Test-Time Defense Against Adversarial Attacks via Stochastic Resonance of Latent Ensembles

Dong Lao* LSU
Yuxiang Zhang UCLA
Haniyeh Ehsani Oskouie UCLA
Yangchao Wu AWS
Alex Wong Yale
Stefano Soatto AWS & UCLA

Abstract

We propose a test-time defense mechanism against adversarial attacks: Unlike existing methods that rely on feature filtering or smoothing, which can lead to information loss, we propose to "combat noise with noise" by leveraging stochastic resonance to enhance robustness while minimizing information loss. Our approach introduces small translational perturbations to the input image, aligns the transformed feature embeddings, and aggregates them before mapping back to the original reference image. This can be expressed in a closed-form formula, which can be deployed on diverse existing network architectures without introducing additional network modules or fine-tuning for specific attack types. The resulting method is entirely training-free, architecture-agnostic, and attack-agnostic. Empirically, the method achieves state-of-the-art robustness on image classification and provides the first generic test-time defense for dense prediction tasks, including stereo matching, optical flow, and monocular depth estimation. Across adversarial attacks, it recovers up to 68.1% of the accuracy loss on image classification, 71.9% on stereo matching (MAE), 29.2% on optical flow (EPE), 68.3% on depth estimation (SqRel), and 83.8% on vision-language alignment (cosine similarity).

1 Introduction

Most deep neural networks in use today are deterministic maps from a fixed-size input to a fixed-size feature vector. In auto-regressive Transformer models, that vector encodes the next element (token) in the input sequence. Similarly, in convolutional architectures, that vector may encode the input data. In either case, the output vector is often highly sensitive to perturbations of the input, and one can intentionally choose these imperceptible perturbations *adversarially* so as to maximize the change in the output [30]. In some cases, the same perturbation can even be disruptive for a large number of possible inputs [48], exploiting the convoluted geometry of the decision boundary imposed by such trained models [68]. This spurious sensitivity could be exploited adversarially to disrupt the operation of a model.

From a classical perspective of signal processing, adversarial perturbations of images appear as small high-frequency "noise" resembling *aliasing* artifacts. These are imperceptible since the human visual system easily discounts them on account of their poor fit to the "ecological statistics" of natural images [29]. Classical sampling theory prescribes anti-aliasing by low-pass filtering the data, removing information along with the artifacts. Low-pass filtering consists of spatial averaging of the perturbed data with respect to a chosen kernel, typically a Gaussian or a constant ("pillbox"). Alternatively, one could "denoise" the embeddings by averaging output vectors, also a lossy operation. The choice of the kernel should match the statistics of the perturbations, which sets up a cat-and-mouse game

*Project page: `https://donglao.github.io/sr/`. Correspondence: dong.lao@lsu.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

where the adversary can easily modify the perturbations to bypass the anti-aliasing filter, namely *adaptive* attacks [67], and the model needs to constantly be updated to "anti-alias" new forms of adversarial perturbations. In the context of deep neural networks, existing defense methods that manipulate feature representations [80, 3, 83, 37] fundamentally adhere to this paradigm. Thus, despite substantial engineering efforts, these methods remain inherently vulnerable to adaptive attacks because they rely on pre-defined filtering strategies that are fixed at inference time.

Desiderata: To break this cycle, we advocate a method to mitigate the effect of adversarial perturbation that (i) operates at test-time, without the need to update the model weights, and (ii) does not entail information loss associated with direct or indirect spatial filtering. In addition, it would be ideal if this method could (iii) be applied to existing network architectures without modifications, and (iv) be agnostic to the specific type of adversarial perturbation.

Stochastic Resonance is a technique that resolves a quantized signal below the quantization level, by quantizing and ensembling perturbed versions of the signal [7]. It has been used extensively in cochlear implants, where the resolution of the digital circuitry is limited[62]. It has also been used to 'super-resolve' Vision Transformer embeddings, where entire patches are encoded into a vector, which is computed at a coarsely subsampled grid [42]. In this paper, instead, we use Stochastic Resonance for the opposite purpose, *not* to super-resolve the quantized signal, but to perform latent ensembling to mitigate the effects of adversarial perturbations.

Rather than averaging the data, or averaging their embedding as in classical denoising, *we average transformed embeddings* in latent space. This averaging is performed over small transformations sampled at random or deterministically from the group of planar translations, by computing the encoding of the transformed image, and then mapping the encoding back through the push-forward of the inverse transformation. This process can be expressed as a single formula in (1). Since the embedding is typically computed on a coarse grid, but the transformations are sampled on the native lattice of the image, the resulting embedding is free of spatial averaging artifacts. As with other uses of Stochastic Resonance, the effect is seemingly paradoxical as *we combat noise with noise*: We apply purposeful perturbations to eliminate the effect of adversarial perturbations.

Our method can be viewed as marginalizing the translation group in latent space with respect to a chosen prior. We choose the simplest, which is the constant prior. The purposeful perturbations alter the spatial sampling, and the ensembling in latent space averages out the effect of sampling artifacts, thwarting the effect of adversarial perturbations. It is as if we were given images with different 'noise', except that the noise in question is not the adversarial perturbation, but the splinters of adversarial perturbations obtained by different spatially quantized versions of the perturbed image, due to the translational perturbations, which are then averaged out by the latent ensembling.

Outcomes: Our method fulfills the desiderata (i)-(iv) laid out earlier: (i) It does not require training or fine-tuning; (ii) it minimizes information loss by latent ensembling of perturbed embeddings; (iii) it can be applied to different network architectures and tasks, *including* networks already equipped with different defense techniques like adversarial training, and (iv) is agnostic to the specific perturbation. We evaluate our method across six diverse tasks, including four dense prediction tasks. Stochastic resonance drastically improves adversarial robustness without training or architecture-specific modifications, recovering up to 68.1% of the accuracy loss in classification, 71.9% in stereo matching (MAE), 29.2% in optical flow (EPE), 68.3% in depth estimation (SqRel), and 83.8% in vision-language alignment (cosine similarity). To our knowledge, this is the first test-time defense to show consistent gains across such a broad range of settings, especially on dense prediciton tasks.

2 Related Work

Adversarial Training as Defense. Adversarial training increases the robustness of the model by training it with adversarially augmented images. The popular attack methods used are Fast Gradient Sign Method (FGSM) [30], and Projected Gradient Descent (PGD) [47]. ALP [36] Minimizes the difference between the logits of pairs of clean and adversarially augmented images. TRADES [86] decomposes prediction error for adversarial examples into natural error and boundary error to improve adversarial robustness. MART [75] improves adversarial robustness by considering misclassified natural examples during training. Subsequent work [9, 88, 73, 35, 28, 82, 31] improves adversarial robustness with curriculum learning [4], model ensembling, critical parameter selection, and gradient tracking. On the other hand, some methods learn robust feature representation through a modified

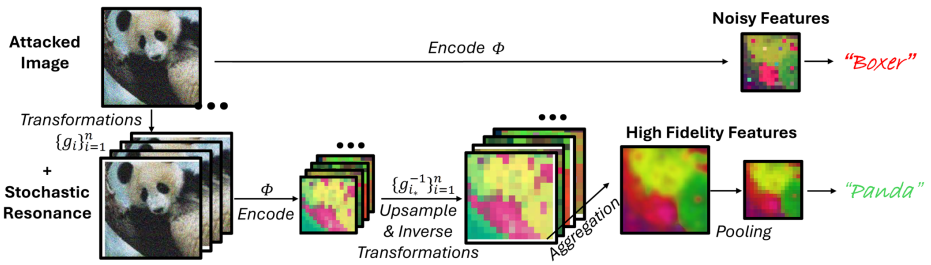

Figure 1: **Diagram.** Neural networks are sensitive to perturbations in the input space, which adversarial attacks exploit to manipulate outputs. Existing defense focuses on filtering out unreliable features or denoising either the input or the features. Instead, we propose a novel defense based on stochastic resonance. Controlled transformations are introduced to the input. Features are then aggregated after inverting these transformations. The resulting test-time defense method requires no training and is compatible with diverse architectures. Notably, it not only improves robustness against attacks but also increases the difficulty of crafting adversarial examples, even when the attacker is aware of whether and how stochastic resonance is being used (adaptive adversary).

architecture or feature manipulation. [23, 5, 72] investigate the effect of batch normalization on adversarial robustness. [20, 46] prunes certain activations in the network that are susceptible to adversarial attacks. [79] keep k-features with the largest magnitude and deactivate everything else. [90] uses an attention mask to highlight robust regions on the feature. Feature Denoising (FD) [80] uses classical denoising techniques to deactivate abnormal activations. [3, 83] proposed Channel Activation Suppression (CAS) and Channel-wise Importance-based Feature Selection (CIFS) to deactivate feature channels that are vulnerable to attacks. [37] improves the robustness with Feature Separation and Recalibration (FSR). Our method also operates in feature space, but purely at test-time. While we choose some as baselines, our method can work in conjunction with any aforementioned methods.

Adversarial Purification as Defense Another line of work focuses on purifying or augmenting the images before they are used as input. [63, 84, 69] train an FGSM robust classifier, a diffusion model, or a mask auto-encoder, respectively, to purify adversarial examples. [71] optimizes both the model and the input to minimize the entropy of model predictions to adapt to changing attacks. [15] trains a random forest predictor to ensemble outputs from test-time augmented images. These works involve training a new model or updating the original model, while our method is purely test-time and does not require any training. [53] ensembles model output from different augmentations, which is a special case of our method, as the ensemble is performed solely on the output, while we can ensemble at any layer, which both saves computational cost and achieves higher performance. Notably, we recognize that above methods focus solely on classification as a task for adversarial attacks. Through extensive experiments, we demonstrate that our method can not only perform well on classification, but also on dense prediction tasks like optical flow, and stereo matching.

Stochastic Resonance (SR) was proposed by [7] and first applied in climate dynamics [6] and later in signal processing [76, 38, 12] and acoustics [61, 74]. While several works [89, 70] introduce stochasticity within feature space, the most closely related prior effort is [42], which applies stochastic resonance to "super-resolve" Vision Transformer (ViT) embeddings. In this work, we instead use SR to mitigate adversarial attacks. Since SR has been developed specifically to address *quantization artifacts*, it has never before been used to mitigate classes of perturbations beyond aliasing. Our novel use of the technique leverages the fact that group transformations and spatial quantization preserve the statistics of natural images, which are heavy-tailed, but do not preserve the statistics of adversarial perturbations. We are also closely related to [22, 25, 26, 41, 42, 78] in use of group transformations.

3 Method

Formalization. An image $x : \Lambda \to \mathbb{R}^k$ is a sampled signal on a finite pixel lattice $\Lambda \subset \mathbb{Z}^2$, with k color channels ($k = 3$ for RGB). At the layer where we apply stochastic resonance, the encoder ϕ produces a spatial feature map

$$\phi(x) : \tilde{\Lambda} \to \mathbb{R}^K,$$

where $\tilde{\Lambda}$ is the feature lattice and K is the number of feature channels.

