

# TEST-TIME DEFENSE AGAINST ADVERSARIAL ATTACKS VIA STOCHASTIC RESONANCE OF LATENT ENSEMBLES

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

We propose a test-time defense mechanism against adversarial attacks: imperceptible image perturbations that significantly alter the predictions of a model. 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. Empirical results show state-of-the-art robustness on image classification and, for the first time, establish a generic test-time defense for dense prediction tasks, including stereo matching and optical flow, highlighting the method’s versatility and practicality. Specifically, relative to clean (unperturbed) performance, our method recovers up to 68.1% of the accuracy loss on image classification, 71.9% on stereo matching, and 29.2% on optical flow under various types of adversarial attacks.

## 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 Goodfellow et al. (2014). In some cases, the same perturbation can even be disruptive for a large number of possible inputs Moosavi-Dezfooli et al. (2017), exploiting the convoluted geometry of the decision boundary imposed by such trained models Tramèr et al. (2017). 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 Gibson (2014). 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 where the adversary can easily modify the perturbations to bypass the anti-aliasing filter, namely *adaptive* attacks Tramer et al. (2020), and the model needs to constantly be fine-tuned to “anti-alias” new forms of adversarial perturbations. In the context of deep neural networks, existing defense methods that manipulate feature representations Xie et al. (2019); Bai et al. (2021); Yan et al. (2021); Kim et al. (2023) fundamentally adhere to this paradigm in principle. 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 Benzi et al. (1981). It has been used extensively in cochlear implants, where power constraints limit the resolution of the digital circuitry Stocks et al. (2002). 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 Lao et al. (2024). 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 remove the effects of adversarial perturbations in the embedding.

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 equation 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 thought of as marginalizing the translation group in latent space with respect to a chosen prior, which is the only design choice in our method. We choose the simplest, which is the constant prior. The purposeful perturbations alter the spatial sampling, and the implicit ensembling in latent space averages out the effect of sampling artifacts, thwarting the effect of adversarial perturbations. It is as if we were given multiple 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. To measure the effectiveness of our method in mitigating the effects of adversarial perturbations, we test it on three vastly different tasks, including image classification, and two other dense prediction tasks: stereo matching and optical flow. where we are the first to show a significant and consistent improvement in robustness to various adversarial attacks.

One could argue that there is still a cat-and-mouse game in our setting, if the adversary knows our technique and tailors the adversarial perturbations to bypass it. To assess this risk, we conduct “worst-case” adaptive tests to measure the performance of our method under adaptive attacks when the attacker knows the exact defense strategy, thus the adversarial perturbation is designed to maximally disrupt the result end-to-end, *including* our stochastic resonance. Our results show that the method is resistant to breaking even when the adversary optimizes adaptively through it end-to-end.

## 2 RELATED WORK

The literature on adversarial attack and defense is extensive. We highlight some of the advances.

**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) Goodfellow et al. (2014), and Projected Gradient Descent (PGD) Madry et al. (2017). ALP Kannan et al. (2018) Minimizes the difference between the logits of pairs of clean and adversarially augmented images. TRADES Zhang et al. (2019) decomposes prediction error for adversarial examples into natural error and boundary error to improve adversarial robustness. MART Wang et al. (2019) improves adversarial robustness by considering misclassified natural examples

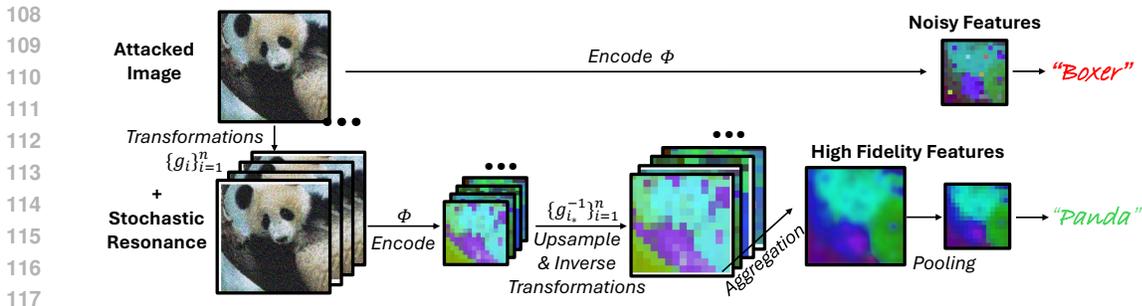


Figure 1: **Defense against adversarial attacks via stochastic resonance.** Neural networks are sensitive to perturbations in the input space, which adversarial attacks exploit to manipulate outputs. Conventional defense primarily focuses on filtering out unreliable features or denoising either the input or the features. Instead of removing noise, we propose a novel defense by introducing noise. 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 adversarial 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.

during training. Subsequent work Cai et al. (2018); Zhang et al. (2020); Wang & Wang (2022); Jin et al. (2022); Ghiasvand et al. (2024); Xu et al. (2024); Gopal et al. (2025) improves adversarial robustness with curriculum learning Bengio et al. (2009), model ensembling, critical parameter selection, and gradient tracking. On the other hand, some methods learn robust feature representation through a modified architecture or feature manipulation. Galloway et al. (2019); Benz et al. (2021); Wang et al. (2022) investigate the effect of batch normalization on adversarial robustness. Dhillon et al. (2018); Madaan et al. (2020) prunes certain activations in the network that are susceptible to adversarial attacks. Xiao et al. (2019) keep k-features with the largest magnitude and deactivate everything else. Zoran et al. (2020) uses an attention mask to highlight robust regions on the feature. Feature Denoising (FD) Xie et al. (2019) uses classical denoising techniques to deactivate abnormal activations. Bai et al. (2021); Yan et al. (2021) proposed Channel Activation Suppression (CAS) and Channel-wise Importance-based Feature Selection (CIFS) to deactivate feature channels that are vulnerable to attacks. Kim et al. (2023) 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. Tang & Zhang (2024); Yeh et al. (2024); Tsai et al. (2023) train an FGSM robust classifier, a diffusion model, or a mask auto-encoder, respectively, to purify adversarial examples. Wang et al. (2021) optimizes both the model and the input to minimize the entropy of model predictions to adapt to changing attacks. Cohen & Giryes (2024) 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. Pérez et al. (2021) 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 Benzi et al. (1981) and first applied in climate dynamics (Benzi et al., 1982) and later in signal processing (Wellens et al., 2003; Kosko & Mitaim, 2001; Chen et al., 2007) and acoustics (Shu-Yao et al., 2016; Wang et al., 2014). While several works Zhang et al. (2022); Ukita & Ohki (2023) introduce stochasticity within feature space, the most closely related prior effort is Lao et al. (2024), 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.

### 3 METHOD

**Notation.** A digital image  $x \in [0, L - 1]^{W \times H}$  can be described as a map from a discrete planar lattice  $\Lambda \subset \mathbb{R}^2$  with  $H$  rows and  $W$  columns to  $L$  discrete levels,  $x : \Lambda \rightarrow [0, L - 1]$ ; a ‘feature’ or ‘embedding’ of an image  $x$  is the output of an encoder  $\phi$  that maps it to a vector space with  $K$  channels, typically through a parametric trained model:

$$\phi : x \mapsto \phi(x) \in \mathbb{R}^K.$$

We represent a group transformation of the image through an operator  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , which can be restricted to the lattice  $\Lambda$  through padding, sampling and quantization at the expense of invertibility:

$$g : x \mapsto g(x) \in \Lambda \subset \mathbb{R}^2.$$

For example, a translation by an integer pixel can be represented by an upper diagonal matrix  $G$ ,  $g(x) = Gx$  with ones above the diagonal. The group  $g$  operating on  $x$  induces an operation on  $\phi$  via

$$g_*\phi(x) \doteq \phi(g(x)).$$

We call the composition of  $\phi$  and  $g$  the encoding of the transformed image

$$\psi(x) \doteq \phi(g(x)) = \phi \circ g(x).$$

The main object of interest in our method is:

$$g_*^{-1}\psi(x) = g_*^{-1} \circ \phi \circ g(x).$$

This is obtained, reading right-to-left, by first transforming the image, then passing it through an encoder, and then transforming back the feature map through the push-forward action  $g_*^{-1}$ .

