mathematics

MDPI

Article

Attack-Dependent Composition Effects and Defense Selection in Adversarial Image Classification

Nataliia Kussul [1,2], Bohdan Omelchenko [2,*] and Andrii Shelestov [2,3]

[1] Department of Geographical Sciences, University of Maryland, College Park, MD 20742, USA; kussul@umd.edu
[2] Department of Mathematical Modelling and Data Analysis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 03056 Kyiv, Ukraine; andrii.shelestov@edu.kpi.ua
[3] Space Research Institute of the National Academy of Sciences of Ukraine and the State Space Agency of Ukraine, 03187 Kyiv, Ukraine
* Correspondence: omelchenko.bohdan_fi31f@edu.kpi.ua or bogdan.r.omelchenko@gmail.com

Abstract

Input-preprocessing defenses are often combined, and studies of such combinations ask whether a composition improves robustness. When the effect of adding a component changes sign across attacks, this leaves open which pipeline should be deployed. We treat the choice among preprocessing pipelines as a decision problem over a finite table of recovery rates and characterize its geometry: a pipeline is optimal for some attack weights exactly when no mixture of the other pipelines strictly dominates it, each pipeline's winning region is a convex polytope, and preferences switch on hyperplanes. We apply this framework to Gaussian perturbation, feature squeezing, spatial smoothing, and their pairwise compositions, using DenseNet169 and ResNet-50 on ImageNetV2 (614 and 569 correctly classified images from as many classes) and four attacks generated against the undefended classifiers. On both architectures, adding Gaussian noise after median smoothing raises recovery under PGD, FGSM, and BIM by 7.0–16.4 percentage points but lowers it under Boundary Attack by 22.1–27.6 points relative to the better constituent. Mean and maximin recovery select this composition, whereas minimax regret selects smoothing alone. On both architectures, removing dominated pipelines discards the identity and feature squeezing with smoothing; for DenseNet169, feature squeezing with Gaussian perturbation wins no individual attack yet is the unique optimum for mixtures of FGSM and Boundary Attack, and it remains optimal for some attack weights in 96% of bootstrap replicates. When clean accuracy is weighted by an attack probability, no preprocessing is preferred below a probability of about 0.26 for DenseNet169 and 0.42 for ResNet-50. An adaptive BPDA/EOT attack reduces the recovery of every tested pipeline to zero, so the empirical results describe a non-adaptive threat model.

Keywords: adversarial examples; image classification; input preprocessing; defense selection; minimax regret; Gaussian perturbation; feature squeezing; spatial smoothing

MSC: 68T07; 68T45; 91B06; 90C05; 62C20

Check for updates

Academic Editor: Hanzhou Wu

Received: 21 September 2026
Revised: 12 November 2026
Accepted: 13 November 2026
Published: 16 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Input preprocessing offers a way to intervene when a classifier is already trained and its weights must remain fixed. Studies of combined input transformations ask whether a composition improves robustness [1,2]. Adding a second transformation, however, can help

against one attack and harm performance against another. When gains and losses differ across attacks, evidence that a composition helps on average leaves open which pipeline should be deployed. This paper addresses that selection problem: Given recovery rates of candidate pipelines under several attacks, which candidates can be discarded, and for which attack mixtures and decision criteria does each remaining candidate become optimal?

Adversarial examples exploit the sensitivity of learned decision functions to small, deliberately chosen input changes [3]. FGSM uses a single gradient-sign step [4]; BIM and PGD use iterative updates [5,6]. The DeepFool and Carlini–Wagner attacks pursue different geometric or optimization objectives [7,8]. Boundary Attack works with predicted labels [9], whereas Square Attack uses prediction scores [10]. These differences matter for preprocessing: a transformation that disrupts one attack's perturbations can preserve those produced by another. Our earlier study assessed the resilience of pattern-recognition systems to malicious perturbations [11]; the present work moves from measuring this vulnerability to selecting among preprocessing defenses.

