ATTACK AND DEFENSE ANALYSIS OF LEARNED IMAGE COMPRESSION

Tianyu Zhu[1], *Heming Sun*[2*], *Xiankui Xiong*[3], *Xuanpeng Zhu*[3], *Yong Gong*[4], *Minge jing*[1], *Yibo Fan*[1]

[1]State Key Lab of Integrated Chip System, Fudan University, China
[2]Faculty of Engineering, Yokohama National University, Japan
[3]State Key Laboratory of Mobile Network and Mobile Multimedia Technology, ZTE Corporation
[4]China Nanhu Academy of Electronics and Information Technology

ABSTRACT

Learned image compression (LIC) is becoming more and more popular these years with its high efficiency and outstanding compression quality. Still, the practicality against modified inputs added with specific noise could not be ignored. White-box attacks such as FGSM and PGD use only gradient to compute adversarial images that mislead LIC models to output unexpected results. Our experiments compare the effects of different dimensions such as attack methods, models, qualities, and targets, concluding that in the worst case, there is a 61.55% decrease in PSNR or a 19.15 times increase in bpp under the PGD attack. To improve their robustness, we conduct adversarial training by adding adversarial images into the training datasets, which obtains a 95.52% decrease in the R-D cost of the most vulnerable LIC model. We further test the robustness of H.266, whose better performance on reconstruction quality extends its possibility to defend one-step or iterative adversarial attacks.

Index Terms— learned image compression, adversarial attack, adversarial training, model robustness

1. INTRODUCTION

Image compression acts as the basement of all the other vision tasks. For the demands of high efficiency, LIC models have attracted more and more attention. However, the forward presentation of neural networks is usually irreversible without quantization error. This provides a fatal flaw of LIC models that could be fooled by tiny perturbations added to the input images because quantization error would be accumulated through layer-by-layer computation, leading to significant changes in the coding and reconstruction of images.

The adversarial attack is one of the attack methods that utilize the abovementioned disadvantage. The attacker intends to modify original images with imperceptible targeted noise for well-pretrained neural networks to cause undesired behaviors. As for LIC models, Chen et al. [1] carried out attacks targeting reconstruction quality by injecting random noise into the original image and optimizing it with a gradient descent algorithm. However, comparing the effects of different attack dimensions has not been researched yet.

Meanwhile, such attacks could cause danger to the applications of LIC models. For example, security monitoring meets the challenge of distinguishing clues on the screen while local users have to pay more storage resources and transmission bandwidth to obtain images. Thus, it's necessary to explore effective defense methods to weaken these attacks' impact and improve the robustness of common LIC models.

In this work, we make the following contributions:

- We conduct iterative FGSM and PGD attacks on 6 LIC models of low and high qualities (Sec. 4). Also, we conduct PGD training [2] [3] as a defense method to finetune the LIC models and raise their robustness (Sec. 5);
- We target not only reconstruction quality but also bit rate in attacks and defense, proposing R-D cost change to measure the finetuning effect of defense.
- We try a transferring attack on the conventional image compression method by using adversarial images generated by one-step FGSM, which concludes its vulnerability on bit rate (Sec. 4.4).

2. RELATED WORK

2.1. Learned Image Compression

Ballé et al. [4] first proposed a LIC model in 2016 with factorized-prior as the entropy model (named '`Fac`'), which established the optimization function aiming at the trade-off between distortion and bit rate (R-D cost). Minnen et al. [5] (named '`Mean`') and Ballé et al. [6] (named '`Hyper`') applied hierarchical hyper-prior entropy models to compact latent structure information, the latter differs in using only zero-mean Gaussian distribution. Another model put forward by Minnen et al. [5] (named '`Mbt`') added an autoregressive context model making entropy estimation even more accurate. Based on `Mbt`, Cheng et al. [7] utilized residual blocks in the analysis and synthesis transforms (named '`Anchor`'). What's more, Cheng et al. [7] added self-attention modules (named '`Attn`') to enhance the performance of image compression and reconstruction. There are quite a few works about LIC. [8] [9] [10] [11] [12] [13] [14] [15] [16]

2.2. Adversarial Attack and Defense

Most adversarial attacks can be classified as white-box attacks, which means attackers have access to all the information of target networks so that they can take advantage of the weakness. [17] The goal of a white-box attack is to generate an adversarial image x' that is similar enough to the original input x but leads to a mistaken result through the target network. Given a neural network model F, it can be formulated as:

$$max\|F(x') - F(x)\| \; s.t. \; \|x' - x\| \leq \epsilon \tag{1}$$

Here $\|\cdot\|$ quantitatively describes the difference between two items. ϵ refers to the maximum perturbation allowed to add to the original input.