Let $\tau_{u,v}(p) = p + (u, v)$ denote a translation of spatial coordinates. We write $g_{u,v}$ for its action on an image:

$$[g_{u,v}(x)](p) = x_{\text{ext}}\big(\tau_{u,v}^{-1}(p)\big), \qquad p \in \Lambda,$$

where x_{ext} equals x on Λ and is zero outside Λ. On the finite image grid, shifting back need not recover samples lost at the boundary.

For an image transformation g, let A_g denote the operation that aligns the resulting spatial feature map to the original image coordinates. The main object of interest in our method is

$$A_g\big(\phi(g(x))\big).$$

Perturbations. We consider two types of perturbations, extraneous and purposeful. The extraneous one could be an additive perturbation to an image, $\tilde{x} = x + n$, designed to maximally change the value of the embedding $\phi(\tilde{x})$:

$$n(x) \in \arg\max_{\|\delta\|_p \leq \epsilon} d\big(\phi(x), \phi(x+\delta)\big), \qquad \tilde{x} = x + n(x)$$

for some small ϵ designed so the perturbation is, ideally, imperceptible by humans (adversarial perturbation).

The purposeful perturbations are small translations $g_i,\ i = 1, \ldots, N$, which could be sampled deterministically or at random according to some chosen distribution $g_i \sim P_g$, either way yielding a set $\{g_1(x), \ldots, g_N(x)\}$. Our goal is to use these purposeful perturbations to combat the effects of extraneous adversarial perturbations.

Smoothing vs. stochastic resonance. The resemblance between adversarial perturbations and aliasing has motivated the use of anti-aliasing, or smoothing, techniques to mitigate them. These consist of *spatial averaging* of the data prior to computing the map ϕ. If we call B_{ij}^{σ} a neighborhood of size $\sigma > 0$ around $(i, j) \in \Lambda$, then the simplest form of smoothing is simply averaging in a neighborhood,

$$\bar{x}_{i,j} = \frac{1}{\sigma^2} \sum_{(i',j') \in B_{i,j}^{\sigma}} x_{i'j'}$$

which we write in terms of translations $[g_{u,v}(x)](p) = x_{\text{ext}}(p-(u, v))$ within the same neighborhood B^{σ},

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} g_i(x).$$

One can then obtain an encoding by

$$\bar{\phi}(x) = \frac{1}{N} \sum_{i=1}^{N} \phi(g_i(x)).$$

This can be interpreted as marginalizing over the chosen set of translations with a prior P_g when computing ϕ. Notice that the average can be computed on a coarser lattice $\tilde{\Lambda}$, but its value still depends on data in the finer grid Λ. Classical sampling theory teaches that smoothing mitigates the effect of high-frequency aliasing at the cost of information loss on x.

Stochastic resonance also marginalizes over the chosen set of translations, but by *averaging transformed data in latent space:*

$$\hat{\phi}(x) = \frac{1}{N} \sum_{i=1}^{N} A_{g_i}\big(\phi(g_i(x))\big). \tag{1}$$

For an encoder like ViT, A_{g_i} reshapes the spatial tokens of $\phi(g_i(x))$ into a two-dimensional feature map, upsamples it to the input-pixel lattice using nearest-neighbor interpolation, and applies the opposite integer translation on that lattice. Equation (1) averages these aligned feature maps on the input-pixel lattice, after which we apply average pooling to return to the original token resolution.

Defense against adversarial attacks (sketch only; please find details in Appendix A).

Consider a first-order expansion at each transformed input

$$\phi(g_i\tilde{x}) = \phi(g_i x + g_i n(x)) \approx \phi(g_i x) + J_{g_i x}\,[g_i n(x)],$$

Table 1: **Defense against adversarial attacks on CIFAR-10.** Unlike baselines, we perform defense purely at test-time, and achieve state-of-the-art robustness under different attacks. Even in an adaptive setting (worst-case) where the attacker is fully aware of how defense is applied, the effectiveness of adversarial attacks is still notably reduced, while the computational cost for executing such attacks is significantly increased.

Method	Architecture	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble	AutoAttack
AT [47]	ResNet-18	85.02	56.21	48.22	46.37	47.38	45.90	44.11
+ FD [80]		85.14	56.81	48.54	46.70	47.72	45.82	44.57
+ CAS [3]		**85.78**	55.57	50.42	49.91	53.47	46.46	44.23
+ CIFS [83]		79.87	56.53	49.80	48.17	49.89	47.26	43.94
+ FSR [37]		81.46	58.07	52.47	51.02	49.44	48.34	46.41
+ TTE [53]		85.25	59.20	53.00	51.65	52.45	50.60	-
+ Ours		84.93	**61.02**	**56.08**	**55.17**	**55.53**	**53.68**	**49.78**
+ Ours (Worst-Case)		84.93	58.81	53.58	52.39	52.73	50.95	-
TRADES [86]	WideResNet-34	84.92	60.87	56.13	55.16	54.02	53.38	46.81
+ Ours		**85.03**	**62.43**	**58.64**	**57.87**	**57.18**	**56.28**	**54.02**
MART [75]	ResNet-18	**83.07**	60.21	54.14	52.90	49.62	48.95	44.27
+ Ours		82.70	**62.62**	**59.03**	**58.13**	**55.51**	**54.61**	**50.72**

where $J_{g_i x}$ is the Jacobian of ϕ at $g_i x$. Substituting with (1) yields the perturbation term

$$\Delta_{\mathrm{SR}}(x) = \hat{\phi}(\tilde{x}) - \hat{\phi}(x) \approx \frac{1}{N} \sum_{i=1}^{N} A_{g_i}\big(J_{g_i x}[g_i n(x)]\big). \tag{2}$$

By contrast, a standard feedforward encoder responds as

$$\Delta_{\mathrm{base}}(x) = \phi(\tilde{x}) - \phi(x) \approx J_x\, n(x),$$

and is therefore dominated by the local worst-case Lipschitz direction at x, exploited by the attacker.

The terms $A_{g_i}\big(J_{g_i x}[g_i n(x)]\big)$ in (2) are generally *not aligned* across i. Under mild assumptions (decorrelation across i, details and empirical evidence in Appendix A),

$$\mathbb{E}\left\| \frac{1}{N} \sum_{i=1}^{N} v_i \right\|^2 \approx \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}\|v_i\|^2, \qquad v_i = A_{g_i}\big(J_{g_i x}[g_i n(x)]\big),$$

provided that the sum of the expected cross terms is small in magnitude relative to $\sum_{i=1}^{N} \mathbb{E}\|v_i\|^2$. Under this condition, the root-mean-square magnitude of the first-order perturbation term in (2) is approximately $1/\sqrt{N}$ times the root-mean-square magnitude of the individual terms v_i across the chosen translations g_i.

Purposeful perturbation. The only design choice in this principle is the set of purposeful perturbations. While that can be optimized for performance, we optimize for simplicity, restricting our attention to translation by integer pixels. The theory of Stochastic Resonance shows that, if we sub-sample a signal from its native granularity Λ to a coarser grid $\tilde{\Lambda} \subset \Lambda$, and choose the purposeful perturbations to act on the finer grid Λ, under certain conditions one can recover the original signal at the finer granularity [7]. Accordingly, we enumerate the bounded integer-pixel translations

$$\{g_i\}_{i=1}^{N} = \big\{g_{u,v} : (u,v) \in \mathbb{Z}^2,\ |u| \le d_x,\ |v| \le d_y\big\}, \qquad N = (2d_x+1)(2d_y+1).$$

Whenever we report a single SR level d, we use $d_x = d_y = d$.

4 Experiments

We evaluate across 6 tasks (image classification, stereo matching, optical flow, semantic segmentation, monocular depth estimation, vision-language alignment), multiple architectures and variants (ResNet, ViT, MLP-Mixer, PSMNet, RAFT), totaling over 100 settings that vary datasets, models, attack types and strengths, and defense strategies. Given the breadth of the evaluation, we focus on the key findings and discussions in the main paper and defer implementation details and additional ablations to the Appendix.

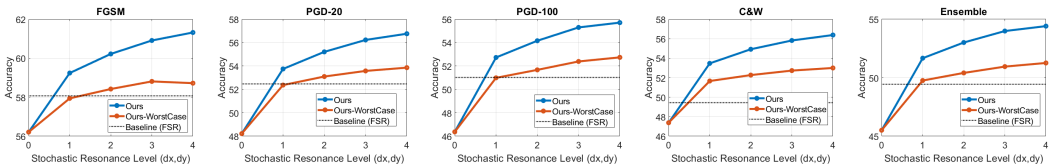

Figure 2: **Varying levels of stochastic resonance.** Increasing the stochastic resonance level enhances robustness across all settings. Notably, our approach achieves superior performance even under adaptive attacks (Ours-WorstCase), despite the baseline (FSR) being evaluated only in the non-adaptive case.

4.1 Image Classification

CIFAR-10. We evaluate on CIFAR-10 [39], building upon the publicly available code base of FSR [37] and accompanying evaluation protocol, including AutoAttack [18]. We apply stochastic resonance purely at test-time, to networks pre-trained with AT [47], TRADES [86], and MART [75], using their publicly released weights. We conduct experiments on ResNet-18 [33] and WideResNet-34 [85], depending on the availability of author-released pre-trained weights. In all cases, feature ensembling is performed before the final linear layer. Furthermore, we consider a worst-case adaptive adversary setting, where the attacker has full knowledge of the model weights, and knows every stochastic resonance transformation by accessing *every single* forward pass and its gradients. Implementation details are in Appendix G.

We compare our method against feature-level manipulation methods (FD [80], CAS [3], CIFS [83]) and ensemble-based approach TTE [53]. The results (Tab. 1), show that stochastic resonance consistently outperforms all baselines across different attacks. Even in the adaptive attack scenario, where the adversary accounts for all stochastic resonance forward passes, the model remains significantly more robust than the baseline methods in non-adaptive settings. Note that, with stochastic resonance, the computational cost for generating adaptive attacks increases substantially: $8\times$ more wall-clock time is required to create adversarial examples with PGD-100. Fig. 2 shows that increasing the resonance level consistently enhances robustness in all attack settings, leading to substantial improvements over the baseline method (FSR).

ImageNet. We further evaluate our approach on the ImageNet [19] using a ViT [21] backbone. We vary the level of stochastic resonance and conduct ablation studies by testing against PGD attacks of different strengths. Tab. 2 reports the results, where increasing the resonance level leads to consistent improvements in adversarial robustness. Notably, our method recovers the accuracy drop under adversarial attacks by up to a relative 55.8% (Top-1, when $\epsilon = 4/255$) and 68.1% (Top-5, when $\epsilon = 2/255$). Our experiments on ResNet-50 also demonstrate a consistent improvement, as in Tab. 12 and Tab. 13 in the Appendix.

