KEEP IT REAL: CHALLENGES IN ATTACKING COMPRESSION-BASED ADVERSARIAL PURIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Previous work has suggested that preprocessing images through lossy compression can defend against adversarial perturbations, but comprehensive attack evaluations have been lacking. In this paper, we construct strong white-box and adaptive attacks against various compression models and identify a critical challenge for attackers: high realism in reconstructed images significantly increases attack difficulty. Through rigorous evaluation across multiple attack scenarios, we demonstrate that compression models capable of producing realistic, high-fidelity reconstructions are substantially more resistant to our attacks. In contrast, low-realism compression models can be broken. Our analysis reveals that this is not due to gradient masking. Rather, realistic reconstructions maintaining distributional alignment with natural images seem to offer inherent robustness. This work highlights a significant obstacle for future adversarial attacks and suggests that developing more effective techniques to overcome realism represents an essential challenge for comprehensive security evaluation.

1 INTRODUCTION

Adversarial attacks on image classification models involve making small perturbations to an image such that the classifier's output changes–even though the image appears semantically unchanged to a human observer. A model can be trained to be robust against such adversarial examples, for example, by augmenting the training data with noise or showing the model adversarial examples during training. Another strategy is applying a transformation to the input image that preserves its semantic content while altering it to render the adversarial noise ineffective. This approach has the advantage of being able to be used for any classification model without requiring retraining.

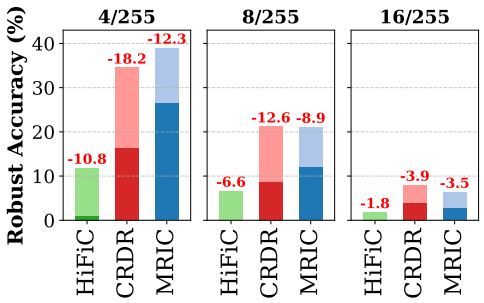

Figure 1: Decrease in robust accuracy when employing a compression defense with reduced realism under different perturbation budgets. Incorporating realism substantially increases the difficulty of successful attacks.

Lossy image compression often discards details deemed perceptually unimportant, which may include the subtle perturbations introduced by adversarial attacks. Early work argued that standard codecs like JPEG yield modest robustness gains (Guo et al., 2017), though they often introduce compression artifacts that push images outside the classifier's training distribution. More recently, learned compression models have promised to bridge this gap by generating visually plausible reconstructions that may also remove stronger adversarial noise (Räber et al., 2025).

However, many proposed defenses—especially preprocessing-based ones—have been criticized for relying on gradient masking; they make general gradient-based attacks harder without genuinely increasing robustness (Shin & Song, 2017). When attackers adapt to circumvent gradient obfuscation (e.g., by approximating or smoothing gradients), many defenses fail (Carlini et al., 2019; Tramer et al., 2020; Sheatsley et al., 2023; Zhang et al., 2025). This raises two critical questions:

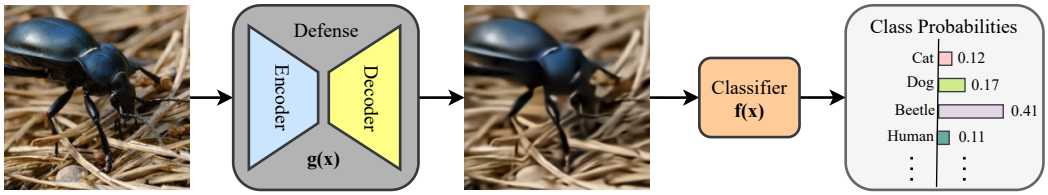

Figure 2: Overview of compression-based adversarial defense. An input image (potentially containing adversarial perturbations) is first processed by the defense module, which consists of an encoder-decoder architecture that compresses and reconstructs the image. This reconstructed image is then passed to a classifier, which outputs class probabilities. The defense aims to indirectly, through the compression process, remove adversarial noise while preserving the semantic content needed for correct classification.

(1) *Do robustness gains from image compression persist under rigorous, adaptive attacks?*
(2) *If so, what underlying mechanism contributes to this robustness?*

We argue that realism in reconstructed images is a key factor underpinning the robustness. Our results show that only compression models capable of producing high-fidelity, realistic images offer meaningful robustness against strong adaptive adversarial attacks, while other compression models can be broken. Realism aids robustness in two ways: It avoids unnatural artifacts that shift images off-distribution, and by hallucinating in semantically plausible details that obscure adversarial noise.

A highly realistic compression model reconstructs images close to the original and free from perceptible signs of compression. Classifier models often perform poorly on out-of-distribution inputs; ensuring realism in reconstructed images helps keep them within the distribution of natural images. High realism can be achieved by hallucinating in plausible details. For example, while the exact texture of tree leaves may be removed during compression, a realistic model will hallucinate plausible leaf-like textures. These added details can help obscure adversarial noise, making it more difficult for an attacker to craft successful perturbations.

In this paper, we systematically evaluate the robustness of compression-based defenses and isolate the role of realism. Our analysis builds on the findings in (Räber et al., 2025), where it was argued that human-aligned compression contributes to robustness. We re-evaluate their claims under rigorous adversarial threat models and identify shortcomings in their evaluation protocol. We show that their observed robustness stemmed not merely from compression, but specifically from realism. We reinforce this conclusion through extensive evaluation across a broader and newer set of learned compression models.

2 Background and Related Work

Our work lies at the intersection of two big fields, and we aim to give complete overviews of both. However, due to the page limit, we had to move the complete related work to Section C.

2.1 Adversarial Robustness

Shortly after the success of AlexNet (Krizhevsky et al., 2012), it was found that neural networks are very susceptible to *adversarial attacks* (Szegedy et al., 2014; Goodfellow et al., 2015). Here, an adversary adds small and imperceptible perturbations to an image such that a model mislabels it.

Attacks Many attacks have been developed over the years with various benefits and drawbacks. Some of the most noteworthy are FGSM (Fast Gradient Sign Method) (Goodfellow et al., 2015), iterated FGSM or iFGSM (Kurakin et al., 2017), CW (Carlini & Wagner, 2017b), and PGD (Projected Gradient Descent) (Madry et al., 2019). These attacks fall into two categories: FGSM, iFGSM, and PGD are l_∞-bounded attacks, and CW is an l_2-bounded attack. We refer the reader to the original papers for specific details, but include some details in Section D. The TLDR is: PGD has a perturbation budget ϵ and an iteration budget n. PGD does projected gradient descent for n iterations to find an adversarial example at most ϵ distance in l_∞ from the original image.

When the defense includes randomness, a common attack augmentation is EoT (Expectation over Transformation), where gradients are averaged over multiple backward passes (Athalye et al., 2018a). If the defense causes gradient masking, Athalye et al. (2018a) suggested Backward Pass Differentiable Approximation (BPDA), where the defense is used during the forward pass, but a differentiable approximation is used during the backward pass. The simplest is to use the identity function, while better methods train a differentiable surrogate model g' that emulates the defense g; i.e., $g(x) \approx g'(x)$.

Lastly, often the best attack results are found using *adaptive attacks* (Carlini et al., 2019; Sheatsley et al., 2023; Zhang et al., 2025). Here, the attacks, for instance, the optimization objective, are adjusted to the defense. We focus on adaptive attacks with PGD as the optimization method.

Defenses Given the prevalence of adversarial attacks, many researchers have explored how to defend against them. We broadly view this in three groups: Architecture improvements (Singh et al., 2023; Wu et al., 2023; Fort & Lakshminarayanan, 2024), adversarial training (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2019; Pang et al., 2020b; Fort & Lakshminarayanan, 2024), and adversarial purification (Nie et al., 2022; Frosio & Kautz, 2023; Zollicoffer et al., 2025; Räber et al., 2025). The architecture branch focuses on making the models more robust by design, for instance, by taking inspiration from biology to mimic the human eye when training a ResNet model (Fort & Lakshminarayanan, 2024). Adversarial training consists of showing adversarial examples during training to ensure correct classifications. This has yielded positive results (Szegedy et al., 2014; Madry et al., 2019), but models can be broken (Zhang et al., 2025).

Adversarial purification aims to remove the adversarial noise in images before passing the cleaned images to classification models. It is motivated by the work of Ilyas et al. (2019), hinting that adversarial examples perturb brittle features in the model. These methods *should work independently* of any robustness applied to the classifier through adversarial training or robust architectures.

Diffusion models have been proposed as a way to remove adversarial noise from input images (Nie et al., 2022; Lee & Kim, 2023). These approaches are conceptually similar to compression-based defenses with realism: both aim to project adversarial examples back onto the manifold of natural images to restore classifier performance. However, prior work on diffusion-based purification has not explicitly investigated the role of realism as a contributing factor to robustness. One major drawback of diffusion-based defenses is their computational cost (Dhariwal & Nichol, 2021; Yang & Mandt, 2023b). But simply making gradients difficult to compute does not equate to genuine robustness. The purification method proposed by Nie et al. (2022) was later defeated by Lee & Kim (2023). However, the latter only evaluated their improved defense under the attack that broke the former. They did not develop new adaptive attacks tailored to their defense—a strategy that past research suggests would likely reduce the effectiveness of the defense. The history of adversarial robustness research shows that defenses that are not tested under strong, tailored attacks often overstate their robustness (Athalye et al., 2018a; Carlini et al., 2019; Tramer et al., 2020).

2.2 Realism in Image Compression

Image compression algorithms are traditionally evaluated using *distortion metrics*, which measure the distance between a restored image $\hat{x}$ and its reference x. Formally, distortion is defined as:

$$\mathcal{D} := \mathbb{E}_{(x,\hat{x})\sim(p_X,p_{\hat{X}})}[\Delta(x,\hat{x})], \tag{1}$$

where $\Delta(\cdot,\cdot)$ is a pointwise distortion measure (e.g., ℓ_2 distance), and p_X, $p_{\hat{X}}$ denote the distributions of ground truth and reconstructed images. Distortion is considered a *full-reference* metric, requiring access to the original image x to compare it to the reconstructed version $\hat{x}$.

However, metrics such as PSNR, MS-SSIM (Wang et al., 2003), or LPIPS (Zhang et al., 2018) correlate poorly with human perception, and directly quantifying perceptual distortion remains a challenging problem. Instead of a perceptual distortion metric, one can measure both distortion and realism and optimize the compression model for both. Realism can be formally defined as:

$$\mathcal{R} := -d(p_{\hat{X}}, p_X), \tag{2}$$

where $d(\cdot,\cdot)$ is a divergence measure, such as Kullback-Leibler. Unlike distortion, realism is a *no-reference metric*, requiring only the generated image distribution to match that of natural images. Although measuring realism is challenging (Theis, 2024), a widely used proxy is the Fréchet Inception

Distance (FID) (Heusel et al., 2017), which compares the distributions extracted by a pretrained network, and has gained popularity for capturing both fidelity and diversity in generated samples.

Compression models are typically trained with the following loss function:

$$\mathcal{L} = \mathcal{L}_{\text{RATE}} + \lambda\mathcal{D} - \beta\mathcal{R} \tag{3}$$

$\mathcal{L}_{\text{RATE}}$ is the estimated rate (the number of bits required to represent the image after compression), λ controls the level of distortion, and β the level of realism. $\mathcal{L}_{\text{RATE}}$ does not play a critical role in this context, as information is not transmitted through an explicit information bottleneck in our approach.

Compression models can be broadly categorized into those optimized for distortion (e.g., JPEG (Wallace, 1991), Hyperprior (Ballé et al., 2016), ELIC (He et al., 2022a)) and those explicitly trained to maximize realism (e.g., HiFiC (Mentzer et al., 2020), PO-ELIC (He et al., 2022b)). More recent approaches, such as MRIC (Agustsson et al., 2023) and CRDR (Iwai et al., 2024), provide explicit control over the realism–distortion tradeoff within a single network by conditioning on λ and β. This controllability enables direct investigation of how realism influences adversarial robustness.

Our work builds on this foundation, exploring the intersection of realistic compression and adversarial robustness. We extend the existing literature by providing experimental evidence that realism, rather than distortion, makes image-compression models (partially) robust against adversarial examples.