**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 (adversarial perturbation)  $\phi(\tilde{x})$ :

$$\tilde{x} = x + n(x) \quad | \quad n(x) = \arg \max d(\phi(x), \phi(\tilde{x})), \quad |n| < \epsilon$$

for some small  $\epsilon$  designed so the perturbation is, ideally, imperceptible by humans.

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

**Averaging, smoothing, and stochastic resonance.** The resemblance between adversarial perturbations and aliasing artifacts 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  $\phi$ . If we call  $B_{ij}^\sigma$  a neighborhood of size  $\sigma > 0$  around  $(i, j) \in \Lambda$ ,

$$B_{ij}^\sigma = \{(i', j') \in \Lambda \mid d((i, j), (i', j')) \leq \sigma\}$$

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  $g(x_{i,j}) = x_{i+u,j+v}$  within the same neighborhood  $B^\sigma$ ,

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

One can then obtain an encoding by smoothing the embedding

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

This can be interpreted as marginalizing the translation group with a prior  $P_g$  when computing  $\phi$ . Notice that the average can be computed on a coarser lattice  $\tilde{\Lambda}$ , but its value still depends on data

in the finer grid  $\Lambda$ . Classical Sampling Theory teaches that smoothing mitigates the effect of high-frequency aliasing  $n$  at the cost of information loss on  $x$ .

Stochastic resonance also marginalizes the translation group, but by *averaging transformed data in latent space*:

$$\hat{\phi}(x) = \frac{1}{N} \sum_{i=1}^N g_{i*}^{-1} \circ \phi \circ g_i(x). \quad (1)$$

More general groups, and more general averaging kernels, can be considered although we find that the simplest case described here already suffices.

**Stochastic resonance as a defense against adversarial attacks.** To analyze the effect of an adversarial perturbation, consider a first-order expansion at each transformed input:

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

where  $J_{g_i x}$  is the Jacobian of  $\phi$  at  $g_i x$ . Substituting with equation 1 yields the perturbation term

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

By contrast, a standard feedforward encoder responds as

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

and is therefore dominated by the local worst-case Lipschitz direction at  $x$ .

The terms  $g_{i*}^{-1} J_{g_i x} [g_i n(x)]$  in the stochastic resonance expansion are generally *not aligned* across  $i$ , for several reasons. First, the adversarial noise  $n(x)$  is constructed for the quantization pattern and receptive field alignment at a *single* input  $x$ ; under translation, it is re-aliased into different patches. Second, the Jacobian  $J_{g_i x}$  changes with  $i$ , so the maximally amplifying direction at  $x$  is not simultaneously maximally amplifying at all transformed inputs. Third, after inverse warping by  $g_{i*}^{-1}$ , the signal terms  $\phi(g_i x)$  align and add coherently, whereas the Jacobian-noise terms do not.

Under mild assumptions (e.g., approximate decorrelation across  $i$ ),

$$\left\| \frac{1}{N} \sum_{i=1}^N v_i \right\|^2 \sim \frac{1}{N} \mathbb{E} \|v_i\|^2, \quad v_i = g_{i*}^{-1} J_{g_i x} [g_i n(x)],$$

so the effective Lipschitz constant is reduced by a factor  $1/\sqrt{N}$  by stochastic resonance, which suppresses adversarially amplified latent-space outliers while preserving scene-consistent features.

**Purposeful perturbation.** The only design choice in the method 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. We know that, for adversarial perturbations,  $d(\phi(\tilde{x}), \phi(x))$  is large, where  $d(\cdot)$  defines the distance between features. Ideally, for stochastic resonance, we want  $d(\hat{\phi}(\tilde{x}), \hat{\phi}(x)) = 0$  while keeping  $\hat{\phi}$  as information-preserving as possible. The theory of Stochastic Resonance shows that, if we sub-sample a signal from its native granularity  $\Lambda$  to a coarser grid  $\tilde{\Lambda} \subset \Lambda$ , and choose the purposeful perturbations to act on the finer grid  $\Lambda$ , under certain conditions one can recover the original signal at the finer granularity Benzi et al. (1981). In this paper, we focus on testing whether  $\hat{\phi}$  is insensitive to adversarial perturbations. We do so empirically in Sect. 4.

## 4 EXPERIMENTS

While  $g$  can be sampled from any invertible group transformation (e.g., rotation, scaling), we implement stochastic resonance using integer-pixel translations, denoted as  $\{g_i\} = \{(x, y) | x \in [-d_x, d_x], y \in [-d_y, d_y]\}$ , following the approach in SRT Lao et al. (2024), which avoids interpolation artifacts. While the networks’ inherent sensitivity to pixel-level shifts is typically regarded as detrimental due to the “flickering problem” Azulay & Weiss (2019); Sundaramoorthi & Wang (2019), our approach, on the contrary, leverages it to defend against adversarial perturbations. Given these perturbations  $\{g_i\}$ , ensembling can be performed at any chosen layer of the network. Features are aggregated as described in Eq. 1 and then passed to downstream network components.

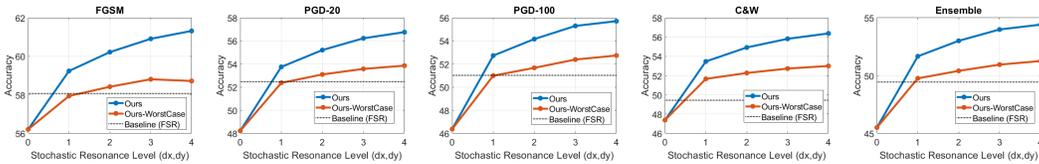


Figure 2: Results on CIFAR-10 under varying levels of stochastic resonance. Increasing the stochastic resonance level consistently enhances robustness across all settings, yielding clear gains over the baseline method (FSR). Notably, our approach achieves superior performance even under adaptive adversarial attacks (Ours-WorstCase), despite the baseline being evaluated only in the non-adaptive case.

Method	Architecture	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble	AutoAttack
AT	ResNet-18	85.02	56.21	48.22	46.37	47.38	45.90	44.11
+ FD		85.14	56.81	48.54	46.70	47.72	45.82	44.57
+ CAS		<b>85.78</b>	55.57	50.42	49.91	53.47	46.46	44.23
+ CIFS		79.87	56.53	49.80	48.17	49.89	47.26	43.94
+ FSR		81.46	58.07	52.47	51.02	49.44	48.34	46.41
+ TTE		85.25	59.20	53.00	51.65	52.45	50.60	-
+ Ours		84.93	<b>61.02</b>	<b>56.08</b>	<b>55.17</b>	<b>55.53</b>	<b>53.68</b>	<b>49.78</b>
+ Ours-WorstCase		84.93	58.81	53.58	52.39	52.73	50.95	-
TRADES	WideResNet-34	84.92	60.87	56.13	55.16	54.02	53.38	46.81
+Ours		<b>85.03</b>	<b>62.43</b>	<b>58.64</b>	<b>57.87</b>	<b>57.18</b>	<b>56.28</b>	<b>51.80</b>
MART	ResNet-18	<b>83.07</b>	60.21	54.14	52.90	49.62	48.95	44.27
+ Ours		82.70	<b>62.62</b>	<b>59.03</b>	<b>58.13</b>	<b>55.51</b>	<b>54.61</b>	<b>50.72</b>

Table 1: Defense against adversarial attacks on classification task (CIFAR-10). Compared to baselines that filter or manipulate features, ours does not modify network architecture or weights. Instead, ours performs an ensemble in the feature space. On CIFAR-10, ours achieves state-of-the-art robustness without requiring any additional training. Moreover, even in a worst-case adaptive adversary setting where the attacker is fully aware of the defense and how stochastic resonance is applied, the effectiveness of adversarial attacks is still notably reduced, while the computational cost for executing such attacks is significantly increased.