Table 2: **ImageNet with ViT-Small.** Increasing the level of stochastic resonance consistently improves both Top-1 and Top-5 accuracy under adversarial attacks. Relative to the clean baseline (72.9 Top-1, 92.91 Top-5), our method recovers up to 55.8% of the Top-1 accuracy loss and 68.1% of the Top-5 accuracy loss.

Att. Strength (ϵ)	8/255		4/255		2/255		1/255	
Metric	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
No Defense	4.51	19.55	12.25	42.15	29.79	65.09	48.09	78.13
Ours ($d = 1$)	11.66	46.27	27.85	65.80	45.94	77.63	57.48	83.43
Ours ($d = 2$)	18.78	58.58	36.42	72.86	51.58	80.85	60.41	84.77
Ours ($d = 3$)	**25.77**	**65.88**	**42.52**	**76.75**	**54.94**	**82.64**	**62.08**	**85.50**

Applying stochastic resonance to early layers. (Details in Appendix E.1.) Since adversarial perturbations often appear as high-frequency noise, early feature extraction is a natural point of intervention. As shown in Tab. 12, applying stochastic resonance through only the first residual block of ResNet-50 already provides strong robustness, while extending it to the second block yields the best performance. This suggests that running stochastic resonance through shallow layers can be sufficient for defense, substantially reducing computational overhead.

Alternative purposeful perturbations. (Details in Appendix E.2.) We also evaluate rotations as an alternative family of purposeful perturbations. As shown in Tab. 14, rotations improve adversarial robustness across all resonance levels, with larger gains at lower resonance levels. We hypothesize that this is partly due to the implementation of standard libraries. Unlike translations with integer displacements, rotations typically map pixel coordinates to non-integer locations and therefore require interpolation-based resampling, which can introduce artifacts such as aliasing and blurring. Higher-fidelity resampling strategies, such as supersampling or Fourier-domain rotation, may mitigate these



Figure 3: **Qualitative results on stereo matching.** We demonstrate a significant reduction in prediction errors. This technique holds significant potential for safety-critical applications where stereo vision must remain reliable under diverse environmental conditions and adversarial threats. Additional visualizations at Appendix N.

Table 3: **Stochastic resonance enhances stereo matching robustness.** Stochastic resonance significantly mitigates errors across all settings, reducing error by up to 71.89% (MAE, attacked by PGD with $\epsilon = 0.005$).

	Attack Strength (ϵ)	0.02			0.01			0.005			0.002		
	Metric	MAE	RMSE	D1-err	MAE	RMSE	D1-err	MAE	RMSE	D1-err	MAE	RMSE	D1-err
FGSM	No Defense	14.83	24.10	97.33	8.49	14.53	90.49	5.05	7.70	74.71	3.01	3.49	38.12
	Latent Smoothing [1]	13.42	22.09	96.61	8.12	13.69	89.25	4.89	7.05	73.01	2.89	3.32	36.25
	Randomized Smoothing [16]	14.20	22.41	97.26	7.38	11.21	89.44	4.26	5.19	70.49	2.78	3.18	30.76
	Ours ($d = 1$)	10.12	15.81	95.80	6.46	9.92	86.30	4.39	5.93	68.55	2.78	3.18	31.99
	Ours ($d = 2$)	**9.22**	**13.88**	**94.61**	**6.13**	**8.87**	**84.49**	**4.19**	**5.43**	**66.74**	**2.73**	**3.12**	**31.18**
	Error Reduced (%)	37.82	42.40	2.79	27.78	38.95	6.63	16.92	29.41	10.66	9.21	10.65	18.21
PGD	No Defense	161.70	162.61	99.99	131.66	140.55	98.64	63.97	88.55	85.31	6.83	17.03	39.24
	Latent Smoothing [1]	161.79	162.69	99.99	131.46	140.41	98.57	63.86	88.21	85.39	7.28	17.44	39.81
	Randomized Smoothing [16]	149.00	153.36	99.95	62.59	83.94	95.09	11.06	19.03	69.85	2.69	3.38	25.01
	Ours ($d = 1$)	107.86	125.14	98.29	69.97	84.72	91.79	20.66	42.29	73.21	4.17	8.32	29.09
	Ours ($d = 2$)	**77.59**	**100.14**	**96.14**	**44.77**	**66.23**	**89.24**	**17.98**	**32.80**	**71.09**	**3.76**	**6.73**	**28.44**
	Error Reduced (%)	52.01	38.41	3.85	66.02	52.87	9.53	71.89	62.94	16.66	44.99	60.52	27.53

artifacts. Nevertheless, the robustness gains achieved by rotations provide a proof of concept: our method extends beyond translations and is not specific to a single transformation class.

Extending to MLP-based architectures. (Details in Appendix F.) We further check whether stochastic resonance extends beyond CNNs and Transformers by evaluating it on MLP-Mixer [65]. On CIFAR-10, stochastic resonance preserves clean accuracy while substantially improving adversarial robustness: under PGD with $\epsilon = 8/255$, accuracy increases from 0.03 without defense to 65.38 with stochastic resonance at $d = 3$. Similar gains are observed under FGSM, where accuracy improves from 21.76 to 74.08 at $\epsilon = 8/255$.

Real-world corruptions. We also test how stochastic resonance behaves under real-world corruptions. To this end, we evaluate on CIFAR-10-C [34], which contains a diverse set of corruptions covering noise, blur, weather effects, and photometric changes. With ResNet-50, following the best configuration identified in prior sections, stochastic resonance recovers an average of 20.8% of lost accuracy, including 32.6% under `defocus_blur`, 34.4% under `fog`, 31.6% under `gaussian_blur`, and 31.9% under `saturate`. Other corruptions also see recovery ranging from 3.4% to 20.8%. These findings highlight a key strength of stochastic resonance: it works across a wide spectrum of *unknown* corruptions, strengthening it as a *general-purpose*, *plug-and-play* defense.

4.2 Dense Prediction

We next evaluate stochastic resonance on dense prediction tasks. This setting highlights a key strength of our method: standard data augmentations are often infeasible because they can alter the input geometry and violate the underlying physics of the task. In contrast, stochastic resonance ensembles features in latent space after alignment, keeping predictions registered to the original input.

Stereo matching. (Details in Appendix H.) Stereopagnosia [8, 77] first introduced adversarial attacks to stereo matching, yet no test-time method has demonstrated an effective defense. We evaluate our

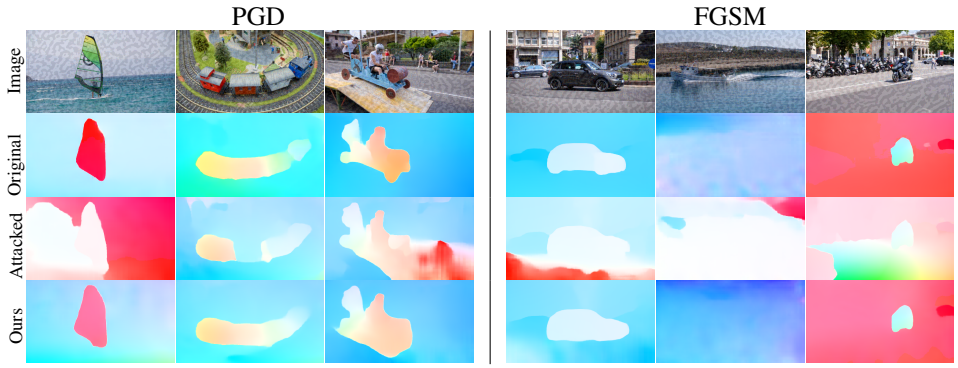

Figure 5: **Qualitative results for optical flow.** Our method substantially mitigates the degradation caused by both PGD and FGSM attacks. Such robustness is critical for perception systems that rely on motion estimation.

method on the standard benchmark used in *Stereopagnosia*, derived from KITTI [27]. Experiments are conducted using FGSM and PGD attacks against a pre-trained PSMNet [11]. Because no established test-time defense is available, we implement latent-space smoothing [1] and randomized smoothing [16] as baseline methods. As shown in Tab. 3, both attacks corrupt network predictions, with PGD proving more effective due to its iterative nature. Nevertheless, stochastic resonance consistently improves robustness across attack strengths. In particular, under PGD with $\epsilon = 0.005$, our method reduces errors by up to 71.89% in terms of MAE. Fig. 3 provides qualitative results. The attacks introduce significant errors, as indicated by bright regions in the predictions. Stochastic resonance effectively mitigates these errors. Note that this is achieved with neither prior knowledge of the attack nor additional robust training. This experiment is particularly relevant for safety-critical applications such as autonomous driving, where adversarial disturbances can arise not only from malicious attacks but also from environmental factors such as adverse weather conditions, varying illumination, or sensor degradation. While some defenses [87, 13, 14, 8] have been proposed to train a more robust network under adverse conditions, test-time defenses remain largely unexplored.

Optical flow. (Details in Appendix K.) We further evaluate our method on optical flow, which computes the dense motion field between two images. The accuracy is measured using the End-Point Error (EPE). Although several adversarial attacks have been proposed [60, 56, 59, 2, 57], there is currently no standardized benchmark for the evaluation of defense. Therefore, we adopt a simple and controlled setup: we employ the RAFT [64] model and focus on global perturbations generated by PGD and FGSM. We apply stochastic resonance to the feature extractor of RAFT, introducing no overhead to the computationally intensive matching module. We test on the DAVIS [54] dataset. As in Fig. 5, our approach effectively removes errors caused by adversarial noise. Quantitative results (Fig. 4) show that increasing stochastic resonance reduces EPE, which aligns with our findings in classification. We also compare our method to an alternative ensembling approach that aggregates predictions in the output space, similar to TTE [53]. We apply the same stochastic transformations, but instead perform an ensemble at the output level (with alignment). This approach remains less effective than our approach. This finding highlights the advantage of having the freedom to ensemble at any layer.

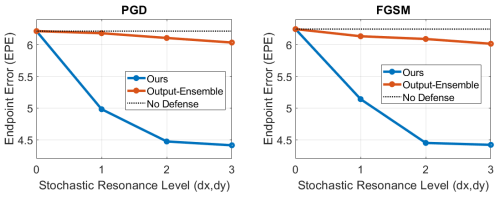

Figure 4: **Improved optical flow robustness.** By ensembling in the latent feature space rather than the output space, stochastic resonance consistently achieves stronger robustness across all resonance levels.