2.3 Compression as Adversarial Defense

Using compression as a defense for neural networks is not a new idea. It has been explored for almost a decade (Dziugaite et al., 2016; Das et al., 2017; Jia et al., 2019) where some authors also explored using iterated compression and decompression cycles (Ferrari et al., 2023; Räber et al., 2025). Two key works in this area are by Guo et al. (2017) and Shin & Song (2017); the former argued that JPEG compression as a preprocessing step is a very effective adversarial defense, while the latter showed that, by making JPEG differentiable, this defense could be bypassed entirely–highlighting the necessity of properly evaluating a defense.

3 Methodology

Our work evaluates the compression models' robustness; thus, we focus on the ImageNet classification benchmark, a well-established benchmark with high-resolution images. This gives us many options for pretrained models, allowing a more exhaustive model evaluation. If otherwise not stated, we use the full validation split of the ImageNet dataset (50000 images). We used two different classification models, ResNet50 (He et al., 2016) and ViT B 16 (Dosovitskiy et al., 2020). For PyTorch models, we use the improved weights `IMAGENET1K_V2` for ResNet50 and the default weights `IMAGENET1K_V1` for ViT B 16. For TensorFlow models, we use the default ResNet50 pretrained weights `imagenet`. Certain ablations are only done for ResNet to reduce the compute load; the results have only minor differences from ViT.

3.1 Defenses

In the evaluated defense strategy, we integrate a compression and decompression step into the image classification pipeline (cf. Figure 2). The image compression model acts as a preprocessing step, transforming the image before classification. The transformation discards certain information from the image by employing lossy compression techniques. This process can mitigate the effect of adversarial perturbation on the classifier's predictions, thereby enhancing the model's robustness.

As we demonstrate later, realism plays a crucial role in the robustness of this pipeline. Therefore, we focus our experiments on models that either explicitly control the level of realism or exist in both standard and enhanced realism variants. In particular, MRIC (Agustsson et al., 2023) and CRDR (Iwai et al., 2024) offer variable realism settings, and we denote their low- and high-realism variants as *MRIC LR*, *MRIC HR*, *CRDR LR*, and *CRDR HR*, respectively. We also consider models available in both standard rate-distortion and rate-distortion-realism versions, such as *Hyperprior* (Ballé et al., 2018) and *HiFiC* (Mentzer et al., 2020). To further validate the importance of realism, we include *JPEG* (Wallace, 1991) and *ELIC* (He et al., 2022a) (noting that the high-realism version, PO-ELIC (He et al., 2022b), does not have publicly available pretrained weights), and show that these

consistently underperform compared to high-realism compression models. For white-box attacks, we require access to gradients; for learned compression models, gradient computation is natively supported, while for JPEG, we employ a continuous relaxation approach (Shin & Song, 2017). We focus on VAE-based methods as diffusion or INR-based approaches are prohibitively expensive to evaluate at scale (Nie et al., 2022; Lee & Kim, 2023).

3.2 Threat models

We define our own threat models that encompass the prior experiments from Räber et al. (2025) to ensure proper evaluation (Carlini & Wagner, 2017a; Carlini et al., 2019; Rando et al., 2025). The threat models follow standard formulations from Biggio et al. (2013); Pang et al. (2020a). Specifically, we consider l_∞ untargeted attacks with perturbation budget ϵ, i.e., for an original image x and perturbed image x', we have $\|x - x'\|_\infty \leq \epsilon$.

Our primary focus is on PGD attacks, widely regarded as among the strongest given sufficient objective formulation and computational resources (see Section D). We also include adaptive attacks, where the adversary, aware of the defense mechanism, tailors the attack accordingly. Adaptation in this context means the adversary can specialize the attack by studying the defense and looking for ways to defeat it (Tramer et al., 2020). These adaptive attacks still use PGD (see Section 3.3 for details). If otherwise not stated, we use 10 PGD iterations.

Lastly, we consider three adversary knowledge levels as described in Carlini & Wagner (2017a). **Black-box (BB)**: The adversary does not know the defense or its existence. The attacker can create adversarial perturbations with respect to the gradients of the classifier or the outputs of the classifier. **Gray-box (GB)**: The adversary knows the defense is present and can use it for the forward pass only; they cannot compute gradients through the defense. This threat variant is used for adaptive attacks. **White-box (WB)**: The adversary has complete knowledge and access to the defense. The gradients of the defense and the output of the combined defense and classifier pipeline are used in the attack.

3.3 Adaptive Attacks Against Compression Defenses

A common pitfall in adversarial robustness papers is the lack of honest effort in implementing proper attacks against one's own defense (Carlini et al., 2019; Tramer et al., 2020; Sheatsley et al., 2023). In our evaluation, we conduct adaptive attacks on all compression models, which means the attacks are tailored to exploit the target model's weaknesses. This section reviews the adaptive attacks used against the compression defenses. We use the following notation: f is the image classifier, g is a compression defense (compression and decompression), and $h = f \circ g$. x is an image with label y, and L is a loss function. We use cross-entropy as the loss function for all our experiments (except ACM). $\nabla_x f$ is the gradient of $L(f(x), y)$ with respect to x.

ST BPDA For the first adaptive attack, we use BPDA with the straight through (ST) approximation; this assumes $g(x) \approx x$. For the forward pass, the defense is used, but the straight-through (ST) estimator is applied in the backward pass, replacing the compression model's gradient with the identity function. Thus, $\nabla_x h := \nabla_x f(x)|_{x=g(x)}$.

U-Net BPDA The second adaptive attack uses BPDA with a U-Net (Ronneberger et al., 2015) trained to approximate the compression defense g (see training details in Section G). The idea is that if g primarily causes gradient masking, then replacing it with a differentiable proxy g' should yield meaningful gradients for the attack. During the forward pass, we use the actual defense, $h(x) = f(g(x))$, while the backward pass substitutes g with g', yielding gradients $\nabla_x h := \nabla_x (f \circ g')(x)$.

Attacks on the Compression Model (ACM) The third adaptive attack focuses only on attacking the compression method. Instead of focusing on misclassifications in the classifier through the cross-entropy loss, the attacker uses the objective function $MSE(x, g(x))$ for the attack. The goal is then to cause significant distortions that confuse the classifier.

Adaptive Realism Attack (ARA) The fourth adaptive attack is specific to models with varying realism. Let g_β be a defense with realism parameter β, the goal of the attack is then for a given β to find the β' that gives the lowest model accuracy when we attack $f(g_\beta(x))$ with $\nabla_x h := \nabla_x (f \circ g_{\beta'})$.

Table 1: ResNet accuracy under adaptive attacks on compression models for ImageNet. Hyperprior and HiFiC results are combined to give the low and high realism. JPEG and ELIC do not have a high realism (HR) version; only low realism (LR) is available. Note, for MRIC, only the PGD attack was applied. We selected each architecture's most effective adaptive attack based on evaluations in Table 3 and Table 6. We see that HR models consistently perform better.

		4/255		8/255		16/255	
CLASSIFIER	DEFENSE	LR	HR	LR	HR	LR	HR
RESNET	HYPERPRIOR	0.98	**11.83**	0.02	**6.65**	0.00	**1.80**
	MRIC	26.68	**39.00**	12.20	**21.10**	2.90	**6.40**
	CRDR	16.30	**34.50**	8.60	**21.18**	4.10	**7.98**
	JPEG	5.19	—	0.24	—	0.01	—
	ELIC	16.43	—	4.98	—	0.34	—
VIT	HYPERPRIOR	0.20	**10.06**	0.00	**2.64**	0.00	**0.16**
	CRDR	14.44	**28.28**	6.98	**14.90**	2.16	**2.46**
	JPEG	0.88	—	0.04	—	0.00	—
	ELIC	14.12	—	2.38	—	0.10	—

Considerations for the Attacks All the VAE-based compression defenses are deterministic; thus, EoT (Athalye et al., 2018b) should not be required for these defenses (Athalye et al., 2018a).

Adaptive attacks aim to craft adversarial examples explicitly tailored to the defense mechanism—in our case, the compression model. White-box PGD (WB PGD) and U-Net BPDA perform best in our evaluated attacks. The primary difference between WB PGD and U-Net BPDA is that U-Net BPDA approximates the backward pass of the compression model with the gradients of U-Net trained to mimic the outputs of the compression model, allowing gradients to flow more freely.

4 EXPERIMENTS

We present a series of experiments to support our claim that realism makes compression-based defenses hard to attack. First, we demonstrate that defenses incorporating realism consistently outperform those that do not (Section 4.1). Next, we verify that the observed robustness does not arise from gradient masking (Section 4.2). Finally, we construct stronger model-specific adaptive attacks, showing that even under these conditions, high-realism defenses remain more robust (Section 4.3).

4.1 ROLE OF REALISM

Table 1 shows that models incorporating realism into their reconstructions consistently achieve higher robust accuracy under strong adaptive attacks. These results provide compelling empirical support for our central hypothesis: *realism plays a key role in reducing the effectiveness of adversarial attacks*. ViT models exhibit lower overall robustness than ResNets across the board.

To isolate the impact of realism, we analyze two compression models where realism can be explicitly controlled. As shown in Figure 6 (Section H.2), increasing realism in reconstructed images monotonically improves robustness. While distortion has been the focus of extensive prior work (Dziugaite et al., 2016; Shin & Song, 2017; Räber et al., 2025), realism remains largely unexplored as a factor in defense performance. Our results highlight a key difference: distortion presents an inherent trade-off. If the distortion is too low, adversarial noise is preserved. If the distortion is too high, the reconstruction discards critical semantic information, making the image unrecognizable. In both cases, the defense fails—either by retaining harmful perturbations or by producing images that are implausible under the natural data distribution. In contrast, increasing realism consistently improves robustness without the trade-offs typically associated with distortion. Prior to this work, the role of realism in adversarial defenses had not been thoroughly investigated.

These findings suggest that compression alone is insufficient as a defense mechanism. Without realism, compression may eliminate adversarial perturbations but also introduce artifacts that push reconstructions off the natural data manifold. Such off-distribution reconstructions can make downstream classifiers more vulnerable to attack. Realistic reconstructions, by contrast, preserve semantic

Table 2: WB PGD attacks with varying numbers of iterations. Only 5000 samples were used due to the increased computational cost. *Hyperprior Noise* refers to a defense in which the gradients through the hyperprior are replaced with a random vector, showing that it suffers from gradient masking.

Iterations	Defense	Standard	2/255	4/255	8/255	16/255
400	Hyperprior	79.20	79.06	78.92	79.18	77.84
	Hyperprior N	79.20	79.10	79.22	78.74	76.58
	JPEG	71.02	1.44	0.26	0.02	0.02
	CRDR LR	46.14	19.60	7.68	1.52	0.26
	CRDR HR	62.02	36.92	19.22	6.46	1.30
200	CRDR LR	46.14	19.52	7.70	1.70	0.28
	CRDR HR	62.02	36.70	19.94	7.46	1.84
100	CRDR LR	46.14	20.14	8.24	2.42	0.38
	CRDR HR	62.02	37.12	20.84	8.86	2.54
50	CRDR LR	46.14	20.44	9.38	3.12	0.70
	CRDR HR	62.02	38.08	23.30	10.86	4.06
10	CRDR LR	46.14	26.40	16.24	8.56	3.84
	CRDR HR	62.02	46.08	36.10	24.36	13.86

content while introducing plausible details, helping keep inputs within the natural data distribution and more effectively mask adversarial signals.

4.2 Gradient Masking

Modifying the attack budget ϵ does not influence gradient masking as it merely changes the size of the space in which attacks can search for adversarial examples. Therefore, if the model consistently fails under higher ϵ values, the robustness observed at smaller ϵ values is more likely to reflect genuine defense rather than gradient obfuscation. In Table 2, we evaluate CRDR with both low and high realism settings under a range of ϵ values and PGD iteration counts. Under a 400-step PGD attack with high ϵ, all models fail except for Hyperprior. To investigate this anomaly, we include a "Hyperprior Noise" variant, which replaces Hyperprior's gradient with random noise. Its comparable performance suggests that Hyperprior's apparent robustness is attributable to gradient masking. As the attack budget ϵ is reduced, the resulting accuracy gains reflect genuine robustness rather than gradient obfuscation. Realism proves critical: at $\epsilon = 4/255$, increasing realism improves CRDR's robust accuracy from 7.68% to 19.22%. This sharp gain underscores realism's essential role in compression-based adversarial defenses.