Input transformations also differ in what they retain and remove. Guo et al. compare resizing, bit-depth reduction, compression, and reconstruction, examining both pretrained classifiers and models trained on transformed inputs [1]. Their comparison makes preservation of class information as important as removal of perturbations. JPEG compression illustrates this trade-off: its benefit depends on perturbation magnitude [12]. Other defenses build multi-stage procedures, including randomized resizing and padding [13], and compression-based randomization with retraining and ensembling in SHIELD [14]. Training-free input transformations are still used in recent test-time defenses, for example, to detect and correct adversarial feature shifts in vision–language models [15]. Xu et al. use feature squeezing to detect adversarial inputs by comparing predictions before and after squeezing [16]. Here, we use the transformed input for classification; our outcome is recovery of the correct label.

The role of Gaussian noise depends on where it enters the procedure. Noise can regularize training [17], hinder black-box query attacks [18], or accompany randomized discretization [19]. Certified smoothing instead aggregates predictions to establish guarantees under specified assumptions [20–22]. The closest compositional study is the preprint by Stalder et al. [2]. They combine Gaussian noise with repeated bilateral filtering, develop a theoretical account of complementary mechanisms, and test preprocessing during training and inference, including integration with adversarial training. Their central question is whether the components yield supralinear robustness gains. We ask instead how to choose among alternative pipelines when the classifier remains fixed and the preferred pipeline changes with the attack, the attack weights, or the decision criterion.

Learned purification and denoising introduce additional models or learned representations, as in PixelDefend, Defense-GAN, representation-guided denoising, feature denoising, diffusion purification, and image translation [23–28]. Diffusion-based purification has been analyzed further [29,30] and extended with purifiers that mimic the clean diffusion process, bridge adversarial and clean distributions, or purify in a single step, as well as with diffusion models used directly as robust classifiers [31–34]; its robustness depends strongly on how it is attacked [35–38]; a recent survey organizes these and other defenses [39]. Ensemble adversarial training and TRADES modify learning itself [40,41], while the comparisons here concern transformations applied at inference to an unchanged classifier.

Composing operations also changes how a defense must be tested. Randomness and nondifferentiability can obstruct attack gradients, making a weak attack overstate robustness [42,43]. Adaptive evaluations address the complete preprocessing–classifier system [44,45], and many test-time defenses lose most of their robustness under such evaluation [46]; AutoAttack and RobustBench support comparisons under standardized

threat models [47,48]. Our main analysis concerns a specific stage of this problem, recovery after transforming fixed adversarial images generated against the original classifier, and we report an adaptive evaluation separately.

Multiple attacks introduce a separate choice of objective. Maini et al. train against the union of several perturbation models, showing why balanced performance across threats matters for worst-case robustness [49]; later work improves such union robustness while retaining clean accuracy [50], and MultiRobustBench evaluates robustness against many attacks simultaneously [51]. Averaging attack-wise accuracies allows a gain against one attack to offset a loss against another. Taking the minimum prevents that compensation, while minimax regret limits the gap to the best available choice in each case. The latter two criteria follow classical decision theory: Wald's minimax principle minimizes the largest loss [52], and Savage's regret criterion uses the loss relative to the best action for each state [53]. We apply these criteria to a finite table of recovery rates. Its minimum is taken across attack-wise averages; robustness to a union of threat models takes a worst case separately for each input.

We pose two research questions. RQ1: Under which attacks does a pair of transformations recover more correct predictions than its better constituent, and where does it reduce recovery? RQ2: How do mean recovery, minimum recovery, and minimax regret change the selected pipeline, and which attack mixtures favor each candidate?

We examine these questions with two pretrained classifiers, DenseNet169 [54] and ResNet-50 [55], on ImageNetV2 [56]. The contributions are as follows:

(i) We characterize the geometry of mixture-optimal selection among a finite set of pipelines (Section 2.5): optimality for some attack weights is equivalent to the absence of strict domination by a mixture, with a winning margin computed by linear programming; winning regions are convex polytopes, and any preference threshold along a family of weights is the crossing of a hyperplane.