We further evaluate stochastic resonance under adverse weather conditions [58]. This setting poses two unique challenges: (i) the corruptions are *not* optimized for the model and instead reflect natural image statistics, and (ii) the perturbations violate the underlying theoretical assumptions in Sec. 3. Despite these challenges, stochastic resonance still recovers 7.5% of the accuracy loss for `snow` and 3.7% for `rain`. While, as expected, the gains are smaller than observed under adversarial attacks, they

Table 4: **Semantic segmentation on Cityscapes under attack.** We report mIoU and boundary mIoU (%) for the baseline model and three levels of stochastic resonance. Higher is better for both metrics.

		mIoU ↑				Boundary mIoU ↑			
Attack	ϵ	No SR	$d=1$	$d=2$	$d=3$	No SR	$d=1$	$d=2$	$d=3$
CosPGD	$2/255$	18.40	25.53	31.14	**32.75**	15.14	21.57	26.71	**28.21**
	$4/255$	17.16	22.42	26.29	**27.38**	13.74	18.61	22.26	**23.28**
	$8/255$	15.39	19.15	21.71	**22.41**	12.18	15.57	17.96	**18.62**
SegPGD	$2/255$	1.02	5.64	18.24	**23.46**	1.08	6.06	17.84	**22.11**
	$4/255$	0.47	1.50	6.43	**9.06**	0.49	1.66	6.79	**9.32**
	$8/255$	0.40	0.67	1.70	**2.30**	0.41	0.72	1.81	**2.41**

nonetheless demonstrate that stochastic resonance provides meaningful robustness against real-world perturbations, consistent with our findings on CIFAR-10-C.

Semantic segmentation. We next evaluate stochastic resonance on semantic segmentation, where the model assigns a class label to each pixel and must localize the boundaries between adjacent regions. We use Cityscapes [17] and evaluate under CosPGD [2] and SegPGD [32] at three perturbation strengths, $\epsilon \in \{2, 4, 8\}/255$. We follow their protocols, except for adding Stochastic Resonance at test time. In addition to mean intersection over union (mIoU), we report boundary mIoU to examine whether the gains from aggregating aligned features extend to pixels near object boundaries. As shown in Tab. 4, both metrics improve with each increase in the resonance level across all tested attack strengths. For example, under CosPGD with $\epsilon = 4/255$, stochastic resonance at $d = 3$ increases mIoU from 17.16 to 27.38 and boundary mIoU from 13.74 to 23.28. SegPGD yields lower scores at the same perturbation strength, but we observe a similar trend: mIoU increases from 0.47 to 9.06 and boundary mIoU from 0.49 to 9.32. At the largest SegPGD budget, $\epsilon = 8/255$, the gains are smaller, with mIoU reaching 2.30 at $d = 3$.

Monocular depth estimation. (Details in Appendix I.) We also evaluate stochastic resonance on monocular depth estimation with larger ViT backbones. We use DINOv2 [51] ViT-B/14 and ViT-L/14 backbones on NYU-V2 [49]. Under PGD-10 attacks with $\epsilon = 8/255$, stochastic resonance consistently recovers performance across all evaluation metrics. The strongest recovery is observed on SqRel, where stochastic resonance recovers 68.3% of the attack-induced degradation for ViT-B/14 and 47.3% for ViT-L/14. These results further indicate that the method remains effective for larger transformer backbones and higher-dimensional embeddings.

Table 5: **Monocular depth on NYU-V2 with larger ViT backbones.** We evaluate under PGD-10 attacks with $\epsilon = 8/255$. Error reduction is measured as relative recovery toward the clean baseline.

Backbone	Method	RMSE ↓	RMSE-log ↓	AbsRel ↓	SqRel ↓
ViT-B/14	Clean	0.396	0.135	0.100	0.061
	No Defense	0.645	0.191	0.168	0.383
	Ours	0.564	0.181	0.155	0.163
	Error Reduced (%)	32.5	17.9	19.1	68.3
ViT-L/14	Clean	0.373	0.127	0.093	0.054
	No Defense	0.588	0.189	0.163	0.236
	Ours	0.573	0.186	0.159	0.150
	Error Reduced (%)	7.0	4.8	5.7	47.3

4.3 Vision-Language Alignment

We further evaluate whether stochastic resonance remains effective for vision-language models. We use CLIP [55] ViT-L/14 and evaluate image-text alignment on COCO-Captions [43] under PGD attack, using cosine similarity between image and text embeddings as the objective. (Details in Appendix J.) The attack substantially the image-text alignment, reducing cosine similarity from 0.302 to -0.025. Stochastic resonance progressively restores the alignment as the resonance level increases, reaching 0.179 with $d = 3$ (Tab. 6). We also evaluate adversarial training (AT). Although AT is not directly comparable to our training-free method, stochastic resonance outperforms it under the same evaluation and remains complementary: combining adversarial training with stochastic resonance yields the strongest result.

Table 6: **Vision-language alignment under attack.** On CLIP ViT-L/14 under PGD attacks, stochastic resonance improves image-text alignment as the resonance level increases, and complements adversarial training (AT).

Defense	Clean	Baseline	$d=1$	$d=2$	$d=3$
No Defense	0.302	-0.025	0.064	0.145	**0.179**
AT	0.302	0.136	0.195	0.231	**0.249**

We also evaluate whether stochastic resonance complements existing CLIP defenses on zero-shot ImageNet classification. Following the TTC protocol [81], we evaluate CLIP ViT-B/32 under PGD-10 with $\epsilon = 1/255$. At $d = 3$, adding SR to TTC increases adversarial Top-1/Top-5 accuracy (%) from 38.49/63.12 to 48.10/75.09, while AOM [66] achieves 44.06/68.98 under the same protocol. For CSR [50], applying SR to classify the rectified image increases adversarial Top-1/Top-5 accuracy from 18.76/38.62 to 31.08/63.54. Adding SR also improves clean Top-1 accuracy for both TTC (49.34% to 54.92%) and CSR (58.65% to 59.43%).

Note that, in modern vision-language models, the vision encoder is typically a small component relative to the language model. For example, in LLaVA-13B [44], the vision encoder occupies roughly 1.1GB in fp32, whereas the language backbone requires more than 24GB even in fp16. Since stochastic resonance is applied only to the vision encoder, the added cost is limited to this smaller component, increasing less than 10% additional memory for models like LLaVA-13B in our best settings.This makes our method practical for current VLMs, where robustness can be improved without modifying or repeatedly evaluating the substantially larger language backbone.

5 Discussion and Conclusion

Speed. Stochastic resonance incurs low overhead due to parallel computing: raising d to 3 with ResNet-50 adds only 0.06 seconds to the inference time on an NVIDIA 1080Ti GPU. Notably, most of the overhead arises from transformations implemented by Python loops, which is expected to be drastically reduced with efficient CUDA implementations. Excluding this `for` loop, forward computation is lightweight: 0.028s with d=3 (49 forward passes), while a single forward pass (d=1) takes 0.007s. In Appendix B we provide a detailed speed-performance tradeoff. Moreover, robustness can be attained by applying stochastic resonance only through shallow layers, offering greater efficiency than existing ensemble-based defenses (e.g. [53]). Moreover, stochastic resonance is fully plug-and-play. In contrast, adversarial training is attack-specific, over 6× slower than standard training, and must be repeated for each type of threat. As such, the computation of our method is well justified by its robustness gains and training-free nature.

Memory. Increasing d from 0 to 3 raises the number of forward passes from 1 to 49, while peak memory grows only moderately (158MB to 650MB) for ImageNet-sized inputs on ResNet-50 (details in Appendix C), which is easily handled by modern GPUs. In contrast, memory usage for *adaptive* attack jumps from 346MB to 10.6GB for a single 224×224 image, rendering such attacks impractical on standard hardware and conferring an inherent defensive advantage. Overall, stochastic resonance provides a favorable cost-robustness trade-off with modest memory growth.

Limitations and future work. Although the computation can be parallelized, the overhead may still be non-negligible in settings with strict memory, latency, or power constraints. More broadly, how to dynamically allocate test-time compute remains an open problem: different inputs, attacks, and deployment environments may require different resonance levels and applying stochastic resonance to different layers. Selecting this budget adaptively is an important direction for future work.

Conclusion. We present a signal-processing perspective on adversarial robustness that connects adversarial perturbations to aliasing effects, and propose stochastic resonance as a principled "combat noise with noise" test-time defense. Implemented through simple pixel-level translations and inverse feature alignment, our approach is training-free, architecture-agnostic, and broadly applicable across tasks and attack models. Extensive experiments demonstrate state-of-the-art robustness on image classification benchmarks, including under adaptive adversaries, and establish the first effective test-time defense for dense prediction tasks, achieving up to a 71% error reduction in stereo matching. Beyond strong empirical performance, our results highlight stochastic resonance as a simple yet powerful mechanism for improving robustness without additional training, offering a practical path toward more reliable deployment of models in adversarial and real-world settings.

Acknowledgments and Disclosure of Funding

Dong Lao is supported by start-up grants from Louisiana State University. Alex Wong is supported by NSF under the NSF AI Institute for Edge Computing and Systems (Athena), Grant No. 2112562. The authors have no additional competing interests to disclose.

References

[1] Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R Venkatesh Babu. Boosting adversarial robustness using feature level stochastic smoothing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2021.

[2] Shashank Agnihotri, Steffen Jung, and Margret Keuper. Cospgd: an efficient white-box adversarial attack for pixel-wise prediction tasks. *arXiv preprint arXiv:2302.02213*, 2023.

[3] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. *arXiv preprint arXiv:2103.08307*, 2021.

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[5] Philipp Benz, Chaoning Zhang, and In So Kweon. Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7818–7827, 2021.

[6] Roberto Benzi, Giorgio Parisi, Alfonso Sutera, and Angelo Vulpiani. Stochastic resonance in climatic change. *Tellus*, 34(1):10–16, 1982.

[7] Roberto Benzi, Alfonso Sutera, and Angelo Vulpiani. The mechanism of stochastic resonance. *Journal of Physics A: mathematical and general*, 14(11):L453, 1981.

[8] Zachary Berger, Parth Agrawal, Tian Yu Liu, Stefano Soatto, and Alex Wong. Stereoscopic universal perturbations across different architectures and datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2022.

[9] Qi-Zhi Cai, Min Du, Chang Liu, and Dawn Song. Curriculum adversarial training. *arXiv preprint arXiv:1805.04807*, 2018.

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[11] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018.

[12] Hao Chen, Pramod K Varshney, Steven M Kay, and James H Michels. Theory of the stochastic resonance effect in signal detection: Part i—fixed detectors. *IEEE transactions on Signal Processing*, 55(7):3172–3184, 2007.

[13] Kelvin Cheng, Christopher Healey, and Tianfu Wu. Towards adversarially robust and domain generalizable stereo matching by rethinking dnn feature backbones. *arXiv preprint arXiv:2108.00335*, 2021.

[14] Kelvin Cheng, Tianfu Wu, and Christopher Healey. Revisiting non-parametric matching cost volumes for robust and generalizable stereo matching. *Advances in Neural Information Processing Systems*, 35:16305–16318, 2022.

[15] Gilad Cohen and Raja Giryes. Simple post-training robustness using test time augmentations and random forest. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3996–4006, 2024.

[16] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[18] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[20] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

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

[22] Ruxiao Duan, Erin Hong, Eric Turner, Alex Wong, and Yunwen Zhou. Fisheye3r: Adapting unified 3d feed-forward foundation models to fisheye lenses. In *European Conference on Computer Vision*. Springer, 2026.