To complement our quantitative evaluations, in Figure 3 we visualize the loss landscapes for the classifier, CRDR with low realism, and CRDR with high realism, following the implementation described by Zhang et al. (2025). CRDR with low realism exhibits some level of gradient masking,

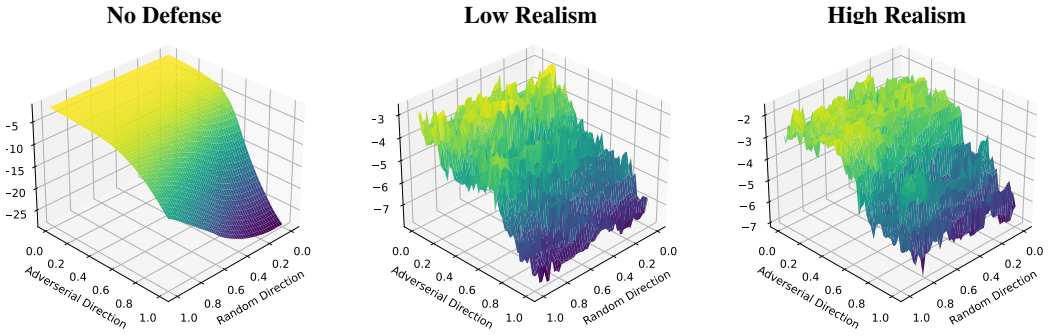

Figure 3: Loss landscapes under successful 100-step PGD attacks on a ResNet with CRDR defense. **Left**: Attacking the classifier directly. **Middle**: Attacking with low realism defense. **Right**: Attacking high realism defense. The standard deviations of the loss surfaces are 0.0544, 0.3343, and 0.3156, respectively. Increasing realism does not make the loss landscape spikier, indicating that it does not contribute to gradient masking.

Table 3: Results for ResNet50. "—" denotes values not implemented for MRIC or evaluations invalid without realism control. Models with strong gradient masking, like Hyperprior, are vulnerable to adaptive attacks. Realism does not increase gradient obfuscation, so both high-realism models retain accuracy with minor drops under adaptive attacks. Due to space, the 16/255 are omitted from this table. See Table 5 for these results.

STRENGTH	DEFENSE	STANDARD	BB PGD	WB PGD	ST BPDA	U-NET BPDA	ACM	ARA
4/255	HYPERPRIOR	78.73	48.76	78.84	10.94	**0.98**	78.82	—
	HIFIC	61.53	59.52	**11.83**	44.65	24.04	59.20	—
	MRIC LR	52.80	51.04	**26.68**	—	—	—	**26.68**
	MRIC HR	63.06	59.28	**39.00**	—	—	—	39.60
	CRDR LR	46.02	44.92	**16.30**	39.36	28.96	41.28	**16.30**
	CRDR HR	61.72	59.80	35.88	56.36	47.62	55.67	**34.50**
	JPEG	70.30	65.68	8.72	18.70	**5.19**	68.67	—
	ELIC	60.03	58.79	**16.43**	40.00	17.98	54.48	—
8/255	HYPERPRIOR	78.73	27.16	78.72	3.18	**0.02**	78.67	—
	HIFIC	61.53	56.84	**6.65**	32.19	7.98	55.37	—
	MRIC LR	52.80	49.44	**12.20**	—	—	—	**12.20**
	MRIC HR	63.06	57.38	21.98	—	—	—	**21.10**
	CRDR LR	46.02	44.14	**8.60**	33.58	12.12	36.96	**8.60**
	CRDR HR	61.72	57.36	23.92	50.77	26.64	49.35	**21.18**
	JPEG	70.30	60.76	5.19	10.01	**0.24**	64.75	—
	ELIC	60.03	57.59	9.01	27.24	**4.98**	51.63	—

as reflected in the spiky and irregular structure of its loss landscape. Importantly, increasing realism does not increase this effect—the loss landscape remains similarly smooth, suggesting that both low- and high-realism models exhibit comparable levels of gradient masking. This observation is supported quantitatively: the standard deviations of the loss surfaces for the classifier, low-realism, and high-realism models are 0.0544, 0.3343, and 0.3156, respectively. Despite this, CRDR with high realism consistently outperforms its low-realism counterpart. The loss landscape analysis further supports this conclusion: while both models may exhibit some degree of gradient masking, increased realism does not exacerbate it and plays a direct role in enhancing robustness.

4.3 ADAPTIVE ATTACKS

Many defenses rely on gradient masking, hiding gradients to appear robust rather than truly resisting attacks. This raises the question: Is realism just gradient masking or genuinely robust? To address this, we perform extensive adaptive attacks designed to overcome gradient obfuscation. Results are presented in Table 3 for ResNet and Table 6 in Section E for ViT.

The Hyperprior model illustrates classic gradient masking behavior. It performs well against attacks relying on gradients computed through the defense, but fails under black-box attacks or gray-box attacks with techniques like U-Net BPDA or ST BPDA. In those settings, its robust accuracy collapses.

In contrast, CRDR HR shows strong robustness across attacks. WB PGD, using true gradients, is the most effective or close second, indicating CRDR isn't just masking gradients. However, at high perturbations ($\epsilon = \frac{16}{255}$), U-Net BPDA sometimes outperforms WB PGD, suggesting that out-of-distribution inputs reduce true gradient effectiveness and increase gradient masking. In such cases, surrogate-based gradients like U-Net BPDA yield stronger attacks.

Table 4: Robust accuracy under PGD 4/255 attack running for ten iterations for diffusion-based defenses and the CRDR compression-based defense. The robust accuracy numbers of the diffusion-based models are taken from Lee & Kim (2023).

DEFENSE METHOD	STANDARD	PGD 4/255
ENGSTROM ET AL. (2019)	62.42	33.20
WONG ET AL. (2020)	53.83	28.04
NIE ET AL. (2022)	75.48	38.71
LEE & KIM (2023)	66.21	42.15
CRDR	61.72	35.88

4.4 Comparison to Diffusion-Based Defenses

As noted earlier, diffusion models can purify adversarial noise; however, they are computationally expensive. For example, Lee & Kim (2023) takes over 60 minutes to process 100 images on an RTX 3090, while CRDR needs only 1.5 seconds. With the ResNet classifier taking 0.5 seconds, CRDR increases inference time 4×, compared to over 7200× for the diffusion method. Nevertheless, we compare CRDR HR to diffusion models due to their high realism. As shown in Table 4, VAE-based models like CRDR HR offer comparable robustness at much lower cost, making them a more practical choice for attack evaluations while maintaining strong defensive performance.

We also conduct our own test to assess the robustness of the diffusion models. Due to space, the results are in the Appendix, but show that a PGD 8/255 attack can get an ASR around 72%, giving a model accuracy around 20% for the diffusion model by Lee & Kim (2023).

4.5 Highlight of Extra Results in the Appendix

Due to space constraints, additional results are provided in the appendix for interested readers.

The section titled "Attacking Adversarial Purification" demonstrates that while it is computationally expensive, diffusion models can be attacked.

Instead of evaluating against standard classifiers, one can use pretrained robust classifiers to assess whether realism offers additional benefits in already robust settings. We ran the compression-based defense on the top 11 models from RobustBench (Croce et al., 2021; Singh et al., 2023; Bai et al., 2024; Amini et al., 2024; Xu et al., 2025; Liu et al., 2025). Our RobustBench results can be seen in Table 11 in Section H.3. The results show that applying CRDR with high quality and realism does not improve performance; it consistently reduces accuracy across all tested robust models.

Iterative compression has shown promising results (Räber et al., 2025), but its robustness largely stems from gradient masking. Transferring attacks from defenses with fewer iterations drops the accuracy significantly, revealing the vulnerability once gradient masking is circumvented (cf. Section H.4).

Interestingly, under a white-box PGD attack, the structure of the adversarial perturbation tends to follow the image structure, see Section H.5. In particular, the perturbation primarily affects object edges. Realism amplifies the model's tendency to hallucinate details. These hallucinated details may change when attacking such models, but the resulting images still appear realistic to the human eye.

5 Discussion and Future Work

In this work, we systematically evaluated attacks against compression-based adversarial defenses, identifying what makes specific compression models difficult to break. Our findings reveal a critical challenge for attackers: high realism in reconstructed images significantly increases attack difficulty. While most compression-based defenses fail under adaptive attacks, models with high realism consistently demonstrate greater resistance.

Through rigorous experimentation, we have shown that realism, not gradient masking or other obfuscation techniques, is the primary obstacle for attackers. When compression models produce realistic reconstructions, they maintain distributional alignment with natural images while discarding adversarial perturbations, creating a fundamental asymmetry that favors the defender. Attackers must find perturbations that survive the compression process and remain effective after high-quality reconstruction, a significantly more challenging task.

Future work should focus on developing more effective attacks against high-realism compression models. This includes designing loss functions that better capture realistic reconstruction quality and crafting attacks that target preserved semantic features rather than merely pixel-level noise. Efficiently attacking diffusion-based compression methods, which naturally achieve high realism, also remains a key challenge for thorough security evaluation. The concept of *perfect realism*, where reconstructions are indistinguishable from natural data and lie on the data manifold, offers an ideal defense by projecting inputs onto the natural image distribution and eliminating adversarial perturbations. Should attacks on these high-realism models, especially diffusion-based approaches, continue to fail, our findings would mark a promising step toward genuinely robust defenses.

Reproducibility Statement

All code used in our experiments is included in the supplementary material, together with a README file that explains how to set up the environment, run the training and evaluation scripts, and reproduce the reported results. The training and test data are publicly available through Huggingface. For the camera-ready version, we will make the code publicly available. In addition, we provide detailed descriptions of the model architectures, training procedures, and datasets in Section 3 to further support reproducibility.

References

Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22324–22333, 2023.

Sajjad Amini, Mohammadreza Teymoorianfard, Shiqing Ma, and Amir Houmansadr. MeanSparse: Post-Training Robustness Enhancement Through Mean-Centered Feature Sparsification, October 2024.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, July 2018a.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018b. URL `https://arxiv.org/abs/1707.07397`.

Yatong Bai, Mo Zhou, Vishal M. Patel, and Somayeh Sojoudi. MixedNUTS: Training-Free Accuracy-Robustness Balance via Nonlinearly Mixed Classifiers. *Transactions on Machine Learning Research*, May 2024. ISSN 2835-8856.

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

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.

Jona Ballé, Luca Versari, Emilien Dupont, Hyunjik Kim, and Matthias Bauer. Good, cheap, and fast: Overfitted image compression with wasserstein distortion. *arXiv preprint arXiv:2412.00505*, 2024.

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*, pp. 387–402. Springer, 2013.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models, February 2018.

Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2023.

Nicholas Carlini and David Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods, November 2017a.

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks, March 2017b.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On Evaluating Adversarial Robustness, February 2019.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, November 2017. doi: 10.1145/3128572.3140448.

Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. AdvDiffuser: Natural Adversarial Example Synthesis with Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4539–4549, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00421.

Yuefeng Chen, Xiaofeng Mao, Yuan He, Hui Xue, Chao Li, Yinpeng Dong, Qi-An Fu, Xiao Yang, Wenzhao Xiang, Tianyu Pang, Hang Su, Jun Zhu, Fangcheng Liu, Chao Zhang, Hongyang Zhang, Yichi Zhang, Shilong Liu, Chang Liu, Wenzhao Xiang, Yajie Wang, Huipeng Zhou, Haoran Lyu, Yidan Xu, Zixuan Xu, Taoyu Zhu, Wenjun Li, Xianfeng Gao, Guoqiu Wang, Huanqian Yan, Ying Guo, Chaoning Zhang, Zheng Fang, Yang Wang, Bingyang Fu, Yunfei Zheng, Yekui Wang, Haorong Luo, and Zhen Yang. Unrestricted Adversarial Attacks on ImageNet Competition, October 2021.

Miguel Costa and Sandro Pinto. David and Goliath: An Empirical Evaluation of Attacks and Defenses for QNNs at the Deep Edge. In *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, pp. 524–541, July 2024. doi: 10.1109/EuroSP60621.2024.00035.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, August 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark, October 2021.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression, May 2017.

Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 321–338, 2019. ISBN 978-1-939133-06-9.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, November 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, October 2020.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images, August 2016.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL `https://github.com/MadryLab/robustness`.