#### 4.1 IMAGE CLASSIFICATION

**CIFAR-10.** We evaluate on CIFAR-10 Krizhevsky et al. (2009), building upon the standard and publicly available code base of FSR Kim et al. (2023) and accompanying evaluation protocol (e.g. AutoAttack Croce & Hein (2020)). We apply stochastic resonance to networks pre-trained with AT Madry et al. (2017), TRADES Zhang et al. (2019), and MART Wang et al. (2019), using their publicly released weights. Our method operates with these methods purely at test time. We conduct experiments on ResNet-18 He et al. (2016) and WideResNet-34 Zagoruyko & Komodakis (2016), 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.

Fig. 2 shows the results varying different levels of stochastic resonance. Increasing the resonance level consistently enhances robustness in all attack settings, leading to substantial improvements over the baseline method (FSR). Importantly, our approach surpasses the baseline even under adaptive adversarial attacks. Additionally, we compare our method against multiple baselines, including feature-level manipulation methods (FD Xie et al. (2019), CAS Bai et al. (2021), CIFS Yan et al. (2021)) and ensemble-based approach TTE Pérez et al. (2021). The results, summarized in Tab. 1, demonstrate that stochastic resonance consistently outperforms all baselines across different attacks. Even in the worst-case adaptive attack scenario, where the adversary accounts for all stochastic resonance forward passes, the model remains significantly more robust than the baseline methods. In addition, stochastic resonance increases attack complexity under the adaptive settings, making adversarial noise generation more challenging for the attacker. As a result, the computational cost for generating adaptive adversarial perturbations increases substantially. For example, with stochastic resonance, 8x more wall-clock time is required to create adversarial examples with PGD-100.

**ImageNet.** We further evaluate our approach on the ImageNet Deng et al. (2009) classification dataset using standard segmentation backbones, including ResNet-50 He et al. (2016) and Vision

Att. Strength ( $\epsilon$ )	8/255		4/255		2/255		1/255	
	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 ( $d = 3$ )	<b>25.77</b>	<b>65.88</b>	<b>42.52</b>	<b>76.75</b>	<b>54.94</b>	<b>82.64</b>	<b>62.08</b>	<b>85.50</b>

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.

Transformer Dosovitskiy et al. (2020), without adversarial training. As in the CIFAR experiments, we vary the level of stochastic resonance and conduct ablation studies by testing against PGD attacks of different strengths. Tab. 2 reports the results for ViT-Small. Consistent with the CIFAR-10 findings, increasing the resonance level leads to consistent improvements in robustness, as measured by both Top-1 and Top-5 accuracy. Notably, the vanilla ViT-Small model without attack achieves 72.9 (Top-1) and 92.91 (Top-5), which means 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  $\epsilon = 2/255$ ). We further evaluate our method on ResNet-50 and observe a consistent trend, as shown in Tab. 4.

We also explored group transformations other than translation, e.g. rotations in Tab. 4. For a fair comparison, we use the same number of augmentations as in the translation experiments. Rotations behave similarly to translations at low levels of stochastic resonance, but performance degrades as the resonance level increases. We hypothesize that since convolutional filters are inherently translation-invariant but not rotation-invariant, aligning features under different rotations reduces feature quality. Moreover, rotations are approximately 30% slower due to interpolation overhead.

Since adversarial perturbations often manifest as high-frequency noise, having a low-pass filter in early layers may form an effective defense. As our method applies to arbitrarily chosen layers, we vary the termination layer of stochastic resonance. As shown in Tab. 3, applying it only through the first residual block already achieves strong adversarial robustness, while extending it to the second block yields the strongest result. This finding is significant: running stochastic resonance through shallow layers can be sufficient as a defense strategy, which reduces overall computational cost.

**CIFAR-10-C.** Although focused on adversarial perturbations, we are also interested in understanding how stochastic resonance behaves under real-world corruptions. To this end, we evaluate our method on the CIFAR-10-C benchmark Hendrycks & Dietterich (2019), which contains a diverse set of test-time corruptions covering noise, blur, weather effects, and photometric changes. We test on a ResNet-50 model trained on the original CIFAR-10 dataset without corruption. Following the best-performing configuration identified in prior sections, with stochastic resonance applied at translation radius  $d = 3$ , and ensemble the features after the Residual Block 2. The dataset provides corruption severity levels from 1 to 5; we report aggregated results across all levels.

Overall, stochastic resonance recovers an average of 20.8% of the performance lost to these perturbations. In particular, our approach achieves substantial gains under several challenging corruptions, recovering 32.6% under `defocus_blur`, 34.4% under `fog`, 31.6% under `gaussian_blur`, and 31.9% under `saturate`. Other corruption types also see improvements ranging from 3.4% to 20.8%. These findings further highlight a key strength of stochastic resonance: it improves robustness across a *wide spectrum* of corruptions that are not specified during training. This property strengthens stochastic resonance as a *general-purpose, plug-and-play* test-time defense. Rather than being tailored to a specific adversarial formulation, it has stronger potential for real-world deployment where the exact perturbation type is often unknown and unpredictable.

Att. Strength ( $\epsilon$ )	8/255		4/255		2/255		1/255	
	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	<b>24.88</b>	<b>48.54</b>	<b>25.90</b>	<b>49.66</b>	<b>27.23</b>	<b>51.19</b>	<b>30.11</b>	<b>54.15</b>
Res. Block 3	21.76	44.28	22.73	45.47	24.14	47.14	27.18	50.20

Table 3: **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, substantially reducing the overall computational cost.

Att. Strength ( $\epsilon$ )	8/255		4/255		2/255		1/255	
	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	39.5	20.36	40.48	22.01	42.50
Ours ( $d = 3$ )	<b>21.17</b>	<b>43.58</b>	<b>22.04</b>	<b>44.78</b>	<b>23.44</b>	<b>46.53</b>	<b>26.58</b>	<b>49.73</b>
w/ Rotation	15.15	32.89	15.61	33.46	16.17	34.19	17.33	35.59

Table 4: **Stochastic resonance using translation v.s. rotation on ResNet-50.** While rotations provide similar gains at low resonance levels, performance degrades as the resonance level increases, likely due to the lack of rotational invariance in convolutional filters.

	Attack Strength ( $\epsilon$ )	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	13.42	22.09	96.61	8.12	13.69	89.25	4.89	7.05	73.01	2.89	3.32	36.25
	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
	Error Reduced (%)	31.76	34.40	1.57	23.91	31.73	4.63	13.07	22.99	7.58	7.64	8.88	16.08
	Ours ( $d = 2$ )	<b>9.22</b>	<b>13.88</b>	<b>94.61</b>	<b>6.13</b>	<b>8.87</b>	<b>84.49</b>	<b>4.19</b>	<b>5.43</b>	<b>66.74</b>	<b>2.73</b>	<b>3.12</b>	<b>31.18</b>
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	161.79	162.69	99.99	131.46	140.41	98.57	63.86	88.21	85.39	7.28	17.44	39.81
	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
	Error Reduced (%)	33.30	23.04	1.70	46.86	39.72	6.94	67.70	52.24	14.18	38.95	51.15	25.87
	Ours ( $d = 2$ )	<b>77.59</b>	<b>100.14</b>	<b>96.14</b>	<b>44.77</b>	<b>66.23</b>	<b>89.24</b>	<b>17.98</b>	<b>32.80</b>	<b>71.09</b>	<b>3.76</b>	<b>6.73</b>	<b>28.44</b>
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	

Table 5: **Stochastic Resonance Enhances Stereo Matching Robustness.** Incorporating stochastic resonance significantly reduces prediction errors induced by adversarial attacks across all evaluation metrics, reducing error by up to 71.89% (MAE, when attacked by PGD with  $\epsilon = 0.005$ ). Notably, this defense mechanism operates entirely at test time without requiring any model re-training, which sets it apart from existing methods.