Biggio et al. [18] first reported adversarial images that deceive neural networks and raised attention to the robustness of

*Corresponding author: Heming Sun (sun-heming-vg@ynu.ac.jp)

research. Many other attack methods designed for image classification tasks were proposed after that like Fast Gradient Sign Method (FGSM) [19], Basic Iterative Method (BIM) [20] [21], Projected Gradient Descent (PGD) [22] and so on.

The defense method is also a crucial topic of adversarial attack. Goodfellow et al. [19] first suggested using adversarial images generated by FGSM to train the network to learn to handle misleading inputs correctly. On their basis, Madry et al. [22] substituted FGSM with PGD so that the finetuned model can improve robustness against most single-step and iterative attacks.

3. ADVERSARIAL ATTACKS

3.1. Loss Functions

Adversarial attacks on LIC models usually focus on reconstruction quality and bit rate, as most LIC models choose R-D cost as their optimization goal.

While targeting reconstruction quality (PSNR attack), based on (1), we take peak signal-to-noise ratio (PSNR) as the explanation of $\|\cdot\|$ so that the attack loss function can be defined as:

$$L_{PSNR} = L_2\ loss(F(x'), x) \tag{2}$$

Another target is to attack bit rate (bpp attack), which is measured by bits per pixel (bpp). Since the entropy coders of LIC models are specially optimized for high efficiency and accuracy, the entropy of the quantized latent nicely estimates its exact bit length. Thus, the attack loss function is shown in (3):

$$L_{bpp} = \frac{\sum \mathbb{E}[-log_2(P(\widehat{y'}))] + \sum \mathbb{E}[-log_2(P(\widehat{z'}))]}{number\ of\ pixels}, \tag{3}$$

where P(·) means the output of the entropy model. As described in [23], for those with a hyper-prior entropy model, $\hat{z}$ is also entropy coded with $\hat{y}$ so that (3) will include an additional term of $\hat{z'}$.

3.2. FGSM Attack

FGSM [19] is a computationally efficient method for generating adversarial images. By considering the sign of the gradients, FGSM determines the direction in which the original image should be modified to create the most significant impact on the model's output in terms of a specific attack target.

Algorithm 1 FGSM attack

Input: original image X, max iteration step T, step size δ, max perturbation size ϵ
Output: adversarial image X'

Let $t = 0$, $X' = X$
while $t < T$ **do**
  Calculate attack loss L by (2) or (3)
  Calculate the gradient of the attack loss: $grad = \frac{\partial L(F(X'),X)}{\partial X}$
  Calculate and add noise: $X' = X' + \delta \cdot sign(grad)$
  Restrict the total perturbation:
  if $\|X' - X\| > \epsilon$ **then**
    Break
  end if
  $t = t + 1$
end while
return X'

As inspired, our attack adopts an iteration version of FGSM. As shown in Algorithm 1, we repeat to add noise calculated by gradients until reaching the max iteration step or upon the maximum allowed perturbation size ϵ. The step size δ is set as 1×10^{-4} so that the noise added to the target image is visually unperceivable. Meanwhile, we choose $\epsilon = 7/255$ in our experiment, which guarantees the PSNR between the adversarial and original image is around 30.

3.3. PGD Attack

PGD [22] is also a powerful and widely used adversarial attack method, which iteratively adjusts the perturbations based on the same calculation formula as FGSM.

In comparison with Algorithm 1, the difference between PGD and iterative FGSM lies in two aspects. First, we introduce random noise to the original image before turning it into optimization iterations to raise the robustness and efficiency of the attack. Second, the way of constraining total perturbation in PGD is to clamp the noise every iteration.