Claudio Ferrari, Federico Becattini, Leonardo Galteri, and Alberto Del Bimbo. (Compress and Restore)^{N} : A Robust Defense Against Adversarial Attacks on Image Classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(1s):1–16, February 2023. ISSN 1551-6857, 1551-6865. doi: 10.1145/3524619.

Stanislav Fort and Balaji Lakshminarayanan. Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness, August 2024.

Iuri Frosio and Jan Kautz. The Best Defense is a Good Offense: Adversarial Augmentation Against Adversarial Attacks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4067–4076, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.00396.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, March 2015.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering Adversarial Images using Input Transformations, October 2017.

Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5718–5727, 2022a.

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*, pp. 1764–1769, 2022b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE Computer Society, June 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, September 2018.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis. High-fidelity image compression with score-based generative models. *arXiv preprint arXiv:2305.18231*, 2023.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Shoma Iwai, Tomo Miyazaki, and Shinichiro Omachi. Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2900–2909, 2024.

Yahya Jabary, Andreas Plesner, Turlan Kuzhagaliyev, and Roger Wattenhofer. Seeing Through the Mask: Rethinking Adversarial Examples for CAPTCHAs, September 2024.

Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. ComDefend: An Efficient Image Compression Model to Defend Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6084–6092, 2019.

Hyunjik Kim, Matthias Bauer, Lucas Theis, Jonathan Richard Schwarz, and Emilien Dupont. C3: High-performance and low-complexity neural compression from a single image or video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9347–9358, 2024.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, February 2017.

Théo Ladune, Pierrick Philippe, Félix Henry, Gordon Clare, and Thomas Leguay. Cool-chic: Coordinate-based low complexity hierarchical image codec. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13515–13522, 2023.

Minjong Lee and Dongwoo Kim. Robust Evaluation of Diffusion-Based Adversarial Purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 134–144, 2023.

Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant Adversarial Purification with Adversarial Consistency Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24331–24340, 2025.

Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A Comprehensive Study on Robustness of Image Classification Models: Benchmarking and Rethinking. *International Journal of Computer Vision*, 133(2): 567–589, February 2025. ISSN 1573-1405. doi: 10.1007/s11263-024-02196-3.

Zhuang Liu and Kaiming He. A Decade's Battle on Dataset Bias: Are We There Yet? In *The Thirteenth International Conference on Learning Representations*, October 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, September 2019.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems*, 33:11913–11924, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal Adversarial Perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 86–94, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.17.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion Models for Adversarial Purification. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16805–16827. PMLR, June 2022.

Tianyu Pang, Kun Xu, and Jun Zhu. Mixup Inference: Better Exploiting Mixup to Defend Adversarial Attacks. In *Eighth International Conference on Learning Representations*, April 2020a.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of Tricks for Adversarial Training. In *International Conference on Learning Representations*, October 2020b.

Samuel Räber, Andreas Plesner, Till Aczel, and Roger Wattenhofer. Human aligned compression for robust models. *arXiv preprint arXiv:2504.12255*, 2025.

Javier Rando, Jie Zhang, Nicholas Carlini, and Florian Tramèr. Adversarial ML Problems Are Getting Harder to Solve and to Evaluate, February 2025.

Lucas Relic, Roberto Azevedo, Markus Gross, and Christopher Schroers. Lossy image compression with foundation diffusion models. In *European Conference on Computer Vision*, pp. 303–319. Springer, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

Ryan Sheatsley, Blaine Hoak, Eric Pauley, and Patrick McDaniel. The space of adversarial strategies. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 3745–3761, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3.

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

Naman D. Singh, Francesco Croce, and Matthias Hein. Revisiting Adversarial Training for ImageNet: Architectures, Training and Generalization across Threat Models, October 2023.

Wei Song, Cong Cong, Haonan Zhong, and Jingling Xue. Correction-based Defense Against Adversarial Video Attacks via {Discretization-Enhanced} Video Compressive Sensing. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 3603–3620, 2024. ISBN 978-1-939133-44-1.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014.

Lucas Theis. What makes an image realistic? *arXiv preprint arXiv:2403.04493*, 2024.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528, June 2011. doi: 10.1109/CVPR.2011.5995347.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On Adaptive Attacks to Adversarial Example Defenses. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1633–1645. Curran Associates, Inc., 2020.

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

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training, 2020. URL `https://arxiv.org/abs/2001.03994`.

Shangxi Wu, Jitao Sang, Kaiyuan Xu, Jiaming Zhang, and Jian Yu. Attention, Please! Adversarial Defense via Activation Rectification and Preservation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(4):142:1–142:18, February 2023. ISSN 1551-6857. doi: 10.1145/3572843.

Song Xia, Wenhan Yang, Yi Yu, Xun Lin, Henghui Ding, Lingyu Duan, and Xudong Jiang. Transferable Adversarial Attacks on SAM and Its Downstream Models. *Advances in Neural Information Processing Systems*, 37:87545–87568, December 2024.

Tongda Xu, Ziran Zhu, Dailan He, Yanghao Li, Lina Guo, Yuanyuan Wang, Zhe Wang, Hongwei Qin, Yan Wang, Jingjing Liu, et al. Idempotence and perceptual image compression. *arXiv preprint arXiv:2401.08920*, 2024.

Xiaoyun Xu, Shujian Yu, Zhuoran Liu, and Stjepan Picek. MIMIR: Masked Image Modeling for Mutual Information-based Adversarial Robustness, April 2025.

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36:64971–64995, 2023a.

Ruihan Yang and Stephan Mandt. Lossy Image Compression with Conditional Diffusion Models. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023b.

Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A Survey on Universal Adversarial Attack. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 4687–4694, Montreal, Canada, August 2021. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-9-6. doi: 10.24963/ijcai.2021/635.

Jie Zhang, Christian Schlarmann, Kristina Nikolić, Nicholas Carlini, Francesco Croce, Matthias Hein, and Florian Tramèr. Evaluating the Robustness of the "Ensemble Everything Everywhere" Defense, February 2025.

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*, pp. 586–595, 2018.

Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable Adversarial Perturbations. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, volume 11218, pp. 471–486. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01263-2 978-3-030-01264-9. doi: 10.1007/978-3-030-01264-9_28.

Geigh Zollicoffer, Minh N. Vu, Ben Nebgen, Juan Castorena, Boian Alexandrov, and Manish Bhattarai. LoRID: Low-Rank Iterative Diffusion for Adversarial Purification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):23081–23089, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i21.34472.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, December 2023.

A Usage of LLMs

In the preparation of this paper, we made use of large language models (LLMs) as supportive tools. ChatGPT, Claude, Gemini, and Grammarly were employed for spellchecking, refining wording, and condensing text to improve clarity and readability. Furthermore, ChatGPT, Claude, and Cursor were used to analyze and explain code, assist with code completion, and generate visualizations that facilitated our development process. These tools served as auxiliary aids for writing and implementation, while all core research ideas, experimental design, and interpretation of results are our own.

B Limitations

The primary limitation of our study is the exclusion of diffusion-based compression models, known to achieve the highest levels of realism. This omission was due to computational constraints but represents an important direction for future research. Additionally, we did not run experiments across multiple random seeds to quantify variance. Although the inherent stochasticity of PGD introduces some variability, incorporating multiple seeds would significantly increase computational costs.

C Extended Related Work and Background

C.1 Adversarial Robustness

Shortly after the success of AlexNet (Krizhevsky et al., 2012), it was found that neural networks are very susceptible to *adversarial attacks* (Szegedy et al., 2014; Goodfellow et al., 2015). Here, an adversary makes, usually small and imperceptible, modifications (perturbations) to, for instance, an image such that a model mislabels it.

Attacks Many attacks have been developed over the years with various benefits and drawbacks. Some of the most noteworthy are FGSM (Fast Gradient Sign Method) (Goodfellow et al., 2015), iterated FGSM or iFGSM (Kurakin et al., 2017), CW (Carlini & Wagner) (Carlini & Wagner, 2017b), and PGD (Projected Gradient Decent) (Madry et al., 2019). Additional attacks that have often been used are APGD (Croce & Hein, 2020), DeepFool (Moosavi-Dezfooli et al., 2016), and ZOO (Chen et al., 2017). We refer the reader to the original papers for specific details, but we include the necessary details to understand this paper in Section D. The main thing to know is that PGD has a perturbation budget ϵ and number of iteration it can perform n. PGD then does projected gradient decent for n iterations to find an adversarial example that is at most ϵ distance in l_∞ from the original image.

The above attacks fall into three broad categories: FGSM, iFGSM, PGD, and APGD are l_∞-bounded attacks, CW and DeepFool are l_2-bounded attacks, and ZOO is a gradient-free attack using the predicted confidence scores for each class. PGD can also be implemented as an l_2-bounded attack.

When the defense includes randomness, a common attack augmentation is EoT (Expectation over Transformation), in which gradients are averaged over multiple forward and backward passes (Athalye et al., 2018a; Zhang et al., 2025). If the defense causes the gradients to be infeasible to compute, Athalye et al. (2018a) suggested Backward Pass Differentiable Approximation (BPDA), where the defense is used during the forward pass, but a differentiable approximation is used during the backward pass. The simplest is to use the identity function, while more advanced methods might train a differentiable surrogate model g' that emulates the defense g; i.e., $g(x) \approx g'(x)$.

Moosavi-Dezfooli et al. (2017) show in their work "Universal adversarial perturbations" that adversarial perturbations transfer to other data samples and models; Szegedy et al. (2014) already hinted at this years before. Several later studies have explored this (Zhou et al., 2018; Zhang et al., 2021; Xia et al., 2024), attempted to explain the effect (Demontis et al., 2019), and shown the effect also exists for language models (Zou et al., 2023). The transfer effect allows for another frequently used technique of black box attacks, where perturbations are generated for one model and applied to a black box model (Carlini et al., 2019). Another branch of black box attacks looks only at the prediction labels of a model to generate attacks (Brendel et al., 2018).

However, as highlighted by Sheatsley et al. (2023), these are only a small set of all the attacks that could be considered. The list of possible attacks is heavily dependent on the threat model used (Carlini et al., 2019; Sheatsley et al., 2023), and if one relaxes the often-used imperceptible assumption, then there exists a wide range of semantic-preserving attacks (Chen et al., 2021; 2023; Jabary et al., 2024), and natural adversarial examples (Hendrycks & Dietterich, 2018; Hendrycks et al., 2021).

Lastly, often the best attack results are found using *adaptive attacks* (Carlini et al., 2019; Tramer et al., 2020; Sheatsley et al., 2023; Zhang et al., 2025). Here, the attacks, for instance, the optimization objective, are adjusted to the defense. It is thus vital to consider adaptive attacks when a defense is resistant to standard gradient-based attacks. This work focuses on adaptive attacks with PGD as the underlying optimization method.

Defenses Given the prevalence of adversarial attacks, many researchers have explored how to defend against them. We broadly view this in three groups: Architecture improvements (Singh et al., 2023; Wu et al., 2023; Fort & Lakshminarayanan, 2024), adversarial training (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2019; Pang et al., 2020b; Fort & Lakshminarayanan, 2024), and adversarial purification (Nie et al., 2022; Frosio & Kautz, 2023; Zollicoffer et al., 2025; Räber et al., 2025).

The architecture branch focuses on making the models more robust by design, for instance, by taking inspiration from biology to mimic the human eye when training a ResNet model (Fort & Lakshminarayanan, 2024).

Adversarial training, as formalized by Madry et al. (2019), is perhaps the most straightforward method and consists of showing the model adversarial examples during training to ensure it classifies these correctly. This method has yielded positive results (Szegedy et al., 2014; Madry et al., 2019), but models with adversarial training can be broken (Zhang et al., 2025). They also experience a drop in clean accuracy (accuracy on images without adversarial perturbations) (Madry et al., 2019; Carlini et al., 2019), and the robustness may not carry over to other attacks (Zhang et al., 2025).

The idea of adversarial purification is to remove the adversarial noise in images before passing the cleaned images to pretrained classification models. It is motivated by the work of Ilyas et al. (2019), hinting that adversarial examples perturb brittle features in the model. These methods *should work independently* of any robustness applied to the classifier through adversarial training or robust architectures. Two noteworthy works using diffusion-based models to remove the adversarial noise are Nie et al. (2022); Lee & Kim (2023).