[23] Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W Taylor. Batch normalization is a cause of adversarial vulnerability. *arXiv preprint arXiv:1905.02161*, 2019.

[24] Luca Gammaitoni, Peter Hänggi, Peter Jung, and Fabio Marchesoni. Stochastic resonance. *Reviews of modern physics*, 70(1):223, 1998.

[25] Suchisrit Gangopadhyay, Jung-Hee Kim, Xien Chen, Patrick Rim, Hyoungseob Park, and Alex Wong. Extending foundational monocular depth estimators to fisheye cameras with calibration tokens. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5198–5209. IEEE, 2025.

[26] Suchisrit Gangopadhyay and Alex Wong. From perspective to fisheye depth estimation and open-vocabulary segmentation. In *European Conference on Computer Vision*. Springer, 2026.

[27] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[28] Sajjad Ghiasvand, Amirhossein Reisizadeh, Mahnoosh Alizadeh, and Ramtin Pedarsani. Robust decentralized learning with local updates and gradient tracking. *arXiv preprint arXiv:2405.00965*, 2024.

[29] James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 11(3):227–235, 1978.

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

[31] Bhavna Gopal, Huanrui Yang, Jingyang Zhang, Mark Horton, and Yiran Chen. Boosting adversarial robustness with clat: Criticality leveraged adversarial training. In *Forty-second International Conference on Machine Learning*, 2025.

[32] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer, 2022.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[34] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[35] Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. Enhancing adversarial training with second-order statistics of weights. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15273–15283, 2022.

[36] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[37] Woo Jae Kim, Yoonki Cho, Junsik Jung, and Sung-Eui Yoon. Feature separation and recalibration for adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8183–8192, 2023.

[38] Bart Kosko and Sanya Mitaim. Robust stochastic resonance: Signal detection and adaptation in impulsive noise. *Physical review E*, 64(5):051110, 2001.

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

[41] Dong Lao, Congli Wang, Alex Wong, and Stefano Soatto. Diffeomorphic template registration for atmospheric turbulence mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25107–25116, 2024.

[42] Dong Lao, Yangchao Wu, Tian Yu Liu, Alex Wong, and Stefano Soatto. Sub-token vit embedding via stochastic resonance transformers. In *International Conference on Machine Learning*, pages 25995–26006. PMLR, 2024.

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[45] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[46] Divyam Madaan, Jinwoo Shin, and Sung Ju Hwang. Adversarial neural pruning with latent vulnerability suppression. In *International conference on machine learning*, pages 6575–6585. PMLR, 2020.

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

[48] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[49] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[50] Sen Nie, Jie Zhang, Zhuo Wang, Shiguang Shan, and Xilin Chen. Contrastive spectral rectification: Test-time defense towards zero-shot adversarial robustness of clip. *arXiv preprint arXiv:2601.19210*, 2026.

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[52] Haniyeh Ehsani Oskouie, Mohammad-Shahram Moin, and Shohreh Kasaei. Attack on scene flow using point clouds. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2024.

[53] Juan C Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 81–91, 2021.

[54] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[56] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J. Black. Attacking optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[57] Erik Scheurer, Jenny Schmalfuss, Alexander Lis, and Andrés Bruhn. Detection defenses: An empty promise against adversarial patch attacks on optical flow. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6489–6498, 2024.

[58] Jenny Schmalfuss, Lukas Mehl, and Andrés Bruhn. Distracting downpour: Adversarial weather attacks for motion estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10106–10116, 2023.

[59] Jenny Schmalfuss, Philipp Scholze, and Andrés Bruhn. A perturbation-constrained adversarial attack for evaluating the robustness of optical flow. In *European Conference on Computer Vision*, pages 183–200. Springer, 2022.

[60] Simon Schrodi, Tonmoy Saikia, and Thomas Brox. Towards understanding adversarial robustness of optical flow networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8916–8924, 2022.

[61] Ji Shu-Yao, Yuan Fei, Chen Ke-Yu, and Cheng En. Application of stochastic resonance technology in underwater acoustic weak signal detection. In *OCEANS 2016-Shanghai*, pages 1–5. IEEE, 2016.

[62] Nigel G Stocks, D Allingham, and Robert P Morse. The application of suprathreshold stochastic resonance to cochlear implant coding. *Fluctuation and noise letters*, 2(03):L169–L181, 2002.

[63] Linyu Tang and Lei Zhang. Robust overfitting does matter: Test-time adversarial purification with fgsm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24347–24356, 2024.

[64] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[65] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[66] Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19921–19930. IEEE, 2025.

[67] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

[68] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

[69] Yun-Yun Tsai, Ju-Chin Chao, Albert Wen, Zhaoyuan Yang, Chengzhi Mao, Tapan Shah, and Junfeng Yang. Test-time detection and repair of adversarial samples via masked autoencoder. *arXiv preprint arXiv:2303.12848*, 2023.

[70] Jumpei Ukita and Kenichi Ohki. Adversarial attacks and defenses using feature-space stochasticity. *Neural Networks*, 167:875–889, 2023.

[71] Dequan Wang, An Ju, Evan Shelhamer, David Wagner, and Trevor Darrell. Fighting gradients with gradients: Dynamic defenses against adversarial attacks. *arXiv preprint arXiv:2105.08714*, 2021.

[72] Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, and Zhangyang Wang. Removing batch normalization boosts adversarial training. In *International Conference on Machine Learning*, pages 23433–23445. PMLR, 2022.

[73] Hongjun Wang and Yisen Wang. Self-ensemble adversarial training for improved robustness. *arXiv preprint arXiv:2203.09678*, 2022.

[74] Jun Wang, Qingbo He, and Fanrang Kong. Adaptive multiscale noise tuning stochastic resonance for health diagnosis of rolling element bearings. *IEEE Transactions on instrumentation and measurement*, 64(2):564–577, 2014.

[75] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019.

[76] Thomas Wellens, Vyacheslav Shatokhin, and Andreas Buchleitner. Stochastic resonance. *Reports on progress in physics*, 67(1):45, 2003.

[77] Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2879–2888, 2021.

[78] Yangchao Wu, Tian Yu Liu, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Augundo: Scaling up augmentations for monocular depth completion and estimation. In *European Conference on Computer Vision*, pages 274–293. Springer, 2024.

[79] Chang Xiao, Peilin Zhong, and Changxi Zheng. Enhancing adversarial defense by k-winners-take-all. *arXiv preprint arXiv:1905.10510*, 2019.

[80] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 501–509, 2019.

[81] Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15172–15182. IEEE, 2025.

[82] Xilie Xu, Jingfeng Zhang, and Mohan Kankanhalli. Autolora: An automated robust fine-tuning framework. In *The Twelfth International Conference on Learning Representations*, 2024.

[83] Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Tan, and Masashi Sugiyama. Cifs: Improving adversarial robustness of cnns via channel-wise importance-based feature selection. In *International Conference on Machine Learning*, pages 11693–11703. PMLR, 2021.

[84] Cheng-Han Yeh, Kuanchun Yu, and Chun-Shien Lu. Test-time adversarial defense with opposite adversarial path and high attack time cost. *arXiv preprint arXiv:2410.16805*, 2024.

[85] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[86] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[87] Howard Zhang, Yunhao Ba, Ethan Yang, Varan Mehra, Blake Gella, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Alex Wong, and Achuta Kadambi. Weatherstream: Light transport automation of single image deweathering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13499–13509, 2023.

[88] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, pages 11278–11287. PMLR, 2020.

[89] Lingyu Zhang, Chengzhi Mao, Junfeng Yang, and Carl Vondrick. Adversarially robust video perception by seeing motion. *arXiv preprint arXiv:2212.07815*, 2022.

[90] Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. Towards robust image classification using sequential attention models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9483–9492, 2020.

A Additional Theoretical Foundation for Stochastic Resonance

The theoretical foundations of Stochastic Resonance have been long established in statistical physics, summarised for instance in [24] . Our work applies the concept to trained embeddings, but we note that, in the context of adversarial perturbations, modern deep networks often exhibit poor local Lipschitz regularity, where small input perturbations induce disproportionately large latent-space changes, exploited by adversaries.

If x is a clean signal (an image in our case) and $\tilde{x} = x + n(x)$ its adversarially perturbed counterpart. Let φ denote the encoder up to the SR layer. Using a discretized set of translations $\{g_i\}_{i=1}^{N} \subset G$ and their corresponding inverse g_i^{-1}, the SR feature is:

$$\hat{\varphi}(x) = \frac{1}{N}\sum_{i=1}^{N} g_i^{-1}\varphi(g_i x), \qquad \hat{\varphi}(\tilde{x}) = \frac{1}{N}\sum_{i=1}^{N} g_i^{-1}\varphi(g_i \tilde{x})$$

The first-order approximation yields:

$$\varphi(g_i\tilde{x}) = \varphi(g_i x + g_i n(x)) \approx \varphi(g_i x) + J_{g_i x}[g_i n(x)],$$

where $J_{g_i x}$ is the Jacobian of φ at $g_i x$. We have

$$\Delta_{\mathrm{SR}}(x) = \hat{\varphi}(\tilde{x}) - \hat{\varphi}(x) \approx \frac{1}{N}\sum_{i=1}^{N} g_i^{-1} J_{g_i x}[g_i n(x)].$$

In contrast, a standard encoder responds as:

$$\Delta_{\mathrm{base}}(x) = \varphi(\tilde{x}) - \varphi(x) \approx J_x n(x),$$

Here $J_x n(x)$ is the first-order response of the encoder to the perturbation $n(x)$. For stochastic resonance, the corresponding terms $v_i = A_{g_i}\big(J_{g_i x}[g_i n(x)]\big)$ in (2) arise from the same image perturbation, evaluated under different translations and aligned to the original image coordinates. Taking the expectation over inputs and their corresponding adversarial perturbations gives

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N} v_i\right\|^2 = \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\|v_i\|^2 + \sum_{i\neq j}\mathbb{E}\langle v_i, v_j\rangle\right).$$

If the second sum is small in magnitude relative to the first, the root-mean-square magnitude of the averaged first-order response is approximately $1/\sqrt{N}$ times the root-mean-square magnitude of the individual terms v_i across the chosen translations. This interpretation assumes that the local first-order expansion is accurate at the perturbation magnitude under consideration. A worst-case Lipschitz bound would require control over all perturbation directions.

To examine alignment empirically, we report the mean pairwise correlation of the aligned perturbation responses at each layer of a pre-trained ViT-S on ImageNet under adaptive PGD-100, where the attack differentiates through all translations (Table 7).

The mean pairwise correlations range from 0.0897 to 0.4056 across the ViT-S layers tested. These measurements characterize alignment in this setting. Verifying the condition above would additionally require measuring the aggregate cross term relative to the summed squared norms. In a separate experiment with a pre-trained DINO encoder, the L_1 distance between adversarial and clean embeddings decreases from 12.30 at $d = 0$ to 2.12 at $d = 4$ (Table 8), showing that the measured feature distortion decreases as the translation radius increases.