Moreover, Madry et al. [22] demonstrate through experiments that the loss value of PGD attack increases fairly consistently and gradually converges, untouched by various start points. Thus, the step of iteration in Algorithm 2 is settled as 40, while the step size is 0.01. Noticed that the max perturbation is 0.03 corresponding to the PSNR of around 30, similar with that in FGSM attack.

Algorithm 2 PGD attack

Input: original image X, iteration step T, step size δ, max perturbation size ϵ
Output: adversarial image X'

Add random noise α: $X' = X + \alpha$
Let $t = 0$
while $t < T$ **do**
  Calculate attack loss L by (2) or (3)
  Calculate the gradient of the attack loss: $grad = \frac{\partial L(F(X'),X)}{\partial X}$
  Obtain noise constrained by max perturbation size: $noise = clamp(X' + \delta \cdot sign(grad) - X, -\epsilon, \epsilon)$
  Obtain the adversarial image: $X' = X + noise$
  $t = t + 1$
end while
return X'

4. EXPERIMENT AND EVALUATIONS

4.1. Experiment Setup

We choose 6 LIC models mentioned in Section 2.1, which have already been trained and published in CompressAI [24]. CompressAI provides 8 quality levels for pre-trained models. For low quality, we test quality = 2 corresponding to a low bit rate and reconstruction quality. For high quality, we test quality = 5 for `Anchor` and `Attn` and quality = 8 for the others. Iterative FGSM and PGD discussed in Section 3 are universal to all these models.

For the convenience of analyzing experiment results, we define the following parameters:

$$\begin{aligned} bpp\ change &= bpp(\widehat{x'})/bpp(\widehat{x}) \\ PSNR\ change &= \frac{PSNR(x,\widehat{x'}) - PSNR(x,\widehat{x})}{PSNR(x,\widehat{x})} \end{aligned} \tag{4}$$

Table 1: Summary of bpp, PSNR, and MS-SSIM change of 6 LIC models under FGSM and PGD attack targeting reconstruction quality.

Method	Model	quality = 2			*quality = 8		
		bpp change	PSNR change	MS-SSIM change	bpp change	PSNR change	MS-SSIM change
FGSM	Fac	0.88	-11.45%	-7.89%	0.99	-26.11%	-7.98%
	Mean	0.62	-12.17%	-12.37%	1.03	-27.73%	-8.71%
	Hyper	0.67	-14.30%	-16.90%	1.01	-26.72%	-7.41%
	Mbt	0.78	-17.65%	-18.54%	1.01	-27.25%	-8.60%
	Anchor	0.67	-12.47%	-8.97%	0.79	**-30.32%**	**-10.16%**
	Attn	0.67	-12.12%	-8.54%	0.77	-20.22%	-6.16%
PGD	Fac	0.97	-10.92%	-5.81%	1.6	-35.71%	-6.20%
	Mean	0.78	-11%	-7.62%	1.92	-35.02%	-5.35%
	Hyper	0.79	-10.84%	-7.74%	1.8	-34.30%	-5.67%
	Mbt	0.72	-10.71%	-8.57%	2.58	-53.50%	-14.02%
	Anchor	0.87	-12.25%	-6.57%	2.52	**-61.55%**	**-28.41%**
	Attn	0.84	-11.51%	-6.10%	1.09	-21.03%	-7.43%

*quality = 5 for Anchor and Attn

Here $\widehat{x}$ refers to the reconstruction of x. As the name implies, bpp change describes the impact on the bit rate. However, it should be noted that in PSNR change, the PSNR of $\widehat{x'}$ is calculated with x instead of x', since what we are concerned with is widening the distance between the adversarial reconstruction and the original image.

Since gradients are necessary for both FGSM and PGD to generate adversarial images, we record the gradients of original and adversarial images separately and draw heatmaps of these gradients for visualization.

(a) original image (b) FGSM adversary (c) PGD adversary

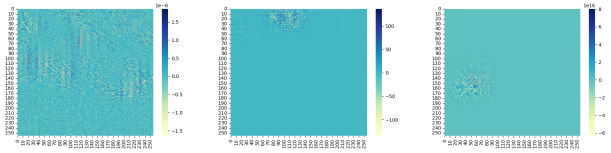
(d) original rebuilding (e) FGSM rebuilding (f) PGD rebuilding

Fig. 1: The results of the FGSM attack (the second column) and the PGD attack (the third column) targeting reconstruction quality on high-quality Anchor.