Figure 3: **Stereo matching robustness via stochastic resonance.** We present visual results on stereo matching under various adversarial attack scenarios, including PGD and FGSM at different perturbation levels. These attacks significantly degrade the network’s predictions, leading to substantial errors. By incorporating stochastic resonance, we demonstrate a significant reduction in prediction errors. This technique holds significant potential for improving robustness in safety-critical real-world applications, such as autonomous driving, where stereo vision must remain reliable under diverse environmental conditions and adversarial threats.

## 4.2 STEREO MATCHING

Stereopagnosia Wong et al. (2021) first introduced adversarial attacks to stereo matching, yet no test-time method has demonstrated an effective defense. The primary challenge arises from the infeasibility of data augmentations, as they risk altering the physics of the input, leading to incorrect estimation. In contrast, our feature-level ensemble is suitable for this task, as transformations introduced by stochastic resonance are “undone” in the latent space, ensuring that features remain aligned precisely with the original input. This process mirrors AugUndo Wu et al. (2024) conceptually.

We evaluate our method on the standard benchmark used in *Stereopagnosia*, derived from KITTI Geiger et al. (2012). Experiments are conducted using FGSM and PGD attacks against a pre-trained PSMNet Chang & Chen (2018). Since no existing test-time defense is available, we adopt latent-space smoothing as a baseline. As shown in Tab. 5, both attacks corrupt network predictions, with PGD proving substantially 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. Crucially, these gains are achieved entirely at test time, without additional training or prior knowledge of the attack.

Fig. 3 provides qualitative results, featuring different attack strengths and methods. The adversarial perturbations introduce significant distortions, as indicated by bright regions in the visualized predictions. Stochastic resonance effectively mitigates these distortions, drastically reducing prediction errors. 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.

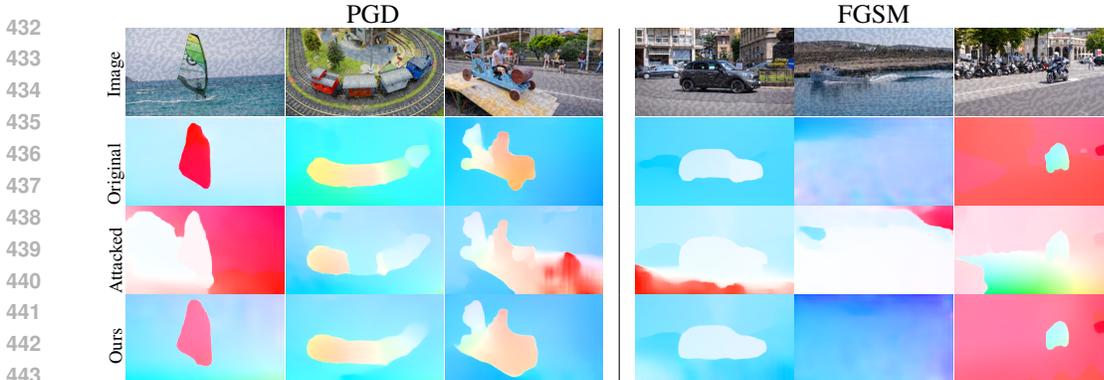


Figure 5: **Optical flow robustness via stochastic resonance.** Qualitative results (visualized with a color wheel) show that our method substantially mitigates the degradation caused by both PGD and FGSM attacks. This robustness is particularly relevant for visual perception systems that rely on accurate motion estimation.

While some defenses Zhang et al. (2023); Cheng et al. (2021; 2022); Berger et al. (2022) have been proposed to train a more robust network under adverse conditions, test-time defenses remain largely unexplored. Our method is the first to provide a viable solution in this setting.

### 4.3 OPTICAL FLOW

We further evaluate our method on optical flow, which computes the dense motion field between two images. The accuracy of optical flow is measured using the End-Point Error (EPE). Although several adversarial attacks on optical flow have been proposed Schrodi et al. (2022); Ranjan et al. (2019); Schmalfluss et al. (2022); Agnihotri et al. (2023); Scheurer et al. (2024), there is currently no standardized benchmark for the evaluation of defense. Therefore, we adopt a simple and controlled setup: we employ the RAFT Teed & Deng (2020) model and focus on global adversarial perturbations generated by PGD and FGSM. We test on the DAVIS Pont-Tuset et al. (2017) dataset.

To defend against adversarial attacks, we apply stochastic resonance to the convolutional feature extractor of RAFT. Since the perturbation is applied only at the feature extraction stage, no additional overhead is introduced in the computationally intensive matching module. Quantitative results (Fig. 4) show that increasing Stochastic Resonance reduces EPE, which aligns with our findings in classification. As in Fig. 5, our approach effectively removes errors caused by adversarial noise.

We further compare our method to an alternative ensembling approach that aggregates predictions in the output space, conceptually similar to TTE Pérez et al. (2021). In this variant, we apply the same stochastic transformations but instead ensemble at the output level rather than in the feature space. While this method provides marginal improvements, it remains less effective than our approach. This finding highlights the advantage of having the freedom to choose from any stage of the model to perform ensemble. In this particular experiment, we demonstrate that ensembling solely at the image encoding sub-module, while leaving the rest of the RAFT network unchanged, yields substantial improvements in robustness, thanks to the flexibility of our method.

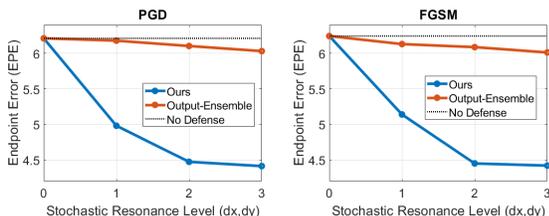


Figure 4: **Enhanced optical flow robustness with stochastic resonance.** Under PGD and FGSM, stochastic resonance significantly reduces endpoint error in optical flow estimation. Notably, our method performs ensembling in the latent feature space rather than the output space, providing greater flexibility. While ensembling in the output space offers minor performance gains, our approach consistently achieves superior robustness across all levels of stochastic resonance.

**Adverse weather corruptions.** For real-world deployment, we further evaluate stochastic resonance under adverse weather conditions Schmalfluss et al. (2023). 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 are highly localized (e.g., rain streaks, snow particles), violating the assumptions underlying our theoretical analysis that rely on small, spatially uncorrelated

486 perturbations. Despite these challenges, stochastic resonance still yields measurable improvements:  
 487 recovering 7.5% of the accuracy loss for `snow` and 3.7% for `rain`. While, as expected, the gains are  
 488 smaller than those observed under adversarial attacks, they nonetheless demonstrate that stochastic  
 489 resonance provides meaningful robustness against real-world perturbations, which is consistent with  
 490 our findings on CIFAR-10-C. This further confirms stochastic resonance’s effectiveness as a generic  
 491 test-time robustness mechanism that extends beyond adversarial settings.