In the realm of adversarial purification, diffusion models have been proposed as a way to remove adversarial noise from input images (Nie et al., 2022; Lee & Kim, 2023). These approaches are conceptually similar to compression-based defenses with realism: both aim to project adversarial examples back onto the manifold of natural images to restore classifier performance. However,

prior work on diffusion-based purification has not explicitly investigated the role of realism as a contributing factor to robustness.

One major drawback of diffusion-based defenses is their computational cost. Diffusion models are typically an order of magnitude more expensive than VAE- or GAN-based learned compression models, making them impractical for many real-world settings (Dhariwal & Nichol, 2021; Yang & Mandt, 2023b). This high cost limits their deployment and complicates the evaluation of robustness: mounting effective adversarial attacks against diffusion models becomes significantly harder due to the computational overhead. However, simply making gradients difficult to compute does not equate to genuine robustness. A well-designed adaptive attacker, given sufficient resources, should still be able to circumvent such defenses (Carlini et al., 2019; Lee & Kim, 2023).

Notably, the purification method proposed by Nie et al. (2022) was later defeated by Lee & Kim (2023). However, the latter only evaluated their improved defense under the attack that broke the former. They did not develop new adaptive attacks tailored to their defense—a strategy that past research suggests would likely reduce the effectiveness of the defense. The history of adversarial robustness research shows that defenses that are not tested under strong, tailored attacks often overstate their robustness (Athalye et al., 2018a; Carlini et al., 2019; Tramer et al., 2020).

RobustBench The considerable interest in adversarial examples has given rise to benchmarks and leaderboards, one of which is RobustBench (Croce et al., 2021). This provides a leaderboard of models that are supposed to be robust to adversarial examples. However, note that Fort & Lakshminarayanan (2024) claimed their model beat the RobustBench leaderboard, but, using adaptive attacks, Zhang et al. (2025) showed that the model is still very vulnerable to adversarial examples.

We will use RobustBench as a source of robust models, and thus enable us to test if compression-based purification would help as claimed in (Räber et al., 2025). For this, we take top-performing models from the leaderboard with open-sourced model weights and test whether the compression-based purification defenses improve the models' robustness.

C.2 Realism

Image compression algorithms are traditionally evaluated using *distortion metrics* such as PSNR or SSIM, which measure the distance between a restored image $\hat{x}$ and its reference x. Formally, distortion is defined as:

$$\mathcal{D} := \mathbb{E}_{(x,\hat{x})\sim(p_X,p_{\hat{X}})}[\Delta(x,\hat{x})], \tag{4}$$

where $\Delta(\cdot,\cdot)$ is a pointwise distortion measure (e.g., ℓ_2 distance), and p_X, $p_{\hat{X}}$ denote the distributions of ground truth and reconstructed images. Distortion is considered a *full-reference* metric, requiring access to the original image x to compare it to the reconstructed version $\hat{x}$.

However, metrics such as PSNR, MS-SSIM (Wang et al., 2003), or LPIPS (Zhang et al., 2018) correlate poorly with human perception, and directly quantifying perceptual distortion remains a challenging problem. Instead of a perceptual distortion metric, one can also measure both distortion and realism and optimize the compression model for both. Realism can be formally defined as:

$$\mathcal{R} := -d(p_{\hat{X}}, p_X), \tag{5}$$

where $d(\cdot,\cdot)$ is a divergence measure, such as Kullback-Leibler. Unlike distortion, realism is a *no-reference metric*, requiring only the generated image distribution to match that of natural images. Although measuring realism remains challenging (Theis, 2024), a widely used proxy is the Fréchet Inception Distance (FID) (Heusel et al., 2017), which compares the distributions extracted by a pretrained Inception network. FID has gained popularity for capturing both fidelity and diversity in generated samples.

Compression models are typically trained with the following loss function:

$$\mathcal{L} = \mathcal{L}_{\text{RATE}} + \lambda\mathcal{D} - \beta\mathcal{R}, \tag{6}$$

where $\mathcal{L}_{\text{RATE}}$ represents the estimated rate, or in other words, the number of bits required to represent the image after it has been compressed, λ controls the level of distortion, and β the level of realism. Since the information is not transmitted through an explicit information bottleneck in our approach, the rate term $\mathcal{L}_{\text{RATE}}$ does not play a critical role in this context.

Blau & Michaeli (2018) prove that there exists an inherent tradeoff between distortion $\mathcal{D}$ and realism $\mathcal{R}$. Specifically, reducing one inevitably increases the other, regardless of the choice of distortion or divergence. This theoretical result underpins the empirical observation that GAN-based methods, which maximize $\mathcal{R}$ through adversarial training, tend to increase $\mathcal{D}$ while producing perceptually convincing outputs.

Our work builds on this foundation, exploring the intersection of realistic compression and adversarial robustness. We extend the existing literature by providing experimental evidence that realism, rather than distortion minimization, makes image-compression models (partially) robust against adversarial examples.

C.3 Compression Models

End-to-End Optimized Image Compression (Ballé et al., 2016) jointly learns analysis and synthesis transforms along with an entropy model over quantized latents, directly optimizing rate–distortion performance in a VAE-style framework. The Hyperprior model (Ballé et al., 2018) extends this approach by introducing a secondary network that predicts spatially adaptive Gaussian scales, better capturing local image statistics. HiFiC (Mentzer et al., 2020) enhances learned compression with GANs to produce realistic reconstructions. By refining network design and training with perceptual losses, it generates outputs that closely resemble natural images.

ELIC (He et al., 2022a) introduced uneven channel grouping and deeper autoregressive context networks for more accurate entropy modeling, yielding faster convergence and lower bitrates. PO-ELIC (He et al., 2022b) then augmented this foundation with adversarial fine-tuning and perceptual losses to enrich texture realism at low bitrates.

More recent work has shifted toward models that offer explicit, user-controllable trade-offs between rate, distortion, and realism within a single network. MRIC (Agustsson et al., 2023) introduced a conditional generator that, at a fixed bit rate, lets users interpolate between low-distortion and high-realism reconstructions by tuning a realism flag β, thereby explicitly navigating the distortion–realism trade-off within a single model. CRDR (Iwai et al., 2024) built on this by adding a discrete quality-level input q alongside β, and embedding interpolation channel attention layers in both encoder and generator to yield true variable-rate compression, allowing joint control over bitrate, distortion, and realism with one network.

Diffusion models are great at generating images, which is also taken advantage of in image compression. Models (Yang & Mandt, 2023a; Careil et al., 2023; Hoogeboom et al., 2023; Xu et al., 2024; Relic et al., 2024) that focus on extra low bitrate or extra high realism without worrying about compute use diffusion. These methods are orders of magnitude more expensive than VAE-based methods, but achieve great realism.

Reducing computational complexity has also been a focus in compression with the line of work like Cool-Chic (Ladune et al., 2023) and C3 (Kim et al., 2024). These methods overfit a latent, auto-regressive latent model and decoder to a single image, and send all three over the channel. High realism can be achieved at this level of complexity (Ballé et al., 2024), but as these models are INRs, computing a gradient through the compression model is not straightforward, and encoding is computationally expensive.

C.4 Compression as adversarial defense

Using compression as an adversarial defense is not new. It has been explored for a decade (Dziugaite et al., 2016; Das et al., 2017; Jia et al., 2019; Ferrari et al., 2023) where some authors also explored using iterated compression and decompression cycles (Ferrari et al., 2023; Räber et al., 2025). And more recently, using video compression to defend video classifiers (Song et al., 2024) and model quantization to defend general models (Costa & Pinto, 2024).

Two key works in this area are by Guo et al. (2017) and Shin & Song (2017); the former argued that JPEG compression as a preprocessing step is a very effective adversarial defense, while the latter showed that, by making JPEG differentiable, this defense could be bypassed entirely–highlighting the necessity of properly evaluating a defense.

D Details on Adversarial Attacks

PGD is a very strong attack, and most defenses in a white-box setting can be broken by it when using the right objective.[1]

For the rest of this section, let f be a neural network, L a loss function, x an input with corresponding output y, and $\nabla_x f(x)$ the gradient of $L(f(x), y)$ with respect to x.

FGSM computes the gradients of the loss function with respect to the image's pixel values, takes the sign of the gradients, and multiplies by ϵ, where ϵ is a small number, usually $\frac{4}{255}$ for ImageNet images and $\frac{8}{255}$ for CIFAR-10 images. The attack can be written as:

$$FGSM(f, x) = x + \epsilon \cdot \text{sign}(\nabla_x f(x)).$$

iFGSM uses two extra parameters $\alpha < \epsilon$ and n. iFGSM then takes n FGSM steps, by default $n = 10$, of size α ensuring the final perturbation is within the l_∞ ball of size ϵ.

$$\begin{aligned} iFGSM(f, x) &= x_n \\ x_0 &= x \\ x_{i+1} &= Clip_{x,\epsilon}\left(x_i + \epsilon \cdot \text{sign}(\nabla_{x_i} f(x_i))\right) \end{aligned}$$

where $Clip_{x,\epsilon}$ clips x_i to be within the l_∞ ball of radius ϵ of x.

PGD works almost entirely like iFGSM; however, there are two key changes. 1) It randomly initializes x_0 in the ϵ l_∞-ball. 2) With the stochasticity from 1), it includes the option for restarts, and thus running the optimization again.

E Extended adaptive attacks

We show in Table 5 complete results for the adaptive attacks applied to the ResNet model.

We show in Table 6 results for the adaptive attacks applied to the ViT model.

F Hyperparameters

This section includes information about the hyperparameters used during our experiments. Unless stated otherwise (e.g, number of steps or epsilon as columns/rows of the table), we use the following in Tables 7 and 8.

For the Adaptive Realism attack, the realism parameter was chosen amongst the beta values specified in Table 8. The results of this experiment can be found in Table 9.

We use the 50000 images from the ImageNet validation set for most of our results. Exceptions are the U-Net BPDA in Table 3 and Table 6, the results in table Table 2 and the results of the ARA ablation in Table 9

G Training U-Nets for BPDA

The U-Net used to approximate the compression and decompression step in the BPDA attack follows a standard U-Net architecture (Ronneberger et al., 2015). We use two downsampling layers, which transform the input images from a resolution of (3,224,224) to an intermediate representation of (256,56,56). For each experiment, we train the U-Net on the whole dataset used in the experiment. The U-Net is trained for 20 epochs using the L1-loss between the input image (original image with or without adversarial noise) and the target image (reconstruction of the input image) using an Adam optimizer with a learning rate of 0.001 and a learning rate schedule. The learning rate scheduler is StepLR from PyTorch, configured with a step size of 5 and a gamma of 0.1. During an epoch, each batch is used 4 times, 3 passes with additional noise added to the images to simulate adversarial noise, and one clean pass with the original images.

[1] Personal communication with Nicholas Carlini with further evidence in (Zhang et al., 2025).

Table 5: Extended version of Table 3 with the 16/255 results. Results for ResNet50. "—" denotes values not implemented for MRIC or evaluations invalid without realism control. Models with strong gradient masking, like Hyperprior, are vulnerable to adaptive attacks. Realism does not increase gradient obfuscation, so both high-realism models retain accuracy with minor drops under adaptive attacks.