(ii) We show that the composition effect is attack-dependent on both architectures, with bootstrap intervals and paired tests: Gaussian noise added after median smoothing gains under the gradient-based attacks and loses under Boundary Attack.

(iii) We show how the decision criterion, the attack weights, and the cost on clean inputs change the selected pipeline: mean and maximin recovery select a different pipeline from the one selected by minimax regret, dominance removes candidates while preserving the optimum for every attack mixture, and we report the thresholds in attack weight and attack probability at which the preferred pipeline changes.

2. Materials and Methods

2.1. Classifiers and Evaluation Samples

We used two classifiers with pretrained ImageNet-1K weights from Keras Applications (Google LLC, Mountain View, CA, USA), DenseNet169 [54] and ResNet-50 [55], in TensorFlow 2.7.0 and Keras 2.7.0 (Google LLC, Mountain View, CA, USA), keeping the weights fixed throughout. Attacks and preprocessing operations were taken from the Adversarial Robustness Toolbox (ART) 1.8.1 (IBM Research, Dublin, Ireland) [57]. Each classifier receives RGB images of size 224×224 on the 0–255 scale; its model-specific input normalization (the Keras `torch` mode for DenseNet169 and the `caffe` mode for ResNet-50) is part of the classifier, so attacks differentiate through it exactly.

The main sample comes from ImageNetV2 [56] (matched-frequency version), a test set with ten images for each of the 1000 ImageNet classes [58,59]. We drew one image per class with a fixed seed, which gives 1000 images. DenseNet169 classified 614 of them

correctly (61.4%), and ResNet-50 classified 569 correctly (56.9%). These correctly classified images, each from a different class, form the full evaluation samples, $N = 614$ and $N = 569$. Boundary Attack requires on the order of 10^4 model queries per image, so it was run on a random subset of $n = 150$ images of each full sample (seed 0). All selection analyses use these subsets, on which every attack and the clean inputs are available for the same images. Composition effects for the gradient-based attacks and clean inputs are also reported on the full samples.

The original submission used 100 images from five classes (beagle, gazelle, koala, malamute, and marmoset) distributed with the `imagenet_stubs` package [60]; DenseNet169 classified 76 of them correctly. We retain this sample for the sensitivity analyses and the adaptive evaluation in Section 3.6.

Let $f_\theta : \mathcal{X} \to \{1, \ldots, 1000\}$ denote top-1 prediction, with fixed weights θ and image domain $\mathcal{X}$. An attack configuration a generates an input $x_i^{(a)}$ from image x_i with reference label y_i, the ImageNet class of the image. An inference-time transformation T produces the prediction $f_\theta(T(x_i^{(a)}))$. A prediction is correct when it equals y_i.

2.2. Attacks and Threat Model

We used four untargeted attacks from ART with the settings in Table 1. Each attack was generated once against the undefended classifier, using the reference labels, and each preprocessing pipeline was then applied to the resulting fixed inputs. This is a non-adaptive threat model: the attacker does not know the preprocessing. Section 3.6 reports an adaptive evaluation.

Table 1. Attack configurations. All attacks are untargeted and use the L_∞ norm where applicable. Budgets are on the 0–255 pixel scale of the ART classifier interface (`clip_values` $= (0, 255)$), so $\epsilon = 8$ corresponds to $8/255 \approx 0.031$ on $[0, 1]$.