## 493 5 DISCUSSION AND CONCLUSION

495 **Speed.** Stochastic resonance incurs low computational overhead when executed in parallel: raising  
 496  $d$  to 3 with ResNet-50 adds only 0.06 seconds to the inference time on an NVIDIA 1080Ti GPU. No-  
 497 tably, most of the overhead arises from transformations (and their inverse) implemented by Python  
 498 loops with `torchvision.transforms`. We expect this overhead to be drastically reduced with  
 499 more efficient CUDA implementations. Excluding this `for` loop, stochastic-resonance-specific for-  
 500 ward computation is lightweight: 0.028s with  $d=3$  (49 forward passes), while a single forward  
 501 pass ( $d=1$ ) takes 0.007s. Moreover, strong robustness can be attained by applying stochastic reso-  
 502 nance only to shallow layers, offering substantially greater efficiency than existing ensemble-based  
 503 defenses (e.g. Pérez et al. (2021)) that require multiple passes through the entire network. More-  
 504 over, stochastic resonance is fully plug-and-play. In contrast, adversarial training is attack-specific,  
 505 over  $6\times$  slower than standard training, and must be repeated for each type of threat. As such, the  
 506 computation of our method is well justified by its robustness gains and training-free nature.

507 **Memory.** Increasing  $d$  from 0 to 3 raises the number of forward passes from 1 to 49 but increases  
 508 peak memory only moderately (from 158MB to 650MB) for ImageNet-sized inputs on ResNet-  
 509 50. Modern GPUs easily accommodate this overhead. An interesting observation emerges when  
 510 performing *adaptive* (worst-case) attacks: memory usage jumps from 346MB to 10.6GB for a single  
 511  $224\times 224$  image. This increase makes adaptive attacks practically infeasible in standard hardware  
 512 settings, providing an inherent defensive advantage. Overall, stochastic resonance offers favorable  
 513 cost-robustness trade-offs with minimal memory growth and low computational footprint.

514 **On-demand scaling.** One of the key strengths of our approach is its flexibility: providing a trade-  
 515 off between robustness and computational cost. We offer a tunable “knob” that allows practitioners  
 516 to adjust the level of resilience based on available resources on the fly: when the system has more  
 517 computational capacity, add a higher level of stochastic resonance, vice versa. Note that, such a  
 518 design does not *rely* on additional computation, yet more computation can bring *extra* performance.  
 519 Moreover, our experiments show that the method generalizes across a wide range of tasks and archi-  
 520 tectures that include an encoder. This on-demand scaling mirrors inference-time scaling in language  
 521 models, where performance can be improved without modifying the underlying pre-trained model.

522 **Limitations.** Despite its strengths, our method has some limitations. First, while we offer parallel  
 523 computation as a remedy, the computational overhead introduced by stochastic resonance may not  
 524 be negligible for scenarios with memory and power constraints. Also, our current study focuses  
 525 on integer-pixel translations. While this choice avoids interpolation artifacts and preserves spatial  
 526 consistency, more generic transformations, including learned transformations, could be explored.

527 **Conclusion.** In this work, we present a signal-processing perspective for defending against adver-  
 528 sarial attacks, motivated by the connection between adversarial perturbations and aliasing artifacts.  
 529 Accordingly, we propose a “combat noise with noise” approach by introducing stochastic resonance  
 530 as a defense mechanism. We formalize the problem and implement stochastic resonance using pixel-  
 531 level translations paired with their inverse transformation in the feature space. The resulting method  
 532 is training-free, agnostic to both tasks and attack types, and independent of network architectures.

533 We evaluate our method across various tasks. Empirical results on image classification demonstrate  
 534 that our stochastic resonance approach achieves state-of-the-art robustness against diverse attack  
 535 types, offering a clear advantage over feature-level denoising and filtering. Even in the adaptive ad-  
 536 versary scenario, where an attacker is aware of the use of stochastic resonance, our method maintains  
 537 strong robustness. Furthermore, we are the first to introduce test-time defense to dense prediction  
 538 tasks. Specifically, we apply this method to stereo matching and optical flow, achieving up to a 71%  
 539 reduction in prediction error. More importantly, these findings highlight the practical potential of  
 stochastic resonance as a universal defense in real-world adversarial scenarios.

## REPRODUCIBILITY STATEMENT

We provide sufficient technical details in the paper to ensure reproducibility. Specifically, we describe the augmentations used for stochastic resonance, including augmentations (e.g. translation, rotation) and their corresponding inverse transformations, as well as the model architectures, datasets, and the network layers where our method is applied. Attack settings and evaluation protocols are drawn directly from standard benchmark datasets and publicly available code base, ensuring comparability with prior work. All implementation details necessary to reproduce our experiments, including parameters and ablation settings, are provided in the main paper and further expanded in the Appendix. Our experiments can be reproduced on a single desktop-level GPU without requiring large-scale computational resources. We will release the complete source code and pre-computed adversarial data upon publication.

## LLM STATEMENT

All technical content of this work, including literature review, methodology, experiments, and analyses, was developed entirely by the authors. Large Language Models (LLMs) were employed as a tool for proofreading, without contributing to the scientific or technical substance of the manuscript.

## REFERENCES

- 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.
- Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.
- 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.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.
- 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*, pp. 7818–7827, 2021.
- Roberto Benzi, Alfonso Sutera, and Angelo Vulpiani. The mechanism of stochastic resonance. *Journal of Physics A: mathematical and general*, 14(11):L453, 1981.
- Roberto Benzi, Giorgio Parisi, Alfonso Sutera, and Angelo Vulpiani. Stochastic resonance in climatic change. *Tellus*, 34(1):10–16, 1982.
- 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*, pp. 15180–15190, 2022.
- Qi-Zhi Cai, Min Du, Chang Liu, and Dawn Song. Curriculum adversarial training. *arXiv preprint arXiv:1805.04807*, 2018.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. Ieee, 2017.
- Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5418, 2018.
- 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.

- 594 Kelvin Cheng, Christopher Healey, and Tianfu Wu. Towards adversarially robust and domain gener-  
595 alizable stereo matching by rethinking dnn feature backbones. *arXiv preprint arXiv:2108.00335*,  
596 2021.
- 597 Kelvin Cheng, Tianfu Wu, and Christopher Healey. Revisiting non-parametric matching cost vol-  
598 umes for robust and generalizable stereo matching. *Advances in Neural Information Processing*  
599 *Systems*, 35:16305–16318, 2022.
- 600  
601 Gilad Cohen and Raja Giryes. Simple post-training robustness using test time augmentations and  
602 random forest. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer*  
603 *Vision*, pp. 3996–4006, 2024.
- 604  
605 Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble  
606 of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–  
607 2216. PMLR, 2020.
- 608  
609 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hi-  
610 erarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*,  
611 pp. 248–255. Ieee, 2009.
- 612  
613 Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi,  
614 Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial de-  
615 fense. *arXiv preprint arXiv:1803.01442*, 2018.
- 616  
617 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas  
618 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An  
619 image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*  
620 *arXiv:2010.11929*, 2020.
- 621  
622 Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W Taylor. Batch  
623 normalization is a cause of adversarial vulnerability. *arXiv preprint arXiv:1905.02161*, 2019.
- 624  
625 Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti  
626 vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*,  
627 pp. 3354–3361. IEEE, 2012.
- 628  
629 Sajjad Ghiasvand, Amirhossein Reisizadeh, Mahnoosh Alizadeh, and Ramtin Pedarsani. Robust  
630 decentralized learning with local updates and gradient tracking. *arXiv preprint arXiv:2405.00965*,  
631 2024.
- 632  
633 James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press,  
634 2014.
- 635  
636 Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial  
637 examples. *arXiv preprint arXiv:1412.6572*, 2014.
- 638  
639 Bhavna Gopal, Huanrui Yang, Jingyang Zhang, Mark Horton, and Yiran Chen. Boosting adversarial  
640 robustness with clat: Criticality leveraged adversarial training. In *Forty-second International*  
641 *Conference on Machine Learning*, 2025.
- 642  
643 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-  
644 nition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.  
645 770–778, 2016.
- 646  
647 Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common cor-  
648 ruptions and perturbations. In *International Conference on Learning Representations*, 2019.
- 649  
650 Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. Enhancing adversarial  
651 training with second-order statistics of weights. In *Proceedings of the IEEE/CVF conference on*  
652 *computer vision and pattern recognition*, pp. 15273–15283, 2022.
- 653  
654 Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint*  
655 *arXiv:1803.06373*, 2018.