B Empirical Analysis for Robustness and Inference Time Trade-off

We analyze the trade-off between the inference time and robustness incurred by our methods empirically. We note that the specific trend depends on the task, the architecture, the evaluation metric,

Table 7: Mean pairwise correlation of the aligned perturbation terms v_i across layers of a ViT-S model under an adaptive PGD-100 attack on ImageNet. The values remain far from perfect alignment, supporting the decorrelation assumption used in the analysis.

Layer	Mean Correlation
`conv_proj`	0.0897
`block_00`	0.2107
`block_01`	0.2518
`block_02`	0.3311
`block_03`	0.3712
`block_04`	0.3781
`block_05`	0.3083
`block_06`	0.3484
`block_07`	0.3683
`block_08`	0.3784
`block_09`	0.4056
`block_10`	0.3670
`block_11`	0.3624

Table 8: L_1 distance between adversarial and clean DINO embeddings for different translation radius d increases.

Translation Radius d	0	1	2	3	4
L_1 Distance (Adv. vs. Clean)	12.30	5.29	3.50	2.64	2.12

and the attack type. Below, we present a consolidated analysis using **ResNet-50** under different translation radii d, illustrating how computational budget and adversarial robustness co-vary.

Table 9: Image Classification performance and computational cost on ImageNet with different translation radius d.

Metric	$d = 0$	$d = 1$	$d = 2$	$d = 3$
# Forward Passes	1	9	25	49
Wall-clock Time (s)	0.095	0.114	0.133	0.158
SR Forward Time (excluding interpolation)	0.007	0.010	0.019	0.028
Top-1 Acc @ $\epsilon = 8/255$	8.51	18.40	20.01	21.17
Top-1 Acc @ $\epsilon = 4/255$	9.45	19.54	21.02	22.04
Top-1 Acc @ $\epsilon = 2/255$	11.10	21.02	22.44	23.44
Top-1 Acc @ $\epsilon = 1/255$	15.21	24.66	25.80	26.58

Even at the most efficient setting $(d = 1)$, the improvement in robustness is substantial. Users can choose the desired operating point based on latency constraints and may also selectively sample transformations to further reduce computation. Notably, a significant portion of the overhead comes from translating images and performing interpolation via Python-level loops in PyTorch. The pure computation time of the network components under SR is reported as **"SR Forward Time,"** demonstrating that the method itself is efficient. Future work will focus on implementing these transformations in a custom CUDA kernel for further speedups

C Memory Footprint for Our Method

We present the memory footprint for our method using **ResNet-50** under different translation radii d, evaluated on both CIFAR-sized inputs (32×32) and ImageNet-sized inputs (224×224). All measurements are obtained using `torch.cuda.max_memory_allocated(device)` to capture peak memory footprint. This illustrates how computational budget scales with SR depth.

Table 10: Memory footprint of ResNet-50 on different size input with different degree of translation radii d

Metric	$d = 0$	$d = 1$	$d = 2$	$d = 3$
# Forward Passes (can be parallelized)	1	9	25	49
Memory (CIFAR)	114.50 MB	141.50 MB	182.93 MB	260.50 MB
Memory (ImageNet)	157.80 MB	225.19 MB	384.03 MB	650.47 MB

As shown, modern GPUs have ample memory headroom, and SR's memory growth under normal inference remains modest. The more interesting observation arises during *adversarial attack generation* on one single 224×224 inputs: memory usage increases dramatically, from **346.19 MB** for standard inference to **10,649 MB** when performing gradient-based adaptive attacks through multi-branch SR. This exponential increase makes strong adaptive attacks ("Worst-Case" in Table 1) extremely difficult in practice, highlighting an important defensive advantage of SR.

D Clean Performance

A potential concern with test-time defenses is that robustness may come at the expense of clean performance, particularly when the defense removes or suppresses high-frequency information. Stochastic resonance differs from smoothing-based methods in that it does not perform spatial averaging on the input image. Instead, it applies small transformations, aligns the corresponding latent embeddings, and ensembles them in feature space. As a result, it mitigates adversarial perturbations without explicitly discarding image content.

We evaluate clean performance on ImageNet using ViT-S and on CIFAR-10 using ResNet-50. As shown in Tab. 11, stochastic resonance preserves, and slightly improves, clean accuracy across both datasets. This suggests that the robustness gains are not obtained by degrading the clean signal. The strong clean-task performance observed in dense prediction experiments, including optical flow and depth estimation, further supports that high-frequency structure useful for prediction is retained.

Table 11: **Clean performance.** Stochastic resonance preserves and slightly improves clean accuracy on both ImageNet and CIFAR-10, indicating that the method does not trade clean performance for robustness.

Dataset / Metric	No Defense	SR ($d = 2$)	SR ($d = 3$)
ImageNet Top-1	74.90	75.09	**75.16**
ImageNet Top-5	92.21	92.30	**92.42**
CIFAR-10 Accuracy	93.26	93.50	**93.64**

E Additional Ablations

E.1 Layer-wise ablation on ResNet-50

We further study where stochastic resonance should be applied within a convolutional backbone. Since adversarial perturbations often behave like high-frequency artifacts, one may expect that early layers play a particularly important role: they are closest to the image lattice, preserve fine spatial structure, and determine how local perturbations are propagated into deeper semantic representations. At the same time, applying stochastic resonance too early or too late leads to different computational and representational trade-offs. Early aggregation is cheaper, while deeper aggregation acts on more semantic features but requires more repeated computation.

To examine this trade-off, we apply stochastic resonance to different termination layers of ResNet-50 on ImageNet under PGD attacks with varying perturbation strengths. As shown in Tab. 12, applying stochastic resonance only through the initial convolution already improves robustness over the undefended model across all attack strengths. The gain becomes substantially larger when stochastic resonance is extended through the first residual block, indicating that a shallow feature ensemble is already sufficient to suppress a large fraction of the attack-induced distortion. The best performance is obtained when aggregating after the second residual block, improving Top-1 accuracy from 8.51 to

Table 12: **Layer-wise ablation on ResNet-50.** Adversarial perturbations resemble high-frequency noise. Applying stochastic resonance through shallow layers is sufficient to defend against adversarial attacks, reducing the overall computational cost.

Att. Strength (ϵ)	8/255		4/255		2/255		1/255	
Metric	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
No Defense	8.51	26.19	9.45	27.92	11.10	30.63	15.21	36.76
Initial Conv.	16.91	34.66	17.36	35.28	18.12	36.30	19.59	38.25
Res. Block 1	22.44	44.95	23.39	46.03	24.84	47.64	27.57	50.80
Res. Block 2	**24.88**	**48.54**	**25.90**	**49.66**	**27.23**	**51.19**	**30.11**	**54.15**
Res. Block 3	21.76	44.28	22.73	45.47	24.14	47.14	27.18	50.20

24.88 under $\epsilon = 8/255$, and from 15.21 to 30.11 under $\epsilon = 1/255$. The same trend holds for Top-5 accuracy.

Interestingly, extending stochastic resonance further to the third residual block does not continue to improve performance. This suggests that the benefit of stochastic resonance is not simply a function of applying it deeper in the network. Instead, there is a favorable intermediate regime where the representation is sufficiently structured to support stable alignment across translated inputs, while still remaining close enough to the spatial signal to attenuate high-frequency adversarial artifacts. Beyond this point, deeper features become increasingly semantic and less spatially equivariant, making inverse-aligned aggregation less effective.

This ablation supports two observations. First, stochastic resonance does not need to be applied through the full network to provide strong adversarial robustness. Second, the strongest setting for ResNet-50 occurs at a relatively shallow layer, namely Residual Block 2, which substantially reduces the computational overhead compared to ensembling the entire model. This makes the method practical as a test-time defense: most of the repeated computation is confined to the front end of the network, while the remaining layers are evaluated only once on the aggregated feature.

E.2 Rotation as Purposeful Perturbations

Our main experiments use integer-pixel translations as the purposeful perturbations for stochastic resonance. This is a design choice within a broader principle: the method can be instantiated with different transformation groups, provided that the transformed features can be mapped back and aggregated in a common reference frame. In this section, we evaluate rotations as an alternative choice and explain why translations are used in the main method.

Tab. 13 compares translations and rotations on ImageNet with ResNet-50 under PGD attacks. Rotations improve robustness over the undefended model across all attack strengths and resonance levels, showing that the stochastic-resonance mechanism is not specific to translations. However, translations provide stronger robustness at higher resonance levels and are also computationally cheaper, since rotations require interpolation. Thus, rotations constitute a valid but less effective instantiation of the same principle, rather than a failure mode of the method.

Table 13: **Translation vs. rotation as purposeful perturbations.** On ResNet-50, rotations improve robustness relative to the undefended model across all attack strengths and resonance levels. However, translations are more effective at higher resonance levels and avoid the interpolation overhead required by rotations. This supports our use of translations as the default instantiation of stochastic resonance.

Att. Strength (ϵ)	8/255		4/255		2/255		1/255	
Metric	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
No Defense	8.51	26.19	9.45	27.92	11.10	30.63	15.21	36.76
Ours ($d = 1$)	18.40	40.36	19.54	42.65	21.02	43.86	24.66	47.81
w/ Rotation	20.01	41.32	20.84	42.39	22.10	43.89	24.72	46.89
Ours ($d = 2$)	20.01	42.22	21.02	43.45	22.44	45.32	25.80	48.97
w/ Rotation	18.86	38.67	19.55	39.50	20.36	40.48	22.01	42.50
Ours ($d = 3$)	**21.17**	**43.58**	**22.04**	**44.78**	**23.44**	**46.53**	**26.58**	**49.73**
w/ Rotation	15.15	32.89	15.61	33.46	16.17	34.19	17.33	35.59

The performance gap between translations and rotations can arise from two related factors. First, convolutional backbones are naturally more compatible with translations than rotations: convolutional

filters are spatially shared, but they are not inherently rotation-equivariant. Second, rotating an image on a discrete pixel grid requires interpolation, which can introduce additional artifacts before features are extracted. These two effects are difficult to disentangle completely, but both predict weaker feature alignment for rotations. To examine this directly, we measure the alignment of transformed features after applying the inverse transformation in latent space. Specifically, for each residual block of ResNet-50, we compare the original feature with the inverse-aligned feature obtained after either translation or rotation. We report both Pearson correlation and cosine similarity in Tab. 14. Across all residual blocks, translated features are more strongly aligned than rotated features. This supports the interpretation that rotations provide weaker defense performance because their inverse-aligned latent features are less consistent before aggregation.

Table 14: **Feature alignment under translations and rotations.** We compare features after each residual block of ResNet-50, after applying the inverse transformation in latent space. Translations yield consistently higher correlation and cosine similarity than rotations, indicating better feature alignment after inversion. This explains why rotations still improve robustness relative to no defense, but are less effective than translations.