Strength	Defense	Standard	BB PGD	WB PGD	ST BPDA	U-Net BPDA	ACM	ARA
4/255	Hyperprior	78.73	48.76	78.84	10.94	**0.98**	78.82	—
	HiFiC	61.53	59.52	**11.83**	44.65	24.04	59.20	—
	MRIC LR	52.80	51.04	**26.68**	—	—	—	**26.68**
	MRIC HR	63.06	59.28	**39.00**	—	—	—	39.60
	CRDR LR	46.02	44.92	**16.30**	39.36	28.96	41.28	**16.30**
	CRDR HR	61.72	59.80	35.88	56.36	47.62	55.67	**34.50**
	JPEG	70.30	65.68	8.72	18.70	**5.19**	68.67	—
	ELIC	60.03	58.79	**16.43**	40.00	17.98	54.48	—
8/255	Hyperprior	78.73	27.16	78.72	3.18	**0.02**	78.67	—
	HiFiC	61.53	56.84	**6.65**	32.19	7.98	55.37	—
	MRIC LR	52.80	49.44	**12.20**	—	—	—	**12.20**
	MRIC HR	63.06	57.38	21.98	—	—	—	**21.10**
	CRDR LR	46.02	44.14	**8.60**	33.58	12.12	36.96	**8.60**
	CRDR HR	61.72	57.36	23.92	50.77	26.64	49.35	**21.18**
	JPEG	70.30	60.76	5.19	10.01	**0.24**	64.75	—
	ELIC	60.03	57.59	9.01	27.24	**4.98**	51.63	—
16/255	Hyperprior	78.73	6.99	76.94	1.96	**0.00**	76.82	—
	HiFiC	61.53	51.67	3.08	25.53	**1.80**	45.17	—
	MRIC LR	52.80	49.40	**2.90**	—	—	—	**2.90**
	MRIC HR	63.06	52.90	7.26	—	—	—	**6.40**
	CRDR LR	46.02	41.40	**4.10**	30.38	2.96	27.82	**4.10**
	CRDR HR	61.72	52.70	13.93	47.10	**7.98**	36.68	11.42
	JPEG	70.30	49.48	2.86	6.58	**0.01**	52.99	—
	ELIC	60.03	54.68	4.10	20.62	**0.34**	36.47	—

H Extended results

H.1 Attacking Adversarial Purification

We ran a limited attack on a diffusion-based adversarial purification defense (Lee & Kim, 2023). Since running (and attacking) this defense is more computationally expensive compared to the compression models used in the other results shown in this paper, we focused on a smaller part of the ImageNet dataset.

We ran a U-Net BPDA attack on 100 images. Table 10 shows the accuracy of an attack with epsilon 8/255. The attacked accuracy of 69% shows that the U-Net attack was not able to produce practical gradients, as the accuracy is higher than the base accuracy. The attack proposed by (Lee & Kim, 2023) showed an accuracy of 42.15% with an even lower epsilon of 4/255. In addition, we apply a very aggressive attack with an epsilon of 64/255. However, Table 10 shows that the results are comparable to adding random noise to the image, showing that the gradients carry no usable signal.

We assume that the U-net failed to capture the large variance shown by the diffusion model. Figure 4 shows the large visible variance in diffusion output images. Since the training, especially the creation of a dataset of diffusion input-output image pairs, was computationally expensive, we did no further experiments with the U-net attack–we utilized approximately 2,000 additional GPU hours for this experiment.

We also ran the PGD + EOT attack from (Lee & Kim, 2023) for epsilon 8/255 and 100 PGD iterations. The weaker defnse we attacked used just 9 diffusion steps. This allows us to compute the gradient through the entire defense. To better evaluate our results, we compute the accuracy of the full defense + classifier after every PGD iteration. We started the attack with 64 correctly classified images. Figure 5 shows the decrease in accuracy over 100 PGD iterations. The non-deterministic nature of

Table 6: Results for ViT

Strength	Defense	Standard	BB PGD	WB PGD	U-Net BPDA	ST BPDA	ACM	ARA
4/255	Hyperprior	80.38	7.47	80.51	1.72	**0.20**	80.46	
	HiFiC	70.76	62.12	**10.06**	39.78	22.56	69.32	
	CRDR LR	52.24	49.24	**14.44**	24.82	33.38	48.64	**14.44**
	CRDR HR	68.82	62.30	28.28	53.22	41.76	63.90	**27.28**
	JPEG	74.80	36.86	2.70	5.54	**0.88**	73.92	
	ELIC	68.24	60.44	16.21	27.36	**14.12**	63.16	
8/255	Hyperprior	80.38	0.61	80.25	0.16	**0.00**	80.38	
	HiFiC	70.76	56.30	**2.64**	24.68	4.00	67.12	
	CRDR LR	52.24	46.14	**6.98**	13.10	21.12	43.16	**6.98**
	CRDR HR	68.82	57.43	16.68	39.60	14.90	60.44	**13.86**
	JPEG	74.80	15.34	0.94	1.14	**0.04**	71.50	
	ELIC	68.24	54.82	6.20	15.22	**2.38**	59.99	
16/255	Hyperprior	80.38	0.01	78.65	0.12	**0.00**	78.48	
	HiFiC	70.76	46.24	0.54	14.32	**0.16**	59.56	
	CRDR LR	52.24	40.60	**2.16**	8.74	14.08	32.32	**2.16**
	CRDR HR	68.82	49.48	9.31	28.38	**2.46**	50.62	6.34
	JPEG	74.80	1.66	0.28	0.54	**0.00**	62.80	
	ELIC	68.24	43.70	1.00	8.52	**0.10**	45.94	

Table 7: Hyperparameters used for the compression models. Weights indicate the designation of the pretrained weights used, quality indicates the quality parameter for compressions with variable quality, and beta indicates the realism parameter for compressions with variable realism.

CRDR LR	quality=0	$\beta = 0$
CRDR HR	quality=0	$\beta = 5.12$
MRIC LR	weights=128	$\beta = 0$
MRIC HR	weights=128	$\beta = 2.56$
Hyperprior	quality=8	
HiFiC	weights=low	
ELIC	weights=0016	
JPEG	quality=25	

Figure 4: Three different diffusion outputs for the same input image. These showcase the large differences the diffusion models can introduce and thus what the adversarial noise must be robust to.

the diffusion model leads to a certain variance in the accuracy at higher iterations. We therefore report the average accuracy of 28.1% over iterations 50 to 100 and the minimal accuracy of 21.9%.

Table 8: Hyperparameters used for the different attacks. Epsilon is the maximum perturbation allowed for an adversarial image, and alpha controls the maximum perturbation added per step.

PGD	$\epsilon = 8/255$	$\alpha = \epsilon/4$	STEPS= 10	RANDOM START=TRUE
ST BDPA	$\epsilon = 8/255$	$\alpha = \epsilon/4$	STEPS= 10	
U-NET BDPA	$\epsilon = 8/255$	$\alpha = \epsilon/4$	TRAINING EPOCHS= 20	ATTACK STEPS = 100
ACM	$\epsilon = 8/255$	LOSS = MSE	STEPS= 20	
ARA	ATTACK=PGD	$\beta \in \{0.0, 0.16, 0.32, 0.64, 1.28, 2.56, 5.12\}$		

Table 9: Comparison of different realism values used in the Adaptive Realism Attack at $\epsilon = \frac{8}{255}$. CRDR LR uses $\beta = 0$ in the defense, CRDR HR uses $\beta = 5.12$. The bold values represent the strongest attack and were used in Tables 3 and 6.

			β						
MODEL	DEFENSE	STANDARD	0	0.16	0.32	0.64	1.28	2.56	5.12
RESNET50	CRDR LR	46.02	**8.60**	14.98	17.30	21.14	23.08	26.72	29.84
	CRDR HR	61.72	28.82	22.12	21.70	**21.18**	21.60	22.90	24.32
	MRIC LR	52.80	**12.20**	22.60	24.60	28.10	30.10	31.40	—
	MRIC HR	63.06	31.90	25.20	22.40	21.80	**21.10**	21.98	—
VIT	CRDR LR	52.24	**6.98**	11.06	12.12	14.68	16.10	19.46	24.50
	CRDR HR	68.82	19.12	14.82	**13.86**	14.20	14.64	15.76	17.02

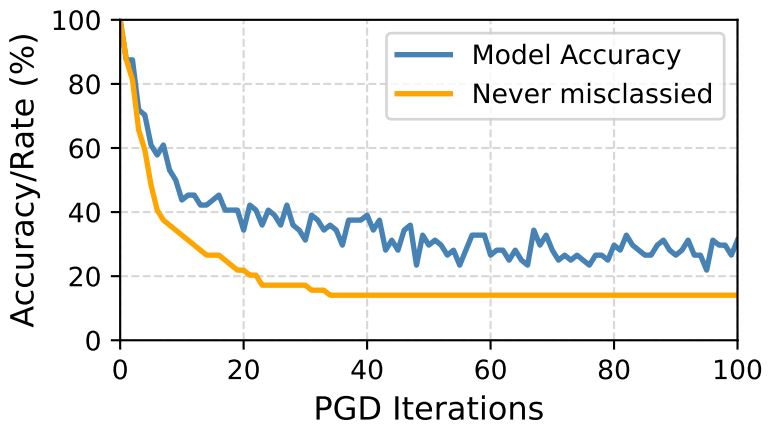

Figure 5: Accuracy of the adversarial purification defense for 100 iterations of PGD. The blue curve shows the accuracy for every iteration, and the orange curve shows the fraction of images that have never been misclassified up to that iteration.

The bottom graph in Figure 5 shows the percentage of images that have never been missclassified up to the current PGD iteration. Evaluating this metric provides an estimate of worst-case adversarial robustness, which is 14.1% after 100 iterations.

The numbers reported above are for a subset with 100% clean accuracy (i.e., they correspond to 100%−ASR). Thus, after accounting for the model accuracy of 70.7%, the average accuracy is 19.9%, placing it only slightly higher in accuracy than CRDR with high realism (cf. Table 2).

The results of this second experiment show that evaluating a diffusion-based defense not only requires significant computational effort but is also inherently more challenging due to the variance in model output, as incorrect predictions can occur for otherwise correctly classified images. This makes direct comparisons of the attack success rate (ASR) to other defenses more challenging.

Table 10: Accuracy of the U-net attack on adversarial purification. The PGD + EOT are from (Lee & Kim, 2023).

		Accuracy	
Attack	Epsilon	Clean	Attacked
U-net	8/255	68.0	69.0
U-net	64/255	63.8	12.0
Noise	64/255	61.3	27.0
PGD + EOT	4/255	70.7	42.15

For the second attack on diffusion-based purification models that was successful, we used two A100s for two days. Thus, scaling up the attacks will be difficult as this puts the processing rate at 16 images *per day* for an A100. Faster diffusion models, such as the model by Lei et al. (2025), might make larger-scale experiments feasible; however, the authors do not provide trained model weights.

H.2 Distortion vs Realism

As shown in Figure 6, robust accuracy increases monotonically with realism. This trend does not hold for distortion: there exists an optimal level of distortion that balances preserving informative content from the original image while removing adversarial perturbations. Interestingly, the optimal distortion level shifts higher as realism increases. This can be attributed to artifacts introduced by compression—excessive artifacts can act as adversarial perturbations themselves. By incorporating realism that mitigates such artifacts, stronger compression-based defenses become possible. For many models, performance gains from increased realism have not yet saturated, although these models were not originally designed to operate at higher values of the β parameter. While prior work has established that a certain distortion (or quality) level yields optimal adversarial robustness with compression, our work is the first to systematically investigate the role of realism. We demonstrate that defenses lacking realism are significantly easier to attack.

H.3 RobustBench

Instead of evaluating against standard (adversarially weak) classifiers, one can use pretrained robust classifiers to assess whether realism offers additional benefits in already robust settings. We ran the compression-based defense on the top 11 models from RobustBench (Singh et al., 2023; Bai et al., 2024; Amini et al., 2024; Xu et al., 2025; Liu et al., 2025). The results in Table 11 show that applying CRDR with high quality and high realism does not improve performance—in fact, it consistently reduces accuracy across all tested robust models. While CRDR benefits standard models by projecting inputs back onto the natural image manifold, we see two reasons for these results. 1) The information loss and subtle degradations introduced by compression can harm standard and robust accuracy when applied to already robust models, as the compression may discard the specific features these robust models have learned to utilize for classification, explaining why performance decreases rather than increases. 2) The robustified models have been overfitted to ImageNet images ((Torralba & Efros, 2011; Liu & He, 2024)) without *any* noise and images with *exactly* the type of noise PGD produces. We leave it for future work to resolve which, if any, of these explanations are correct.

H.4 Iterative Defenses

It was claimed in prior work that applying compression iteratively strengthens adversarial defenses (Räber et al., 2025). However, we show that this effect is primarily due to gradient masking rather than true robustness (cf. Table 12 in Section H.4). When attacking an iterative defense by approximating gradients using fewer defense iterations, the gradients remain informative, and accuracy can be reduced to levels comparable to using a single defense iteration.