4.2. Experiment Results

In Fig. 2(a), it is easy to observe that abnormal values are usually located in the areas of complicated graphics, which vary from different models and qualities.

While targeting the reconstruction quality, abnormal values are concentrated into small areas after being attacked, which embodies the weakness adversarial attack chooses. Fig. 1(e) and Fig. 1(f) show



(a) original grade (b) FGSM grade (c) PGD grade

Fig. 2: Gradient heatmaps of original and adversarial images, corresponding to Fig. 1(a), 1(b) and 1(c).

the adversarial reconstructions through the LIC model, whose distortions appear in almost the same positions as these abnormal values in Fig. 2(b) and Fig. 2(c).

The top half of Table 1 and Table 2 summarizes the results of FGSM attack. Among all the conditions, the attack on high-quality Anchor reaches the greatest distortion loss of 30.32% change while the attack on low-quality Mbt creates the highest increase of bpp by 18.95 times.

We notice that the bpp is saved while attacking the reconstruction quality, especially on low-quality models. This is probably due to the theory of FGSM that partly smooths the gradients on the original graphic boundaries to lower the entropy estimation. There are also certain losses in PSNR and MS-SSIM while attacking bit rate, which better reflects the effectiveness of our iterative FGSM.

The results of PGD attack are shown in the bottom half of Table 1 and Table 2, whose worst cases have just the same conclusions as those in the FGSM attack. While attacking reconstruction quality on high-quality models, the bpp increases, for the reason that PGD introduces greater perturbation than FGSM, resulting in a longer bit length to express the adversarial image. Meanwhile, PSNR and MS-SSIM get better especially when attacking bit rate on low-quality models, rather than becoming worse in FGSM, due to the stability of PGD.

4.3. Discussion

Attack methods. Compared with FGSM, adding random noise initially improves the error-tolerant rate of PGD so that it can take a larger step size and fewer iterations to reach the convergence, and it is also one-tenth less time-consuming than FGSM.

Table 2: Summary of bpp, PSNR, and MS-SSIM change of 6 LIC models under FGSM and PGD attack targeting bit rate.

Method	Model	quality = 2			*quality = 8		
		bpp change	PSNR change	MS-SSIM change	bpp change	PSNR change	MS-SSIM change
FGSM	Fac	1.75	-5.23%	-3.18%	2.61	-22.80%	-3.58%
	Mean	2.79	-1.69%	-0.29%	16.32	-26.80%	-5.60%
	Hyper	3.28	-2.71%	-0.89%	6.59	-29.66%	-5.72%
	Mbt	**18.95**	-1.69%	-0.19%	11	-28.54%	-5.97%
	Anchor	3.57	-5.16%	-1.93%	10.79	-24.51%	-7.35%
	Attn	2.35	-4.51%	-2.52%	2.94	-15.75%	-5.46%
PGD	Fac	1.58	-1.96%	-1.55%	2.44	-17.72%	-2.21%
	Mean	2.48	0.23%	-0.03%	17.73	-20.20%	-2.35%
	Hyper	2.62	1.04%	0.56%	5.34	-17.28%	-2.02%
	Mbt	**19.15**	0.85%	0.51%	10.6	-19.68%	-2.34%
	Anchor	3.04	-0.69%	-0.63%	14.79	-10.85%	-2.81%
	Attn	2.27	-1.09%	-1.86%	3.73	-8.40%	-3.04%

*quality = 5 for Anchor and Attn

The MS-SSIM change of high-quality Anchor in PGD attack is greater than that in FGSM attack because MS-SSIM emphasizes human perception. It can be confirmed by their reconstructions, as shown in Fig 1(e) and Fig 1(f), more visible distortions exist in the reconstruction after the PGD attack. Meanwhile, abnormal values over the gradient heatmaps in the PGD attack also appear more conspicuous if we compare Fig. 2(b) with Fig. 2(c).

However, iterative FGSM shows its advantage in attacking low-quality models, perhaps because the smaller step size makes iterative FGSM fine to discover weaknesses in low-bit-rate images.