Layer	Corr. (Trans.)	Corr. (Rot.)	CosSim (Trans.)	CosSim (Rot.)
Block 1	0.7540	0.6487	0.8357	0.7739
Block 2	0.5933	0.4569	0.6756	0.5766
Block 3	0.6752	0.5095	0.7771	0.6756
Block 4	0.8808	0.7719	0.8954	0.8032

These results clarify the role of rotations in our study. The relevant comparison is not whether rotations outperform translations, but whether the stochastic-resonance principle remains meaningful under a different family of purposeful perturbations. The answer is positive: rotations improve robustness over the undefended baseline. At the same time, the lower feature-alignment scores in Tab. 14 explain why translations are the preferred design choice for convolutional architectures. Consequently, we use translations throughout the main experiments because they instantiate the same principle more effectively, preserve latent alignment more reliably, and avoid the interpolation overhead associated with rotations.

F MLP-Mixer

To test whether stochastic resonance depends on convolutional translation equivariance, we evaluate MLP-Mixer on CIFAR-10. We initialize the model with ImageNet pre-trained weights, fine-tune a linear head on CIFAR-10, and apply stochastic resonance at the layer before the prediction head. As shown in Tab. 15, stochastic resonance preserves clean accuracy while substantially improving robustness under both FGSM and PGD attacks. In particular, under PGD with $\epsilon = 8/255$, accuracy improves from 0.03 to 65.38 with $d = 3$. These results suggest that while equivariance under the chosen transformations can improve feature alignment, it is not strictly required for stochastic resonance to suppress attack-induced feature distortion.

Table 15: **Stochastic resonance on MLP-Mixer.** Even without explicit convolutional translation equivariance, stochastic resonance substantially improves robustness under both FGSM and PGD attacks while preserving clean accuracy.

Setting	Clean	FGSM			PGD		
		$2/255$	$4/255$	$8/255$	$2/255$	$4/255$	$8/255$
No Defense	89.13	31.41	24.01	21.76	0.00	0.01	0.03
Ours ($d = 1$)	**89.99**	71.68	71.51	71.55	58.19	57.03	49.87
Ours ($d = 2$)	89.94	73.40	73.30	73.44	69.08	68.14	61.41
Ours ($d = 3$)	89.92	**74.05**	**73.99**	**74.08**	**73.32**	**72.30**	**65.38**

G Implementation for Classification

For our classification experiments, we built our implementation on top of standard network architectures, implementing SR on two main architectures derived from the FSR codebase [37]:

- ResNet-18 [33]: A standard residual network with 18 layers organized in four main blocks with increasing channel dimensions (64, 128, 256, 512).

- WideResNet-34 [85]: A wider variant of ResNet with depth 34 and width factor 10, resulting in higher representational capacity with three main blocks with channel dimensions scaled by the width factor (160, 320, 640).

For both architectures, we apply SR at the bottleneck layer (after the final convolutional block).

Given an input image batch $x \in \mathbb{R}^{b\times 3\times h\times w}$ (where b is the batch size), our SR approach operates as follows. First, we create a set of $(2d_x + 1) \times (2d_y + 1)$ perturbed versions of the input by applying pixel-level translations within the range $[-d_x, d_x] \times [-d_y, d_y]$ pixels:

$$X_{\text{perturbed}} = \{g_{i,j}(X) \mid i \in [-d_y, d_y], j \in [-d_x, d_x]\} \tag{3}$$

where $g_{i,j}$ translates the image by (i, j) pixels. These transformations are applied using PyTorch's "transforms.functional.affine" function with translation parameters while preserving the original image properties.

All images are concatenated into one batch and processed through the network in parallel up to the bottleneck layer:

$$F = \phi(X_{\text{perturbed}}) \tag{4}$$

where ϕ represents the network up to the bottleneck layer. This batch processing approach significantly improves computational efficiency compared to individual forward passes.

After obtaining feature maps for all perturbed inputs, we aggregate them to create a single enhanced feature map:

$$F_{\text{ensembled}} = \frac{1}{n} \sum_{i,j} T_{-i,-j}(F_{i,j}) \tag{5}$$

where $T_{-i,-j}$ represents the inverse spatial shift operation that realigns the feature map and n the number of augmentations.

Our implementation requires $(2d_x + 1) \times (2d_y + 1)$ forward passes through the network up to the bottleneck layer.

For evaluation, we tested our approach against standard adversarial attacks (FGSM [30], PGD-20 and PGD-100 [47], and C&W [10]), all bounded within $\epsilon = 8/255$ under ℓ_∞-norm. We also report an Ensemble metric that measures the worst-case performance across all attacks for each test example, providing a comprehensive robustness assessment.

H Implementation for Stereo Matching

For our stereo matching experiments, we built our implementation on top of standard stereo network architectures to ensure our approach remains model-agnostic and requires no training or fine-tuning. We integrated SR with PSMNet [11], a pyramid stereo matching network with a stacked hourglass architecture that uses 3D convolutions on a cost volume constructed by concatenating features. We apply SR at the feature extraction stage, before cost volume construction, where stereo correspondences are first established.

Given a pair of input stereo images $x_L, x_R \in \mathbb{R}^{b\times 3\times h\times w}$ (where b is the batch size), our SR approach operates as follows. First, we create a set of $(2d_x + 1) \times (2d_y + 1)$ perturbed versions of each input image by applying translations within the range $[-d_x, d_x] \times [-d_y, d_y]$ pixels:

$$X_{L,\text{perturbed}} = \{g_{i,j}(x_L) \mid i \in [-d_y, d_y], j \in [-d_x, d_x]\} \tag{6}$$

$$X_{R,\text{perturbed}} = \{g_{i,j}(x_R) \mid i \in [-d_y, d_y], j \in [-d_x, d_x]\}. \tag{7}$$

All images are concatenated into batches and processed through the feature extraction component of the network:

$$F_L = \phi(X_{L,\text{perturbed}}) \tag{8}$$

$$F_R = \phi(X_{R,\text{perturbed}}) \tag{9}$$

where ϕ represents the feature extraction component of the stereo network. This batch processing approach significantly improves computational efficiency compared to individual forward passes.

After obtaining feature maps for all perturbed inputs, we aggregate them to create a single enhanced feature map.

Our implementation requires $(2d_x + 1) \times (2d_y + 1)$ forward passes through the feature extraction component of the network for each stereo image.

For evaluation, we tested our approach against adversarial attacks generated using FGSM [30] and I-FGSM [40] (a special case of PGD), bounded within various ϵ values ($\{0.002, 0.005, 0.01, 0.02\}$) under ℓ_∞-norm. We measured performance using three standard stereo matching metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and D1-error (percentage of pixels with disparity error greater than 3 pixels or 5% of the ground truth) [45].

I Details about attack on Monocular Depth Estimation

We evaluate 648 images from the NYU-V2 validation dataset. For attacks, we performed PGD attacks with $\epsilon = 8/255$, $\alpha = 2/255$ for 10 iterations with an L1 loss between the predicted depth and the ground truth depth. We obtain the DINOv2 features using DINOv2 backbone and perform the aggregation with these features before passing the aggregated features into the depth prediction head. We used a linear head pretrained by DINOv2 to generate the final depth prediction[2].

J Details about attack on vision language alignment

We split the COCO-Caption[3] validation set into the first 500 instances for evaluation and the remaining is used for generating adversarial samples for the adversarial training. For the CLIP model, we used the same CLIP model used for building the vision tower of LLaVA-13B, which is a vision transfomer with patch size 14 and input image size of 336 by 336[4]. Since CLIP aligns the feature output for the [CLS] token with the text embedding, we extracted spatial feature from the second last layer of the vision transformer to perform aggregation and use the last layer as it is to generate the final image embedding used for vision language alignment. We obtain the ground truth text embedding by averaging the embedding from 5 ground truth text caption for each image provided by the dataset. For attack, we use PGD attack with $\epsilon = 0.1$, $\alpha = 0.02$, and 10 iterations with a cosine similarity lost between the CLIP image embedding and the ground truth text embedding.

K Details about Attack on Optical Flow

To find an adversarial attack for optical flow estimated by a given neural network f, we adapt the formulation proposed by [52]. Specifically, we aim to find a perturbation δ for given frames F_1 and F_2, maximizing the discrepancy between predicted and ground-truth optical flow OF. If the ground-truth optical flow is unavailable, we use the predicted optical flow from the unattacked frame as our surrogate ground-truth. Our method applies δ to the first input frame, then uses a deep neural network to estimate optical flow from the perturbed frames. The objective is to maximize the average end-point error (EPE) between the predicted and ground-truth optical flow, calculated as the mean Euclidean distance between corresponding 2D flow vectors. In other words, the ϵ-norm bounded adversary δ for optical-flow is calculated by optimizing the following

$$\max_{\delta:\|\delta\|\leq\epsilon} \text{EPE}\big(OF, f(F_1 + \delta, F_2)\big). \tag{10}$$

One l_∞-bounded adversary A for the aforementioned optimization problem is Fast Gradient Sign Method (FGSM) [30] which can be obtained by

$$\begin{aligned} \text{L} &= \text{EPE}(OF, f(F_1, F_2)), \\ A &= F_1 + \epsilon \cdot \text{sign}\big(\nabla_{F_1}\text{L}\big). \end{aligned} \tag{11}$$

Projected gradient descent (PGD) [47] represents an enhanced and more complex version of FGSM. This attack method generates adversarial examples through an iterative process and the formulation for this attack is as following

$$F_1^{(t+1)} = \Pi_{F_1+\mathcal{S}}\big(F_1^{(t)} + \alpha \cdot \text{sign}(\nabla_{F_1}\text{L})\big). \tag{12}$$

[2] https://github.com/facebookresearch/dinov2#pretrained-heads---depth-estimation

[3] https://huggingface.co/datasets/lmms-lab/COCO-Caption

[4] https://huggingface.co/openai/clip-vit-large-patch14-336

Since adversarial perturbations are constrained to be imperceptible, we assume they are too small to affect the actual physical motion between frames. Thus, we treat the ground-truth optical flow as invariant and do not update it during the optimization process.

In our experimental setup, we set the norm value to $\epsilon = \frac{10}{256}$. To generate the adversarial perturbations, we employed a 10-step PGD attack. We calculated the step size α as 2.5ϵ divided by the iteration count, ensuring a robust traversal of the ϵ-ball during optimization.

L Feature-Space Attack and Defense

Most adversarial attacks considered in this paper operate in pixel space and constrain the perturbation magnitude to remain visually imperceptible. Another related line of work studies attacks and defenses based on feature-space stochasticity [70]. This setting is conceptually different from ours: the feature-space attack in [70] can substantially alter the input appearance and is not constrained by the same small pixel-space perturbation assumption used in standard adversarial robustness benchmarks. Nevertheless, we include this comparison to test whether stochastic resonance can be used together with feature-space defenses and whether it remains useful when the attack mechanism differs from the pixel-space attacks used in the main paper.