Attack	Settings
PGD	$\epsilon = 8$, step 1, 20 iterations, and random start in the ϵ-ball (`num_random_init=1`)
FGSM	$\epsilon = 2.4$ and a single gradient-sign step
BIM	$\epsilon = 8$, step 1, 20 iterations, and start at the clean image
Boundary	`max_iter=1000`, `num_trial=25`, `sample_size=20`, and `init_size=100`; run on the $n = 150$ subset

For gradient-based attacks, a projected iterative construction can be written as

$$x^{t+1} = \Pi_{\mathcal{B}(x,\epsilon)}\left(x^t + \alpha \operatorname{sign} \nabla_{x^t} \ell_\theta(x^t, y)\right), \tag{1}$$

where $\mathcal{B}(x, \epsilon)$ combines the norm constraint with the valid input domain, and $\ell_\theta(x, y)$ is the cross-entropy loss computed from the network outputs. FGSM uses a single step. PGD starts from a uniformly random point $x^0 \in \mathcal{B}(x, \epsilon)$, as in Madry et al. [6], whereas BIM starts from $x^0 = x$ [5]. In ART 1.8.1, BIM is the same projected iteration without random initialization [61], so the random start is what distinguishes the two configurations; on the DenseNet169 subset, none of the 150 PGD and BIM adversarial inputs coincide. In the original submission, PGD had no random start and therefore coincided with BIM up to numerical differences; this applies only to the 76-image sample in Section 3.6.

The FGSM budget, $\epsilon = 2.4$ (≈ 0.0094 on $[0, 1]$), is smaller than the iterative budget; Section 3.6 repeats FGSM with $\epsilon \in \{4, 8, 16\}$. Boundary Attack uses the ART defaults for `num_trial`, `sample_size`, and `init_size`, with 1000 instead of the default 5000 iterations. Its search for an image stops earlier when no step within `num_trial` attempts remains adversarial. It produced a misclassified input for 130 of the 150 images (86.7%) with DenseNet169 and for 99 of 150 (66.0%) with ResNet-50; the remaining images are classified

correctly without preprocessing, which is reflected in the recovery of I. The mean L_2 distortions on the 0–255 scale were 1984 for PGD, 1722 for BIM, 920 for FGSM, and 1484 for Boundary Attack (median 834) with DenseNet169, and 1999, 1756, 919, and 1462 (median 827) with ResNet-50. The mean L_∞ distortions of Boundary Attack were 17.7 and 17.3, respectively.

2.3. Preprocessing Pipelines

We use G for inference-time Gaussian perturbation, F for feature squeezing by bit-depth reduction, and S for spatial smoothing by a median filter. Table 2 lists their implementations and parameters. The candidate set is

$$\mathcal{T} = \{I,\ G,\ F,\ S,\ F+G,\ G+S,\ F+S\}, \tag{2}$$

where I is the identity and each composition is applied in a fixed order: $F+G = G \circ F$ (bit-depth reduction and then noise), $G+S = G \circ S$ (median smoothing and then noise), and $F+S = S \circ F$ (bit-depth reduction and then smoothing). A label names the two components of a pipeline, and the order of application is the one defined here; we use these labels throughout the text, tables, and figures. In $G+S$, the noise is added after the median filter and therefore reaches the classifier unfiltered; Section 3.6 reports the reverse order $S \circ G$.

The Gaussian operation is

$$G_\sigma(x;\xi) = x + \sigma\xi, \qquad \xi \sim \mathcal{N}(0, I_d), \tag{3}$$

with no clipping to the input domain. Every Gaussian-containing pipeline was evaluated over $R = 20$ independent noise draws with fixed seeds. All three operations act during inference, with the classifier weights held fixed.

Table 2. Preprocessing components. All operations act on RGB images on the 0–255 scale, before the classifier's input normalization. Implementations from ART 1.8.1.

Symbol	Operation	Implementation	Parameter
G_σ	Additive Gaussian noise	`GaussianAugmentation`	$\sigma = 9$ ($\approx$0.035 on $[0,1]$), with no clipping
F_ϕ	Bit-depth reduction	`FeatureSqueezing`	$\phi = 3$ bits (8 levels) on $[0,255]$
S_ψ	Median filter	`SpatialSmoothing`	$\psi = 5 \times 5$ window

2.4. Recovery and Decision Criteria

For a stochastic transformation evaluated over R noise draws, define the mean correct-image count and the corresponding recovery percentage

$$\overline{n}_{T,a} = \frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{n}\mathbf{1}\{f_\theta(T(x_i^{(a)};\xi_{ir})) = y_i\}, \qquad A_a(T) = 100\frac{\overline{n}_{T,a}}{n}. \tag{4}$$

For a deterministic transformation, $R = 1$. All quantities below are computed from the unrounded means. We write $A_{\text{clean}}(T)$ for the same quantity on unattacked inputs.