- 648 Woo Jae Kim, Yoonki Cho, Junsik Jung, and Sung-Eui Yoon. Feature separation and recalibration  
649 for adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and*  
650 *pattern recognition*, pp. 8183–8192, 2023.
- 651 Bart Kosko and Sanya Mitaim. Robust stochastic resonance: Signal detection and adaptation in  
652 impulsive noise. *Physical review E*, 64(5):051110, 2001.
- 653 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.  
654 2009.
- 655 Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world.  
656 In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.
- 657 Dong Lao, Yangchao Wu, Tian Yu Liu, Alex Wong, and Stefano Soatto. Sub-token vit embedding  
658 via stochastic resonance transformers. In *International Conference on Machine Learning*, pp.  
659 25995–26006. PMLR, 2024.
- 660 Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial  
661 example attacks against neural networks. In *Proceedings of the AAAI conference on artificial*  
662 *intelligence*, volume 32, 2018.
- 663 Divyam Madaan, Jinwoo Shin, and Sung Ju Hwang. Adversarial neural pruning with latent vul-  
664 nerability suppression. In *International conference on machine learning*, pp. 6575–6585. PMLR,  
665 2020.
- 666 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.  
667 Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*,  
668 2017.
- 669 Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal  
670 adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern*  
671 *recognition*, pp. 1765–1773, 2017.
- 672 Haniyeh Ehsani Oskouie, Mohammad-Shahram Moin, and Shohreh Kasaei. Attack on scene flow  
673 using point clouds. In *2024 IEEE 34th International Workshop on Machine Learning for Signal*  
674 *Processing (MLSP)*, pp. 1–6. IEEE, 2024.
- 675 Juan C Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem,  
676 and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling.  
677 In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 81–91, 2021.
- 678 Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and  
679 Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint*  
680 *arXiv:1704.00675*, 2017.
- 681 Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J. Black. Attacking optical flow. In  
682 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October  
683 2019.
- 684 Erik Scheurer, Jenny Schmalfuss, Alexander Lis, and Andrés Bruhn. Detection defenses: An empty  
685 promise against adversarial patch attacks on optical flow. In *Proceedings of the IEEE/CVF Winter*  
686 *Conference on Applications of Computer Vision*, pp. 6489–6498, 2024.
- 687 Jenny Schmalfuss, Philipp Scholze, and Andrés Bruhn. A perturbation-constrained adversarial at-  
688 tack for evaluating the robustness of optical flow. In *European Conference on Computer Vision*,  
689 pp. 183–200. Springer, 2022.
- 690 Jenny Schmalfuss, Lukas Mehl, and Andrés Bruhn. Distracting downpour: Adversarial weather  
691 attacks for motion estimation. In *Proceedings of the IEEE/CVF international conference on com-*  
692 *puter vision*, pp. 10106–10116, 2023.
- 693 Simon Schrodi, Tonmoy Saikia, and Thomas Brox. Towards understanding adversarial robustness  
694 of optical flow networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*  
695 *Pattern Recognition*, pp. 8916–8924, 2022.

- 702 Ji Shu-Yao, Yuan Fei, Chen Ke-Yu, and Cheng En. Application of stochastic resonance technology  
703 in underwater acoustic weak signal detection. In *OCEANS 2016-Shanghai*, pp. 1–5. IEEE, 2016.  
704
- 705 Nigel G Stocks, D Allingham, and Robert P Morse. The application of suprathreshold stochastic  
706 resonance to cochlear implant coding. *Fluctuation and noise letters*, 2(03):L169–L181, 2002.  
707
- 708 Ganesh Sundaramoorthi and Timothy E Wang. Translation insensitive cnns. *arXiv preprint*  
709 *arXiv:1911.11238*, 2019.
- 710 Linyu Tang and Lei Zhang. Robust overfitting does matter: Test-time adversarial purification with  
711 fgsm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,  
712 pp. 24347–24356, 2024.  
713
- 714 Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer*  
715 *Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings,*  
716 *Part II 16*, pp. 402–419. Springer, 2020.
- 717 Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space  
718 of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.  
719
- 720 Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks  
721 to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–  
722 1645, 2020.
- 723 Yun-Yun Tsai, Ju-Chin Chao, Albert Wen, Zhaoyuan Yang, Chengzhi Mao, Tapan Shah, and Junfeng  
724 Yang. Test-time detection and repair of adversarial samples via masked autoencoder. *arXiv*  
725 *preprint arXiv:2303.12848*, 2023.  
726
- 727 Jumpei Ukita and Kenichi Ohki. Adversarial attacks and defenses using feature-space stochasticity.  
728 *Neural Networks*, 167:875–889, 2023.
- 729 Dequan Wang, An Ju, Evan Shelhamer, David Wagner, and Trevor Darrell. Fighting gradients with  
730 gradients: Dynamic defenses against adversarial attacks. *arXiv preprint arXiv:2105.08714*, 2021.  
731
- 732 Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, and Zhangyang Wang. Removing  
733 batch normalization boosts adversarial training. In *International Conference on Machine Learn-*  
734 *ing*, pp. 23433–23445. PMLR, 2022.
- 735 Hongjun Wang and Yisen Wang. Self-ensemble adversarial training for improved robustness. *arXiv*  
736 *preprint arXiv:2203.09678*, 2022.  
737
- 738 Jun Wang, Qingbo He, and Fanrang Kong. Adaptive multiscale noise tuning stochastic resonance  
739 for health diagnosis of rolling element bearings. *IEEE Transactions on instrumentation and mea-*  
740 *surement*, 64(2):564–577, 2014.  
741
- 742 Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving  
743 adversarial robustness requires revisiting misclassified examples. In *International conference on*  
744 *learning representations*, 2019.
- 745 Thomas Wellens, Vyacheslav Shatokhin, and Andreas Buchleitner. Stochastic resonance. *Reports*  
746 *on progress in physics*, 67(1):45, 2003.  
747
- 748 Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with  
749 adversarial perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol-  
750 *ume 35*, pp. 2879–2888, 2021.
- 751 Yangchao Wu, Tian Yu Liu, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Au-  
752 gundo: Scaling up augmentations for monocular depth completion and estimation. In *European*  
753 *Conference on Computer Vision*, pp. 274–293. Springer, 2024.  
754
- 755 Chang Xiao, Peilin Zhong, and Changxi Zheng. Enhancing adversarial defense by k-winners-take-  
all. *arXiv preprint arXiv:1905.10510*, 2019.