We evaluate this setting on CIFAR-10. First, we apply the feature-space stochastic defense of [70] and then combine it with stochastic resonance. As shown in Tab. 16, stochastic resonance provides an additional improvement when used together with the feature-space defense. This indicates that our method is complementary to existing feature-space defenses, since it operates by aggregating aligned latent responses at test time rather than modifying the training objective or replacing the defense mechanism.

Table 16: **Combination with feature-space defense.** Stochastic resonance further improves robustness when combined with the feature-space stochastic defense of [70], suggesting that the two mechanisms are complementary.

Setting	Accuracy (%) ↑
Attacked	1.72
Feature-space defense [70]	8.13
Feature-space defense [70] + Ours	**8.56**

We also evaluate a constrained variant of the feature-space attack, where the perturbation magnitude is bounded to better match the small-perturbation setting considered in this paper. In this setting, stochastic resonance alone achieves higher accuracy than the feature-space defense baseline, as shown in Tab. 17. This result suggests that even when the attack is generated through feature-space stochasticity, the latent ensembling induced by stochastic resonance still suppresses a portion of the attack-induced feature distortion.

Table 17: **Constrained feature-space attack.** When the feature-space attack is constrained by a maximum perturbation magnitude, stochastic resonance alone provides a reasonable level of robustness and outperforms the feature-space defense baseline in this setting.

Setting	Accuracy (%) ↑
Attacked	2.14
Feature-space defense [70]	11.09
Ours only	**12.36**

Overall, these results show that stochastic resonance is not restricted to the pixel-space attack mechanisms evaluated in the main paper. While our method is designed for standard imperceptible pixel-space adversarial perturbations, it can also be combined with feature-space defenses and remains effective under a constrained feature-space attack. This supports the attack-agnostic nature of stochastic resonance as a test-time defense.

M Boundary Handling

Our implementation uses zero filling for samples outside the image grid. To assess sensitivity to this choice, we evaluated five alternative boundary treatments under PGD-100 on CIFAR-10 using the

adversarially trained ResNet-18 from Table 1. Robust accuracy was 54.29% with reflection, 54.56% with replication, 54.55% with circular extension, 54.40% with valid masking, and 55.24% with valid cropping. The five treatments differed by at most 0.95 percentage points in this setting.

N Additional Visualizations

Here we provide additional visualizations from our experiments comparing SR under both FGSM and PGD attack. We also provide a visual showing the results of FGSM and PGD pertubation on various images.

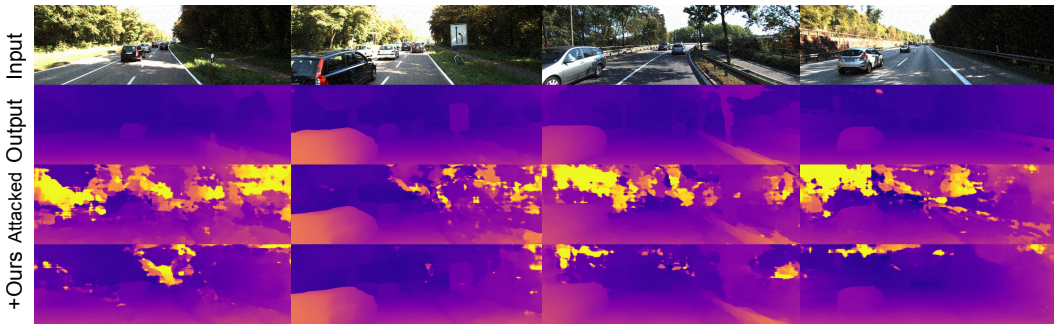

Figure 6: Visual results on stereo matching against FGSM attack, without and with SR, $\epsilon = 0.02$

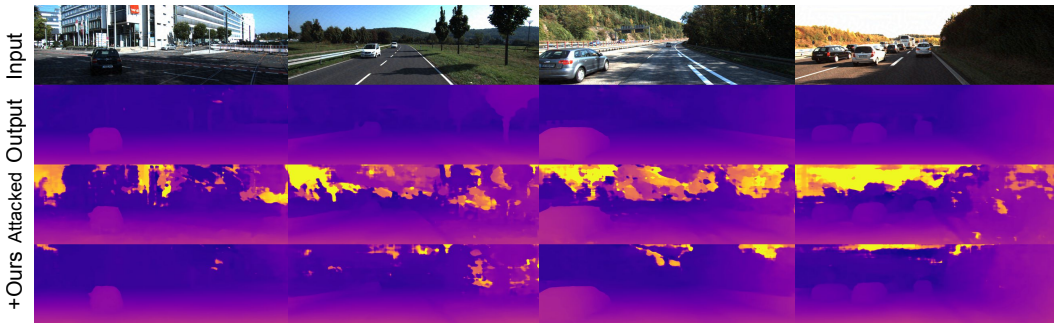

Figure 7: Visual results on stereo matching against FGSM attack, without and with SR, $\epsilon = 0.01$



Figure 8: Visual results on stereo matching against FGSM attack, without and with SR, $\epsilon = 0.005$

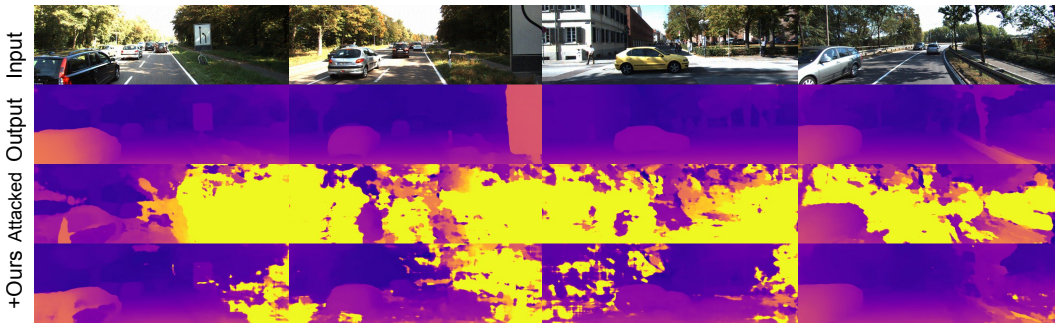

Figure 9: Visual results on stereo matching against PGD attack, without and with SR, $\epsilon = 0.01$

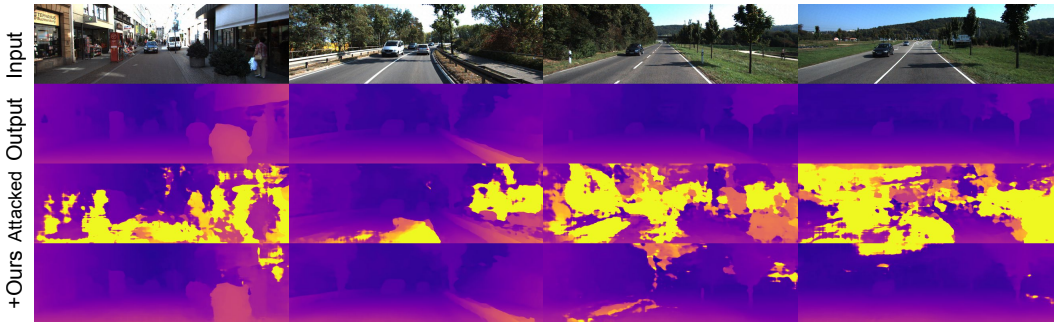

Figure 10: Visual results on stereo matching against PGD attack, without and with SR, $\epsilon = 0.005$

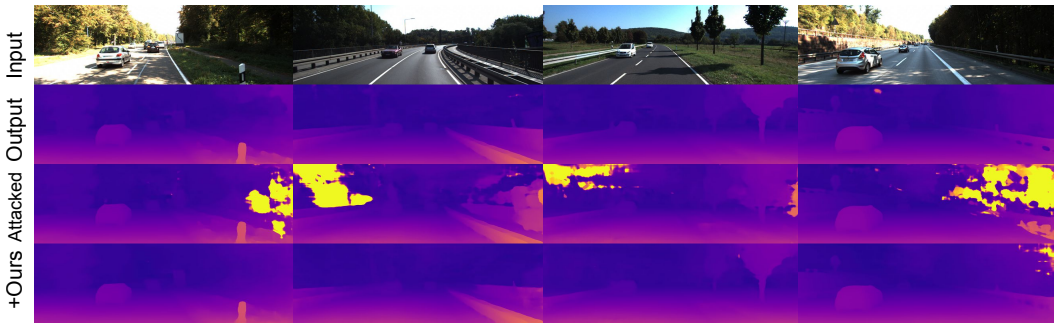

Figure 11: Visual results on stereo matching against PGD attack, without and with SR, $\epsilon = 0.002$

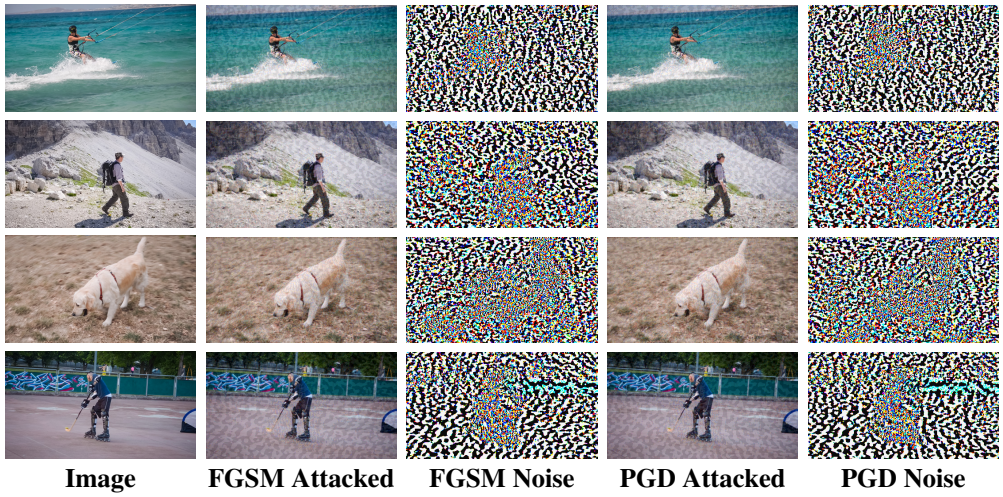

Figure 12: Original images and their corresponding attacked images and perturbations using FGSM and PGD methods on optical flow. The attacks mostly target the main object observed in the image.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims match the proposed training-free stochastic resonance defense and its evaluated scope.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss compute, memory, latency, and adaptive budget limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide assumptions, derivations, and empirical validation for the SR effect.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We report datasets, models, attacks, metrics, and implementation details.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is publicly available.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the settings across all experiments.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Given the scope of the experiments, We do not report error bars or statistical significance tests.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the relevant compute and memory requirements.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Yes.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We improve robustness and overall performance, with no societal impacts beyond those typical of standard machine learning research.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No special release safeguards are applicable.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite existing open-source datasets and models whose licenses permit use for this research.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are introduced.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human-subject research is involved.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: IRB approval is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM use is not part of the core methodology.