The empirical advantage of a composition over its better constituent is

$$\Delta_a(P,Q) = A_a(P+Q) - \max\{A_a(P), A_a(Q)\}, \tag{5}$$

where $P+Q$ is the composition of components P and Q defined in Section 2.3. A positive $\Delta_a(P,Q)$ means that the composition has higher recovery than either constituent in configuration a; a negative value measures the loss relative to the better constituent.

Let $\mathcal{A} = \{\text{PGD}, \text{FGSM}, \text{BIM}, \text{Boundary}\}$, $m = |\mathcal{A}|$, and let $w = (w_a)_{a\in\mathcal{A}}$ be a vector of non-negative weights summing to one. Weighted recovery and its equal-weight special case are

$$U_w(T) = \sum_{a\in\mathcal{A}} w_a A_a(T), \qquad \overline{A}(T) = \frac{1}{m}\sum_{a\in\mathcal{A}} A_a(T). \tag{6}$$

Equal weights summarize recovery across the attacks; other weights express the relative importance assigned to each configuration.

For comparison, define the minimum attack-wise recovery

$$M(T) = \min_{a\in\mathcal{A}} A_a(T) \tag{7}$$

and the maximum regret relative to the best pipeline for each attack,

$$B_a = \max_{T\in\mathcal{T}} A_a(T), \qquad L(T) = \max_{a\in\mathcal{A}}\{B_a - A_a(T)\}. \tag{8}$$

Mean-based selection maximizes $\overline{A}$; maximin selection maximizes M, and minimax-regret selection minimizes L, following Wald and Savage, respectively [52,53]. The benchmarks B_a are the column maxima of the recovery table in Section 3.1. Because B_a is a maximum over the candidate set, regret depends on which pipelines are compared: adding or removing a candidate that attains a column maximum changes B_a and can change the minimax-regret choice [62]. The criterion $M(T)$ takes the minimum of attack-wise recovery averages, whereas robustness to a union of threat models, as studied by Maini et al. [49], takes the worst case separately for each input.

2.5. Geometry of Mixture-Optimal Selection

Write $\mathbf{A}(T) = (A_a(T))_{a\in\mathcal{A}} \in \mathbb{R}^m$ for the recovery profile of T and $\Delta^m = \{w \in \mathbb{R}^m_{\geq 0} : \sum_a w_a = 1\}$ for the simplex of attack weights. The winning region of $T \in \mathcal{T}$ is

$$W(T) = \{w \in \Delta^m : \; U_w(T) \geq U_w(T') \text{ for all } T' \in \mathcal{T}\}. \tag{9}$$

We say that T is strictly dominated by a mixture if there is a probability vector λ on $\mathcal{T}\setminus\{T\}$ with $\sum_{T'} \lambda_{T'} A_a(T') > A_a(T)$ for every $a \in \mathcal{A}$.

Proposition 1. *Define the winning margin*

$$\tau(T) = \max_{w\in\Delta^m} \min_{T'\neq T} \left[U_w(T) - U_w(T')\right]. \tag{10}$$

Then,

$$\tau(T) = \min_{\lambda\in\Delta(\mathcal{T}\setminus\{T\})} \max_{a\in\mathcal{A}} \Big[A_a(T) - \sum_{T'\neq T} \lambda_{T'} A_a(T')\Big]. \tag{11}$$

Consequently, $W(T) \neq \emptyset$ if and only if T is not strictly dominated by a mixture of the other candidates, and T is the unique maximizer of U_w for some $w \in \Delta^m$ if and only if $\tau(T) > 0$.