As shown in Figure 7, the most effective defense configuration for CRDR involves multiple defense iterations with low compression quality and high realism. Interestingly, the optimal attack against this configuration uses fewer iterations to approximate the gradients. In contrast, attacks that match

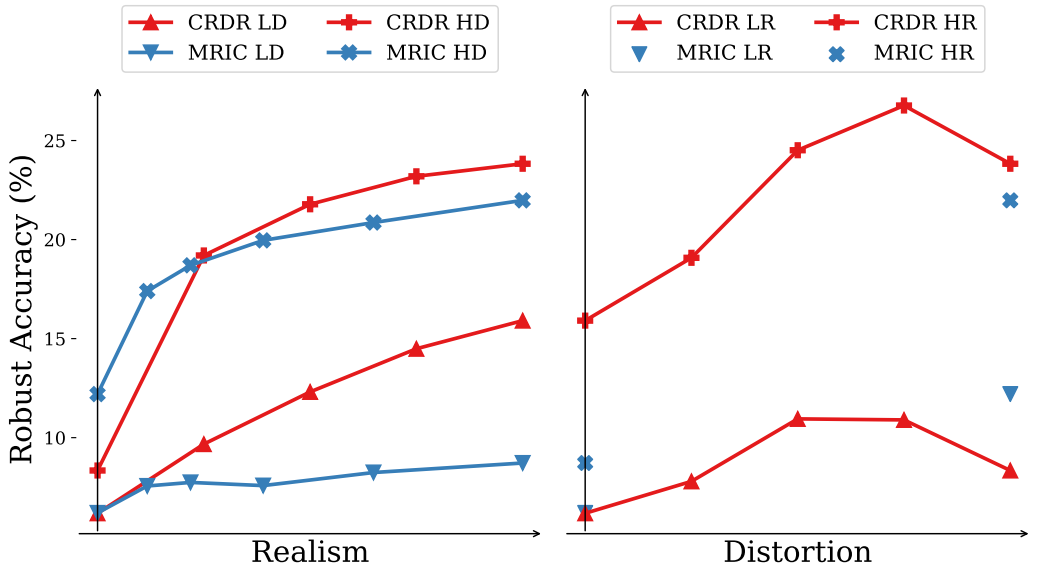

Figure 6: Impact of realism and distortion on robust accuracy for CRDR and MRIC. Realism and distortion are measured by the training parameters λ and β, respectively, and normalized to their minimum and maximum values. Note that for MRIC, pretrained weights are only available at two distortion levels. Higher realism in reconstructed images improves robustness against adversarial attacks. However, excessive distortion can degrade accuracy, suggesting the existence of an optimal distortion level that balances detail preservation and defense effectiveness.

Table 11: Accuracy (%) for RobustBench models attacked by PGD and then defended using iterative CRDR in a white-box setting. Bold highlights better performance between Base and CRDR for each pair. We take the 11 top-performing models from `https://github.com/RobustBench/robustbench` under ImageNet.

Model	Standard		4/255		8/255		16/255	
	Base	CRDR	Base	CRDR	Base	CRDR	Base	CRDR
Amini ConvNeXt-L (Amini et al., 2024)	**78.58**	76.04	**61.98**	57.08	**43.98**	36.20	**18.30**	11.84
Amini Swin-L (Amini et al., 2024)	**78.98**	77.10	**65.12**	60.18	**49.42**	41.66	**25.06**	17.28
Bai NUTS (Bai et al., 2024)	**81.48**	80.12	**70.66**	66.76	**51.10**	46.04	**14.28**	11.62
Liu ConvNeXt-B (Liu et al., 2025)	**77.16**	74.48	**58.40**	53.34	**39.44**	33.06	**18.34**	13.48
Liu ConvNeXt-L (Liu et al., 2025)	**78.62**	76.18	**60.48**	55.26	**41.44**	34.32	**19.80**	14.24
Liu Swin-B (Liu et al., 2025)	**76.78**	74.62	**59.80**	53.80	**41.08**	34.60	**19.80**	13.84
Liu Swin-L (Liu et al., 2025)	**79.00**	77.12	**61.94**	57.04	**43.04**	36.76	**22.24**	16.00
Singh ConvNeXt-B (Singh et al., 2023)	**75.94**	73.52	**58.00**	52.18	**38.84**	32.62	**16.18**	12.08
Singh ConvNeXt-L (Singh et al., 2023)	**77.66**	75.16	**60.32**	54.84	**41.86**	35.24	**19.12**	15.02
Xu Swin-B (Xu et al., 2025)	**77.26**	74.90	**58.58**	53.62	**39.18**	32.88	**16.32**	12.40
Xu Swin-L (Xu et al., 2025)	**79.40**	77.30	**62.32**	57.00	**42.20**	35.86	**19.06**	14.72

the defense in the number of iterations perform worse. This further indicates that CRDR still exhibits some gradient masking, which standard PGD attacks fail to overcome fully (cf. Section 4.3).

H.5 Structure of the adversarial noise

As shown in Figure 8, CRDR structures the adversarial noise. The image cannot be altered unstructured; the compression model ensures that the perturbation follows the image's inherent structure. When attacking CRDR with high realism, the attack can also modify the generated texture, compared to low realism.

Table 12: Accuracy (%) for iterative JPEG defenses as in Räber et al. (2025). Performing the attack (PGD with epsilon 8/255) on a defense with just one iteration defeats the iterative version. The robust accuracy seems connected to the number of iterations in the attack, not the defense.

Iterations			
Defence	Attack	Standard	PGD
50	50	69.92	67.92
50	25	69.92	65.74
50	10	69.92	58.76
50	5	69.92	43.56
50	1	69.92	7.10
10	10	69.94	58.94
10	5	69.94	43.20
10	1	69.94	7.02
1	1	71.54	5.50
50	25	69.92	65.74
25	25	69.92	65.72
10	25	69.94	65.90
5	25	69.98	66.04
1	25	71.54	68.14
0	25	80.64	78.94

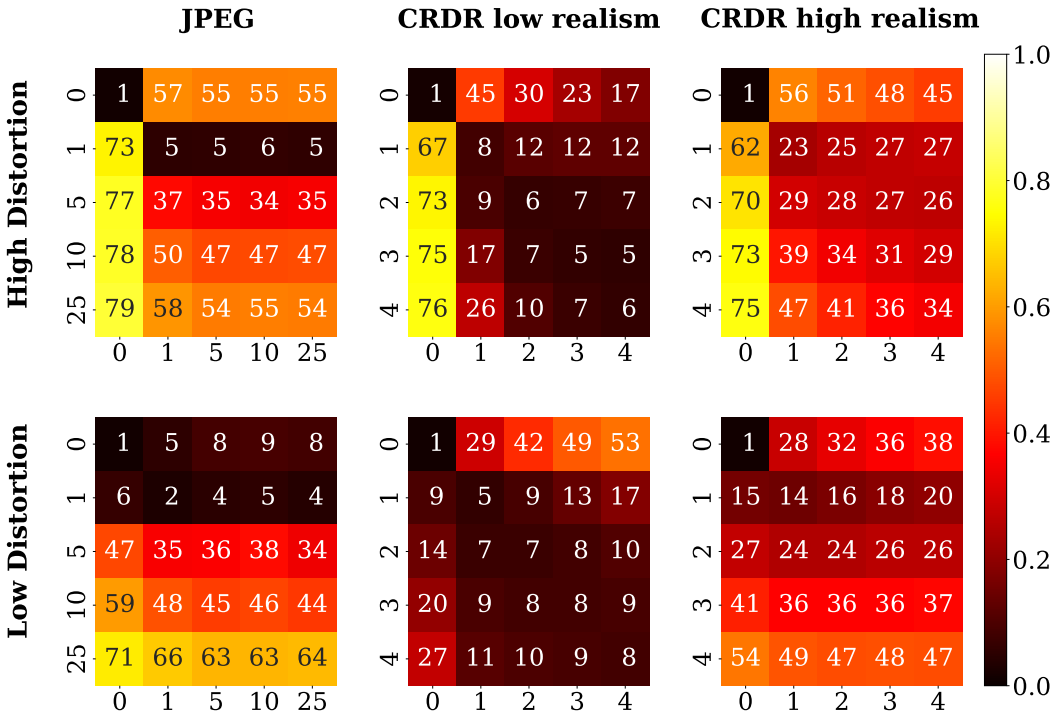

Figure 7: Robust accuracy (%) under iterative defenses and PGD attacks with ϵ 8/255. Rows indicate the number of attack iterations; columns indicate the number of defense iterations. Darker colors represent lower robust accuracy. Increasing the number of JPEG defense iterations leads primarily to gradient masking. In contrast, CRDR shows modest gains from iterative defense under weaker attacks.

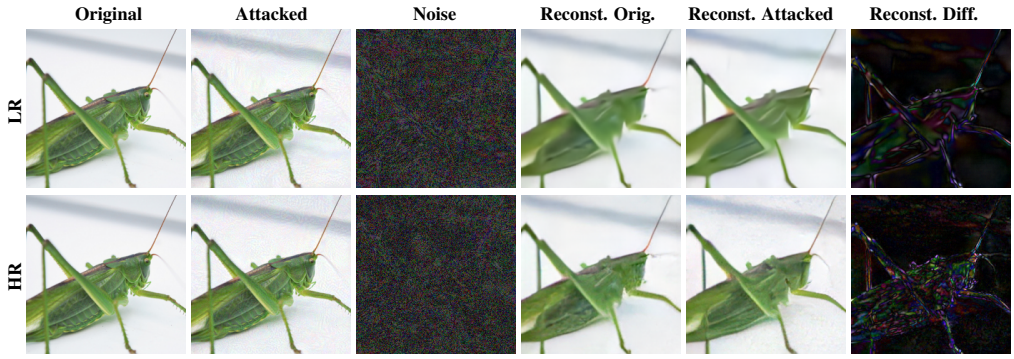

Figure 8: Comparison of original and attacked images, their differences, and reconstructions. CRDR with low realism and high realism, original, attacked, adversarial noise, reconstructed, reconstructed attacked, and the difference between the two reconstructed. For better visualisation, the magnitude of the adversarial noise and reconstructed difference is multiplied by 10 and 3, respectively. We used our default PGD attack with $\epsilon = 8/255$ and $n = 10$ iterations.

I Segmentation

J Computational Resources

The experiments were conducted on an internal cluster equipped with RTX 3090s and RTX 2080 TIs. In total, we have logged almost 5000 GPU hours for the experiments and testing. Most of the compute was spent on exploration and the diffusion experiments, with over 2000 hours being spent on the latter alone.