Models. For the PSNR attack, Anchor using residual blocks with small convolutions is the most vulnerable while the PSNR change of Attn is only one-third of the former. For the bpp attack, Fac is the most robust despite its simplicity. Low-quality Mbt with autoregressive context model makes the greatest loss in bit rate.

In general, Attn with self-attention block shows the best robustness in all the conditions, which means the self-attention block contributes to comprehending the original information of adversarial images.

Qualities. The PSNR change of high-quality models is more dramatic since high bpp retains more information including noise during the compression process. The results of the bpp attack follow the above regularities except that low-quality Mbt is more vulnerable than the high-quality one.

4.4. Attack on Conventional Methods

Conventional methods use reversible transform to compress image information, which is simpler than LIC models. Generally speaking, gradient-based methods are invalid for conventional methods. Thus, take H.266 as an example, we consider if the high transferability of one-step FGSM discovered by Papernot et al. [2] and Dong et al. [3] is effective in transferring attack to H.266.

In the black-box setting, we adopt one-step FGSM on the most vulnerable LIC model to generate adversarial images and use VTM to encode them, while our VTM encoder version is 19.0. Compared with high-quality Anchor, we set the quantization parameter Q=20 in the PSNR attack. Similarly, Q is set as 40 in the bpp attack. The overall coding results are shown in Table 3.

As for the PSNR attack, H.266 performs much better than LIC models, since frequency domain transform smooths the high-frequency noise FGSM adds. However, the cost is that the bpp of H.266 also raises in the PSNR attack, as the high-frequency noise is more difficult to compress and occupies a longer bit length. That's why H.266 also shows its vulnerability in the bpp attack. Even so, conventional methods can probably be used as a defense measure against the PSNR attack if higher bpp is allowable.

Table 3: Results of transferring attack on H.266 targeting reconstruction quality and bit rate

	PSNR attack				**bpp attack**			
	Anchor (quality=5)		H.266 (Q=20)		Mbt (quality=2)		H.266 (Q=40)	
	PSNR	bpp	PSNR	bpp	PSNR	bpp	PSNR	bpp
Origin	35.12	0.59	46	0.66	29.64	0.19	34.44	0.04
Adversary	8.52	1.55	43.58	2.4	23.6	0.56	26.97	0.15

5. DEFENSE METHOD

5.1. Adversarial Training

Known as PGD training, the defense method proposed by Madry et al. [22] is one of the most effective ways to weaken the misdirection of neural networks caused by adversarial attacks. PGD training uses adversarial images generated by PGD to train the network. Thus, except for its effectiveness, PGD training requires numerous computing resources and high time complexity (relevant to the iteration steps of the PGD attack).

However, Madry et al. designed this method for image classification tasks rather than compression. The difference is embodied in that classification tasks do not pursue reconstruction quality, but only demand recognition accuracy. In contrast, a good reconstruction must be close enough to the original image for image compression. Hence, the finetuned LIC models should learn the features from not only adversarial but also original images. We modify the algorithm by adding adversarial images into the training dataset instead of replacing it, as shown in Algorithm 3. The finetuning goal can be changed by the target of the PGD attack (line 4 in Algorithm 3).

Our training set uses randomly cropped 256*256*3 patches of 1633 images from the clic2020 dataset [25] as the inputs, with a train batch size of 16. The λ in R-D cost is the same as the corresponding pre-trained models in CompressAI. The initial learning rate is set as 0.0002 while the max epoch is 200.

Table 4: Finetuning results of adversarial training against PSNR attack and bpp attack

	Denfense against PSNR attack				**Denfense against bpp attack**			
	PSNR change (attacked)	R-D cost on original images	R-D cost on adversarial images	R-D cost loss (%)	bpp change (attacked)	R-D cost on original images	R-D cost on adversarial images	R-D cost loss (%)
Pre-trained model	-61.55%	1.095	74.05	6662.00%	19.15	0.434	3.813	778.00%
Finetuned model	-12.28%	1.69	3.316	96.21%	1.44	0.583	0.556	-4.63%
Finetuning effect*	-	54.34%	**-95.52%**	-	-	34.33%	**-85.42%**	-

*Finetuning effect is represented by the R-D cost change, which is calculated similarly to the PSNR change as in (4).