- 756 Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising  
757 for improving adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer  
758 vision and pattern recognition*, pp. 501–509, 2019.
- 759 Xilie Xu, Jingfeng Zhang, and Mohan Kankanhalli. Autolora: An automated robust fine-tuning  
760 framework. In *The Twelfth International Conference on Learning Representations*, 2024.
- 762 Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Tan, and Masashi Sugiyama. Cifs:  
763 Improving adversarial robustness of cnns via channel-wise importance-based feature selection. In  
764 *International Conference on Machine Learning*, pp. 11693–11703. PMLR, 2021.
- 765 Cheng-Han Yeh, Kuanchun Yu, and Chun-Shien Lu. Test-time adversarial defense with opposite  
766 adversarial path and high attack time cost. *arXiv preprint arXiv:2410.16805*, 2024.
- 768 Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint  
769 arXiv:1605.07146*, 2016.
- 770 Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan.  
771 Theoretically principled trade-off between robustness and accuracy. In *International conference  
772 on machine learning*, pp. 7472–7482. PMLR, 2019.
- 774 Howard Zhang, Yunhao Ba, Ethan Yang, Varan Mehra, Blake Gella, Akira Suzuki, Arnold Pfahnl,  
775 Chethan Chinder Chandrappa, Alex Wong, and Achuta Kadambi. Weatherstream: Light trans-  
776 port automation of single image deweathering. In *Proceedings of the IEEE/CVF Conference on  
777 Computer Vision and Pattern Recognition*, pp. 13499–13509, 2023.
- 778 Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankan-  
779 halli. Attacks which do not kill training make adversarial learning stronger. In *International  
780 conference on machine learning*, pp. 11278–11287. PMLR, 2020.
- 782 Lingyu Zhang, Chengzhi Mao, Junfeng Yang, and Carl Vondrick. Adversarially robust video per-  
783 ception by seeing motion. *arXiv preprint arXiv:2212.07815*, 2022.
- 784 Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli.  
785 Towards robust image classification using sequential attention models. In *Proceedings of the  
786 IEEE/CVF conference on computer vision and pattern recognition*, pp. 9483–9492, 2020.
- 787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

## 810 A IMPLEMENTATION FOR CLASSIFICATION

811  
812 For our classification experiments, we built our implementation on top of standard network archi-  
813 tectures, implementing SR on two main architectures derived from the FSR codebase Kim et al.  
814 (2023):

- 815 • ResNet-18 He et al. (2016): A standard residual network with 18 layers organized in four  
816 main blocks with increasing channel dimensions (64, 128, 256, 512).
- 817 • WideResNet-34 Zagoruyko & Komodakis (2016): A wider variant of ResNet with depth  
818 34 and width factor 10, resulting in higher representational capacity with three main blocks  
819 with channel dimensions scaled by the width factor (160, 320, 640).

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

821  
822 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  
823 follows. First, we create a set of  $(2d_x + 1) \times (2d_y + 1)$  perturbed versions of the input by applying  
824 pixel-level translations within the range  $[-d_x, d_x] \times [-d_y, d_y]$  pixels:

$$825 X_{\text{perturbed}} = \{g_{i,j}(X) \mid i \in [-d_y, d_y], j \in [-d_x, d_x]\} \quad (2)$$

826 where  $g_{i,j}$  translates the image by  $(i, j)$  pixels. These transformations are applied using PyTorch’s  
827 “transforms.functional.affine” function with translation parameters while preserving the original im-  
828 age properties.

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

$$831 F = \phi(X_{\text{perturbed}}) \quad (3)$$

832 where  $\phi$  represents the network up to the bottleneck layer. This batch processing approach signifi-  
833 cantly improves computational efficiency compared to individual forward passes.

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

$$836 F_{\text{ensembled}} = \frac{1}{n} \sum_{i,j} T_{-i,-j}(F_{i,j}) \quad (4)$$

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

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

841 For evaluation, we tested our approach against standard adversarial attacks (FGSM Goodfellow  
842 et al. (2014), PGD-20 and PGD-100 Madry et al. (2017), and C&W Carlini & Wagner (2017)),  
843 all bounded within  $\epsilon = 8/255$  under  $\ell_\infty$ -norm. We also report an Ensemble metric that measures  
844 the worst-case performance across all attacks for each test example, providing a comprehensive  
845 robustness assessment.

## 852 B IMPLEMENTATION FOR STEREO MATCHING

853  
854 For our stereo matching experiments, we built our implementation on top of standard stereo network  
855 architectures to ensure our approach remains model-agnostic and requires no training or fine-tuning.  
856 We integrated SR with PSMNet Chang & Chen (2018), a pyramid stereo matching network with a  
857 stacked hourglass architecture that uses 3D convolutions on a cost volume constructed by concate-  
858 nating features. We apply SR at the feature extraction stage, before cost volume construction, where  
859 stereo correspondences are first established.

860 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  
861 operates as follows. First, we create a set of  $(2d_x + 1) \times (2d_y + 1)$  perturbed versions of each input  
862 image by applying translations within the range  $[-d_x, d_x] \times [-d_y, d_y]$  pixels:

$$863 X_{L,\text{perturbed}} = \{g_{i,j}(x_L) \mid i \in [-d_y, d_y], j \in [-d_x, d_x]\} \quad (5)$$

$$X_{R,\text{perturbed}} = \{g_{i,j}(x_R) \mid i \in [-d_y, d_y], j \in [-d_x, d_x]\}. \quad (6)$$

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

$$F_L = \phi(X_{L,\text{perturbed}}) \quad (7)$$

$$F_R = \phi(X_{R,\text{perturbed}}) \quad (8)$$

where  $\phi$  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 Goodfellow et al. (2014) and I-FGSM Kurakin et al. (2018) (a special case of PGD), bounded within various  $\epsilon$  values ( $\{0.002, 0.005, 0.01, 0.02\}$ ) under  $\ell_\infty$ -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) Luo et al. (2018).

## C DETAILS ABOUT ATTACK ON OPTICAL FLOW

To find an adversarial attack for optical flow estimated by a given neural network  $f$ , we utilize a similar approach to Oskouie et al. (2024) that aims to find a perturbation  $\delta$  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  $\delta$  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  $\epsilon$ -norm bounded adversary  $\delta$  for optical-flow is calculated by optimizing the following

$$\max_{\delta: \|\delta\| \leq \epsilon} \text{EPE}(OF, f(F_1 + \delta, F_2)). \quad (9)$$

One  $\ell_\infty$ -bounded adversary  $A$  for the aforementioned optimization problem is Fast Gradient Sign Method (FGSM) Goodfellow et al. (2014) which can be obtained by

$$\begin{aligned} L &= \text{EPE}(OF, f(F_1, F_2)), \\ A &= F_1 + \epsilon \cdot \text{sign}(\nabla_{F_1} L). \end{aligned} \quad (10)$$

Projected gradient descent (PGD) Madry et al. (2017) 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

$$F_1^{(t+1)} = \Pi_{F_1+S}(F_1^{(t)} + \alpha \cdot \text{sign}(\nabla_{F_1} L)). \quad (11)$$

Note that in PGD, since the perturbations are considered to be too minimal to significantly alter the flow dynamics, the ground-truth optical flow is not updated by intermediate perturbations applied to the input data.

For our experimental setup, we chose to set the norm value  $\epsilon$  at  $\frac{10}{256}$ . Furthermore, we configured the PGD algorithm to run for 10 iterations. The step size  $\alpha$  was determined by dividing  $2.5 \cdot \epsilon$  by the total number of iterations, ensuring a balanced progression throughout the optimization process.