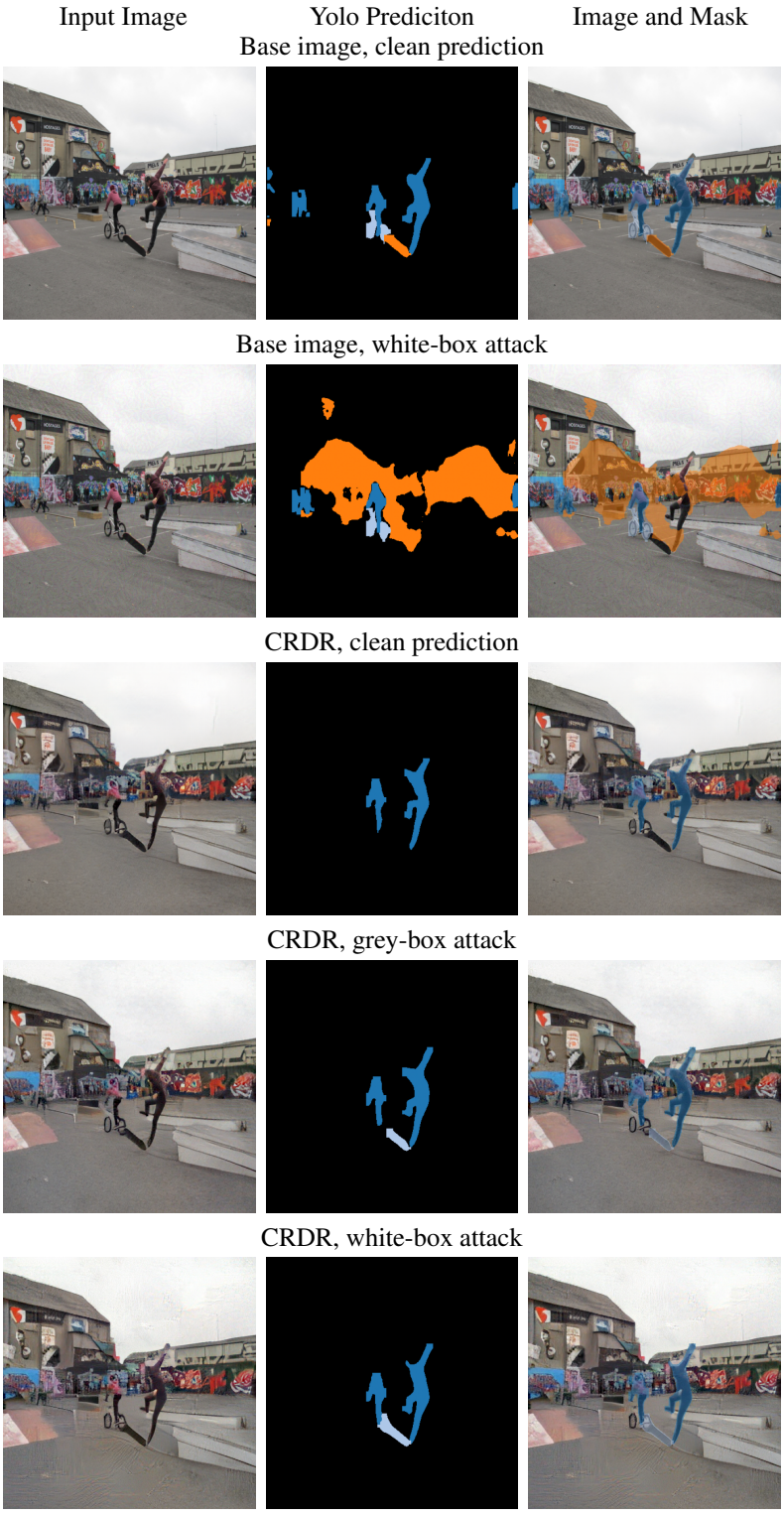

Figure 9: Visualization on an attack on the Yolo11 segmentation model. We used PGD with l_∞ norm 8/255. Using a compression based defense leads to a loss of some objects but preserves many of the objects present in the mask under grey-box and white-box attacks.

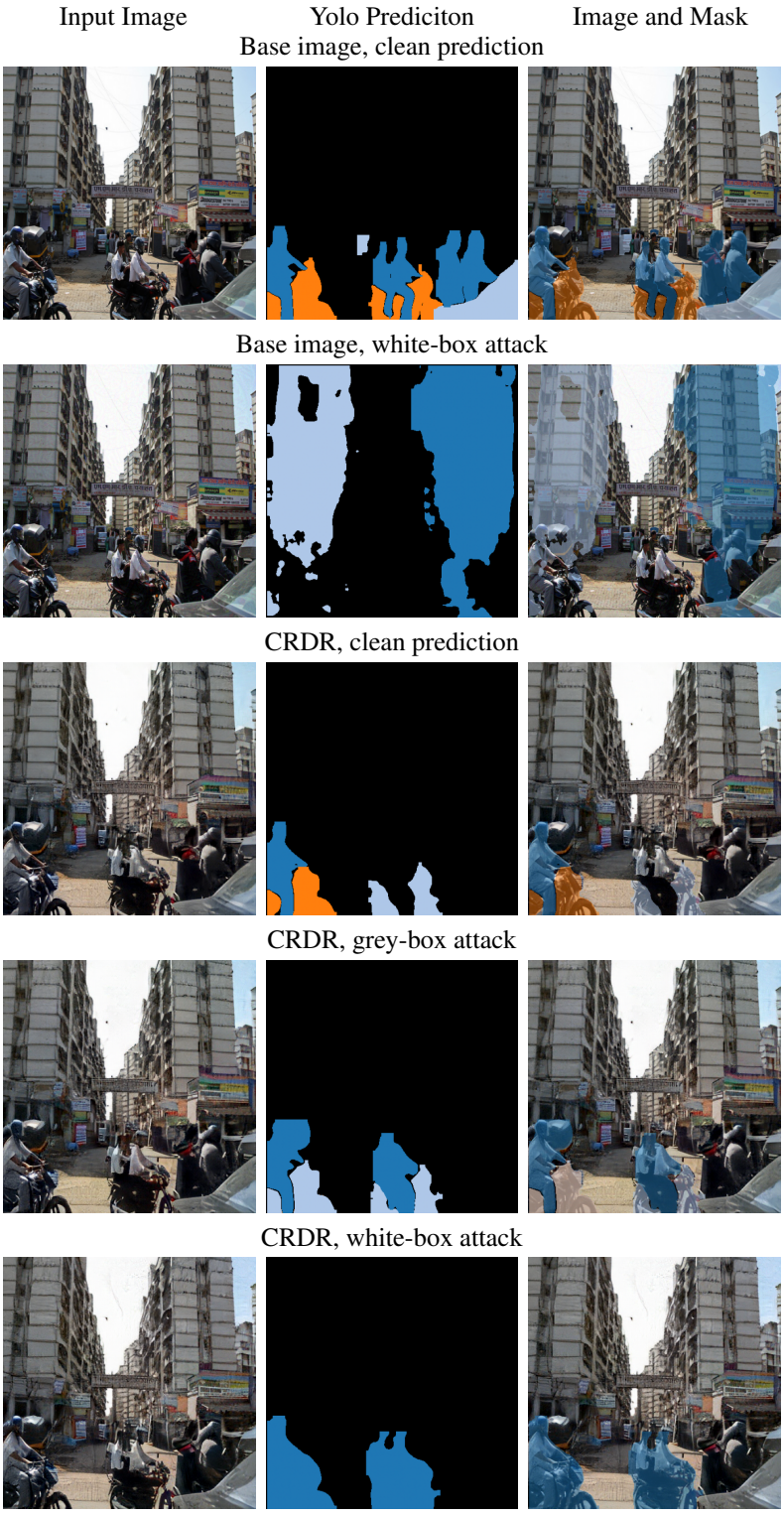

Figure 10: Visualization on an attack on the Yolo11 segmentation model. We used PGD with maximum l_∞ norm 8/255. Using a compression based defense leads to a loss of some objects but preserves many of the objects present in the mask under grey-box and white-box attacks.

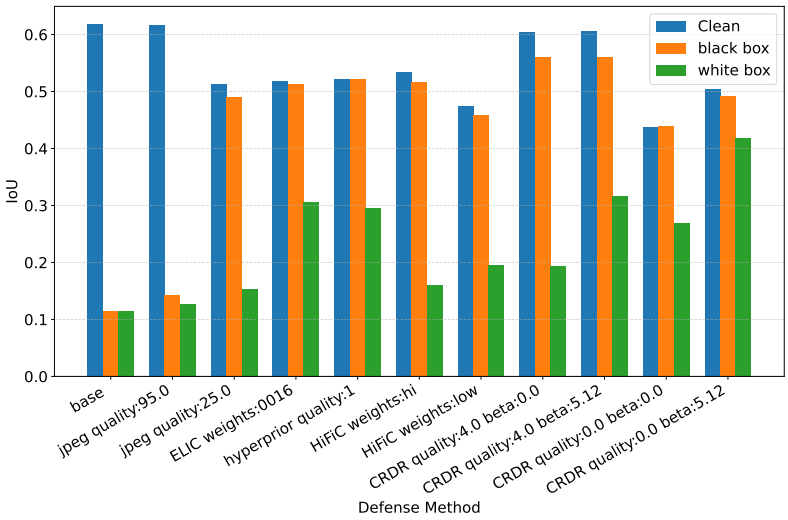

Figure 11: IoU of Yolo11 with different defenses against a PGD attack with maximum l_∞ norm 4/255

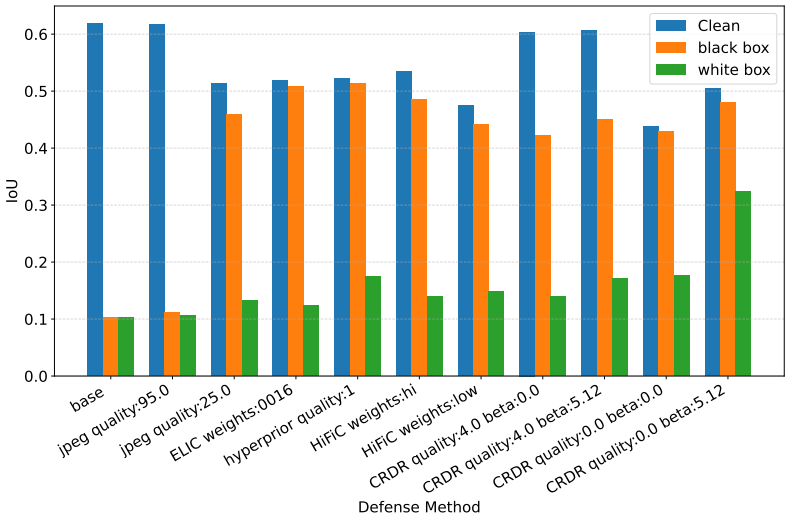

Figure 12: IoU of Yolo11 with different defenses against a PGD attack with maximum l_∞ norm 8/255

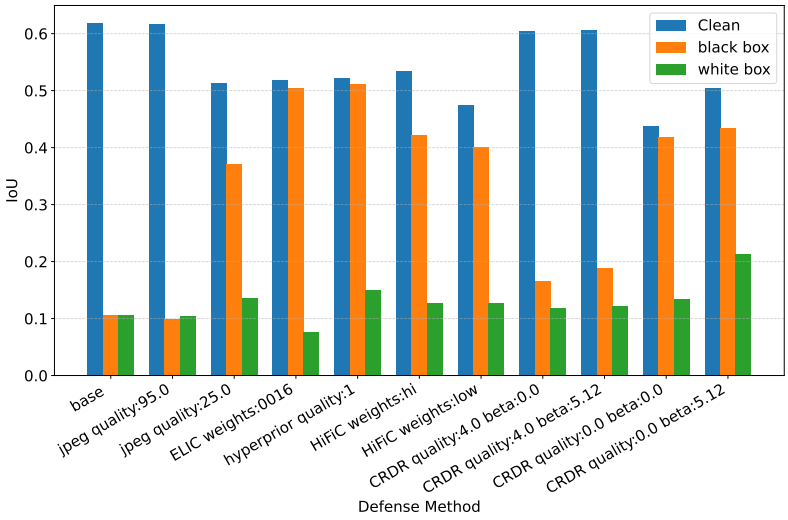

Figure 13: IoU of Yolo11 with different defenses against a PGD attack with maximum l_∞ norm 16/255

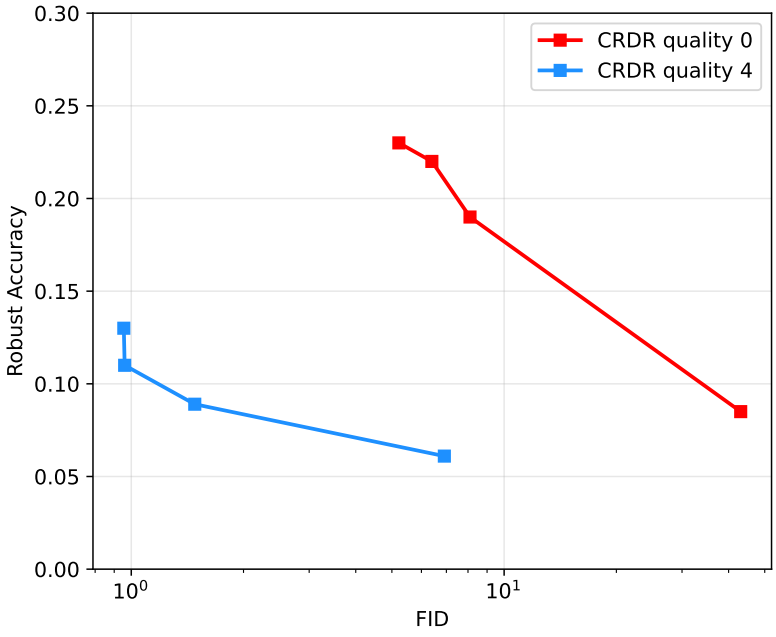

Figure 14: Robust accuracy of ResNet50 against a PGD attack with maximum l_∞ norm 8/255. CRDR with quality 0 and 4 and realism values [0.0, 1.28, 2.56, 3.84].