	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	↓ PSNR	↓ MS-SSIM
				Anchor				
Low	32.01	0.9838	23.96	0.9315	11.57	0.4006	**12.39**	**0.5309**
Middle	33.16	0.989	21.68	0.9015	5.510	0.1274	**16.17**	**0.7741**
High	33.81	0.9914	21.78	0.9006	5.421	0.1285	**16.35**	**0.7721**
Average	32.99	0.9880	22.47	0.9112	7.500	0.2188	**14.97**	**0.6923**
				Factorized				
Low	29.97	0.9831	20.36	0.9002	7.786	0.4178	**12.57**	**0.4824**
Middle	34.07	0.9929	19.76	0.8945	8.270	0.4779	**11.49**	**0.4166**
High	35.84	0.9944	19.59	0.9115	6.787	0.4374	**12.80**	**0.4741**
Average	33.29	0.9901	19.90	0.9020	7.614	0.4443	**12.28**	**0.4577**

Table 3. Results on LIC models with different quality levels.

which verifies the effectiveness of our proposed attack. We also evaluate the average degradation of PSNR and MS-SSIM measured between the adversarial examples and their reconstructed counterparts, marked as bold. Notably, the attack on the `Joint` model with $\epsilon = 32/255$ can achieve an MS-SSIM degradation of 0.7572 (decreased from 0.9205 to 0.1633). This demonstrates that the reconstructed images are almost completely destroyed.

Effect of High-frequency Perturbation. We evaluate the impact of our proposed IC on the `Anchor` model, with a perturbation budget of $\epsilon = \{32/255, 64/255\}$. The low, middle, and high models are trained with quality level coefficients λ 0.0130, 0.0250, and 0.0483, respectively. The results of our proposed attack on the high-quality model with $\epsilon = 32/255$ in Fig. 3, which contains three Kodak images. Table 1 presents the average MS-SSIM of the adversarial examples across 24 images. As observed in the column of adversarial examples and perturbations, LIC successfully disrupts the reconstructed adversarial examples, while the adversarial examples generated by our proposed IC achieve imperceptible perturbations. In contrast, adversarial examples without the IC show conspicuous noises. As shown in Table 2, an attack with IC can increase the average MS-SSIM by 0.078 and 0.146, respectively, demonstrating a substantial difference in image reconstruction quality.

Effect on LIC Models with Different Quality Levels. We further evaluate our proposed attack method on `Anchor` and `Factorized` LIC models across low, middle, and high reconstruction quality levels. Table 3 presents the results with varying quality levels. For example, for the low, middle, and high-quality models of `Factorized`, our proposed method consistently achieves an MS-SSIM degradation of 0.4824, 0.4166, and 0.4741, respectively, averaging 0.4577. The results illustrate that the proposed reconstruction distortion can affect all quality levels.

Findings. Our experiments have led to several intriguing observations. ❶ Besides arbitrary noises, the generated adversarial perturbation also contain certain irregular patterns. For instance, it can be observed in Fig. 3 that there are small square patterns within each generated perturbation. We hypothesize that these specific areas may have a significant impact on reconstruction quality. Future work may investigate designing countermeasures for detecting and defending the adversarial attack leveraging these patterns. ❷ Different LIC models demonstrate varying levels of robustness. From Fig. 2, we find that `Hyperprior` and `Joint` appear to be more robust than others. Based on this, we hypothesize that LIC models with higher-quality reconstruction capabilities also have superior robustness.

5. Conclusion

In this paper, we explore the robustness of LIC by launching adversarial quality attacks based on the Frobenius norm-based loss function to create adversarial examples that maximize the deviation between the original and the reconstructed images and introduce the IC to ensure the perturbations are invisible to human perception. Experiments on the Kodak dataset and various LIC models illustrate the effectiveness and reveal intriguing findings, including irregular perturbation patterns and varying levels of robustness across different LIC models.

References

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=rJxdQ3jeg`.

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.

Bégaint, J., Racapé, F., Feltman, S., and Pushparaja, A. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

Bellard, F. Bpg image format (2014). *Volume*, 1:2, 2016.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939–7948, 2020.

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., and Zhu, J. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7714–7722, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*, 2015.

Li, L., Lei, J., Gan, Z., and Liu, J. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2042–2051, 2021.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*, 2018.

Minnen, D., Ballé, J., and Toderici, G. D. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018.

Sharma, Y., Ding, G. W., and Brubaker, M. A. On the effectiveness of low frequency perturbations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3389–3396, 2019.

Shin, R. and Song, D. Jpeg-resistant adversarial images. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, volume 1, pp. 8, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 2014.

Tan, Y.-P., Kot, A. C., Yu, Y., Wang, Y., Yang, W., and Lu, S. Backdoor attacks against deep image compression via adaptive frequency trigger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12250–12259, 2023.

Taubman, D. S. and Marcellin, M. W. Jpeg2000: Image compression fundamentals. *Standards and Practice*, 11 (2), 2002.

Wallace, G. K. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., and Yuille, A. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1369–1378, 2017.