## D 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 perturbation on various images.

918

919

920

921

922

923

924

925

926

927

928

929

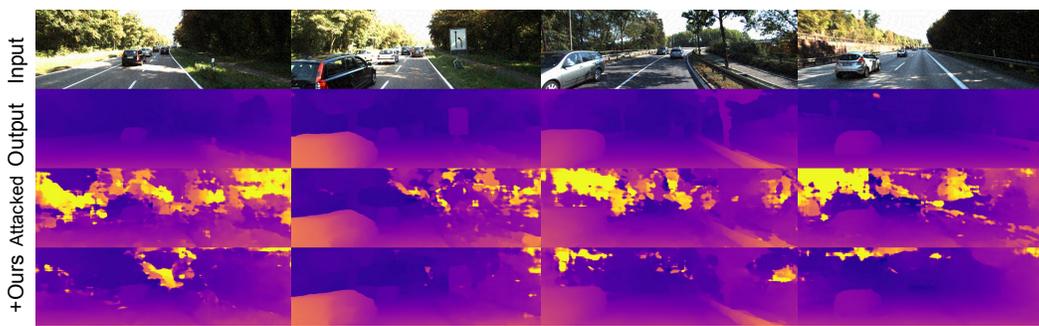


Figure 6: Visual results on stereo matching against FGSM attack, without and with SR,  $\epsilon = 0.02$

930

931

932

933

934

935

936

937

938

939

940

941

942

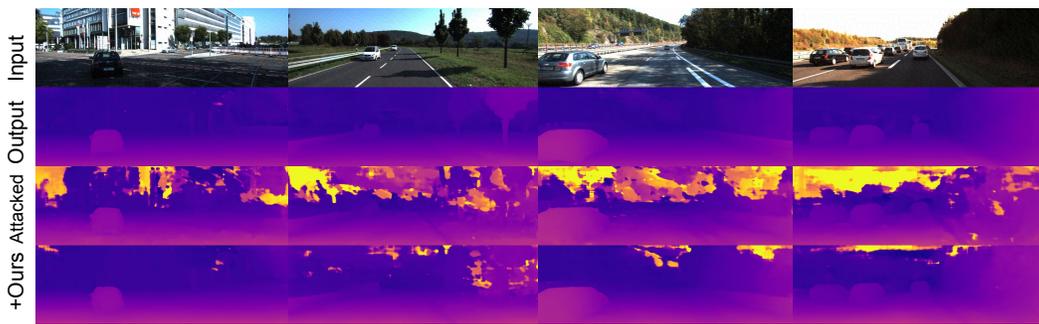


Figure 7: Visual results on stereo matching against FGSM attack, without and with SR,  $\epsilon = 0.01$

943

944

945

946

947

948

949

950

951

952

953

954

955

956



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

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

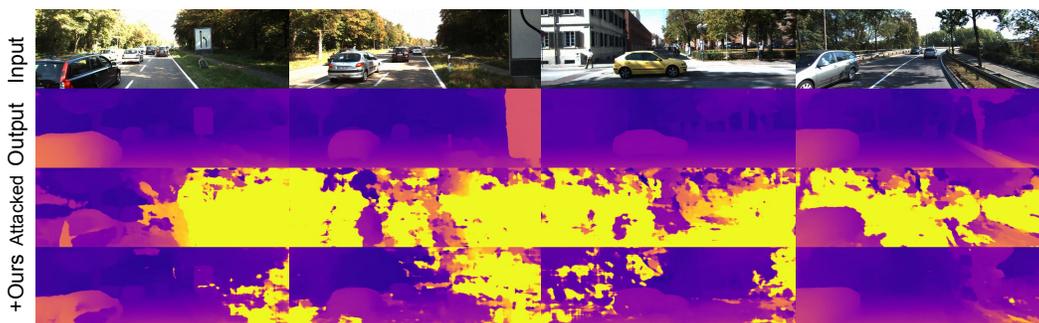


Figure 9: Visual results on stereo matching against PGD attack, without and with SR,  $\epsilon = 0.01$

972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984

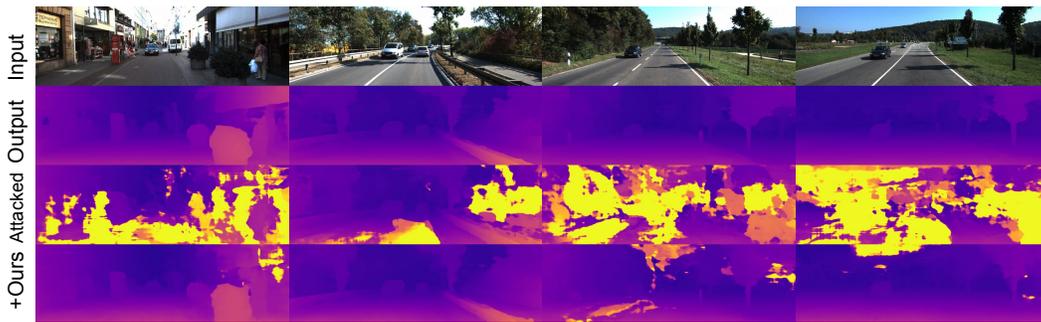


Figure 10: Visual results on stereo matching against PGD attack, without and with SR,  $\epsilon = 0.005$

985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000

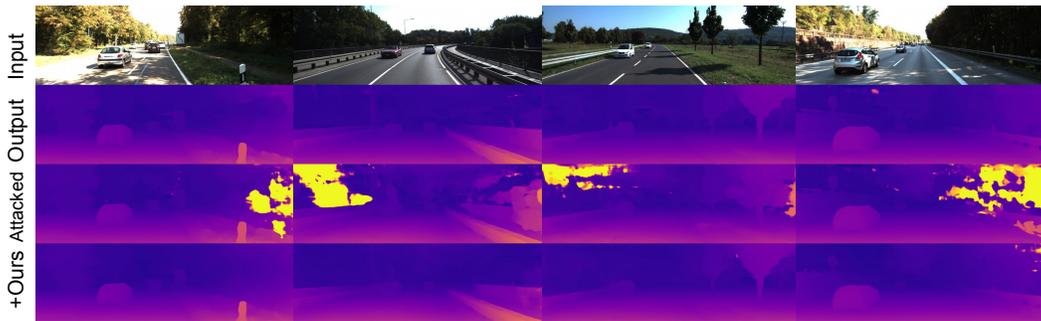


Figure 11: Visual results on stereo matching against PGD attack, without and with SR,  $\epsilon = 0.002$

1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021

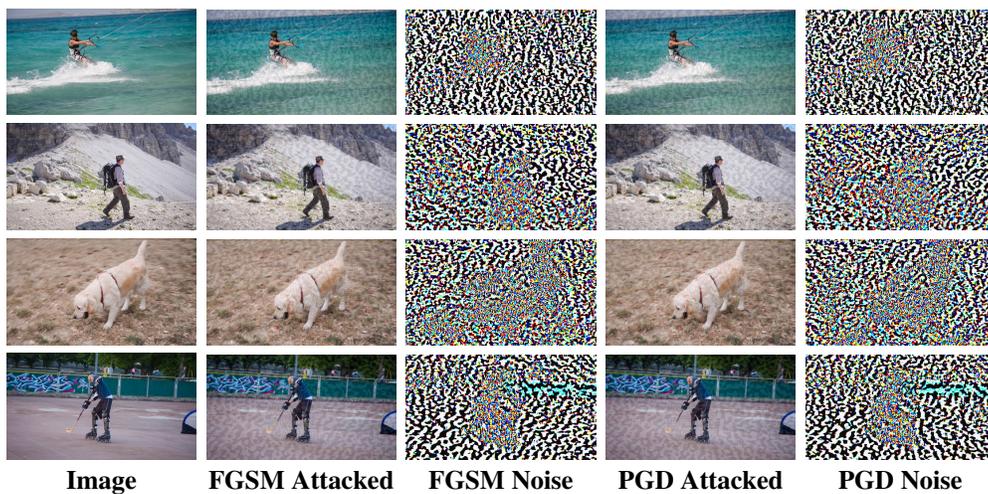


Figure 12: Original images and their corresponding attacked images and perturbations using FGSM and PGD methods on optical flow. The attacks mostly target the main object observed in the image.

1022  
1023  
1024  
1025