Algorithm 3 PGD training

Input: Randomly initialized network F, original training set X

1: **repeat**
2: Read minibatch $B = \{x^1, \cdots, x^m\}$ from X
3: Train one step of F with minibatch B
4: Generate adversarial minibatch $B' = \{x^1_{adv}, \cdots, x^m_{adv}\}$ by PGD attack
5: Train one step of F with minibatch B'
6: Valid F with both minibatch B and B'
7: **until** training convergence

5.2. Defense Efficiency

To evaluate the effect of our modified defense method, we attack pre-trained and finetuned models with PGD and compare their R-D costs. Settings of the PGD attack are the same as in Sec. 3.3. Although PGD training is specifically designed for one finetuning goal, sacrificing image quality to obtain lower bpp is unacceptable.

As shown in Table 4, the horizontal comparison describes the loss brought by adversarial images while the vertical comparison indicates the effect of model finetuning.

We choose high-quality `Anchor` as the target for defending against the PSNR attack because it is the most vulnerable model while attacking the reconstruction quality. Table 4 shows that the loss from adversarial images is greatly weakened. Furthermore, the R-D cost of the finetuned model on adversarial images decreases by 95.52%. Similarly, we use low-quality `Mbt` to defend against the bpp attack. The fine-tuned model reached an 85.42% R-D cost decrease on adversarial images, even lower than the R-D cost on original images as the PGD attack raises the PSNR of reconstructions. However, the R-D cost of the finetuned model increases on original images compared to the pre-trained model, which is inevitable due to the 'polluted' dataset in the training process.

Since the PGD attack is to find the most adversarial image in the neighbor ball around the original image, our experiment results prove that the modified algorithm is effective in defending both one-step and iterative adversarial attacks, whether targeting for reconstruction quality or bit rate. This method could be generalized to most LIC models because of its successful defense of the most vulnerable ones in our experiment.

6. CONCLUSION

In this paper, we research the robustness of learned image compression networks under gradient-based adversarial attack methods. We demonstrate the decrease of compression ability against either PSNR or bpp attacks. We also test the robustness of H.266, which prompts future research that conventional methods could be used to defend against adversarial attacks. Furthermore, we supplement the defense method of adversarial training to enhance the performance of LIC models under attack.

7. REFERENCES

[1] Tong Chen and Zhan Ma, "Towards robust neural image compression: Adversarial attack and model finetuning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[2] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[3] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[5] 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.

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[7] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.

[8] Haojie Liu, Tong Chen, Qiu Shen, Tao Yue, and Zhan Ma, "Deep image compression via end-to-end learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2575–2578.

[9] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 221–231.

[10] Yueyu Hu, Wenhan Yang, and Jiaying Liu, "Coarse-to-fine hyper-prior modeling for learned image compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11013–11020.

[11] David Minnen and Saurabh Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3339–3343.

[12] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Transactions on Image Processing*, vol. 30, pp. 3179–3191, 2021.

[13] Yueqi Xie, Ka Leong Cheng, and Qifeng Chen, "Enhanced invertible encoding for learned image compression," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 162–170.

[14] Zongyu Guo, Zhizheng Zhang, Runsen Feng, and Zhibo Chen, "Causal contextual prediction for learned image compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2329–2341, 2021.

[15] Yueyu Hu, Wenhan Yang, Zhan Ma, and Jiaying Liu, "Learning end-to-end lossy image compression: A benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4194–4211, 2022.

[16] Meng Li, Shangyin Gao, Yihui Feng, Yibo Shi, and Jing Wang, "Content-oriented learned image compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 632–647.

[17] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, pp. 151–178, 2020.

[18] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*. Springer, 2013, pp. 387–402.

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[21] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[23] Heming Sun, Lu Yu, and Jiro Katto, "Improving latent quantization of learned image compression with gradient scaling," in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2022, pp. 1–5.

[24] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja, "Compressai: a pytorch library and evaluation platform for end-to-end compression research," *arXiv preprint arXiv:2011.03029*, 2020.

[25] George Toderici, Wenzhe Shi, Lucas Theis Radu Timofte, Johannes Balle, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer, "Workshop and challenge on learned image compression (clic2020)," 2020.