Proof. Let $\Phi(w,\lambda) = \sum_a w_a\big[A_a(T) - \sum_{T'\neq T}\lambda_{T'}A_a(T')\big]$. For fixed w, Φ is linear in λ, so its minimum over the simplex $\Delta(\mathcal{T}\setminus\{T\})$ is attained at a vertex and equals $\min_{T'\neq T}[U_w(T) - U_w(T')]$. For fixed λ, Φ is linear in w, so its maximum over Δ^m equals the largest coordinate, $\max_a[A_a(T) - \sum_{T'}\lambda_{T'}A_a(T')]$. Since Φ is bilinear and both domains are nonempty compact convex polytopes, von Neumann's minimax theorem [63] (equivalently, linear programming duality) gives $\max_w \min_\lambda \Phi = \min_\lambda \max_w \Phi$, which is Equation (11). Now, $W(T) \neq \emptyset$ if and only if $\tau(T) \geq 0$, that is, if and only if for every λ, some coordinate

of $\mathbf{A}(T) - \sum_{T'} \lambda_{T'} \mathbf{A}(T')$ is non-negative, which means that no mixture strictly dominates T. Likewise, a w at which T is the unique maximizer exists if and only if $\tau(T) > 0$. □

Proposition 1 is the decision-theoretic form of a classical result on best responses in games [64]. The margin $\tau(T)$, in percentage points, is the largest lead that T can hold over its best competitor under the most favorable attack weights. It is the value of a linear program with $|\mathcal{T}|$ variables and $m + 1$ inequality constraints; a negative value measures how far T is from being optimal for any weights.

Proposition 2. *Each $W(T)$ is a convex polytope, $\bigcup_{T \in \mathcal{T}} W(T) = \Delta^m$, and for $\mathbf{A}(T) \neq \mathbf{A}(T')$, the set $W(T) \cap W(T')$ lies in the hyperplane $H_{TT'} = \{w : \sum_a w_a [A_a(T) - A_a(T')] = 0\}$. The optimal value $V(w) = \max_T U_w(T)$ is convex and piecewise linear on Δ^m.*

Proof. $W(T)$ is the intersection of Δ^m with the closed half-spaces $\{w : U_w(T) - U_w(T') \geq 0\}$, $T' \neq T$, each defined by a linear inequality in w. Every w has at least one maximizer, so the regions cover Δ^m. A point of $W(T) \cap W(T')$ satisfies $U_w(T) = U_w(T')$; that is, it lies in $H_{TT'}$. Finally, V is a maximum of finitely many linear functions. □

Corollary 1. *Let $\mathcal{T}' \subset \mathcal{T}$ be such that every $T \in \mathcal{T} \setminus \mathcal{T}'$ is weakly dominated by some $T'' \in \mathcal{T}'$; that is, $A_a(T) \leq A_a(T'')$ for all a. Then, $\max_{\mathcal{T}'} U_w = \max_{\mathcal{T}} U_w$ for every $w \in \Delta^m$, $\max_{\mathcal{T}'} M = \max_{\mathcal{T}} M$, and the regret benchmarks B_a are the same over $\mathcal{T}'$ and $\mathcal{T}$; hence, $L(T)$ is unchanged for every retained T.*

Proof. Each removed T satisfies $U_w(T) \leq U_w(T'')$, $M(T) \leq M(T'')$, and $A_a(T) \leq A_a(T'') \leq \max_{\mathcal{T}'} A_a$ for a retained T''. □

Corollary 2. *For $u, v \in \Delta^m$, let $w(q) = q\,u + (1 - q)\,v$, $q \in [0, 1]$, and $D(q) = U_{w(q)}(T) - U_{w(q)}(T')$. Then, $D(q) = (1 - q)D(0) + qD(1)$, and if $D(0)D(1) < 0$, the preference between T and T' switches exactly once, at*

$$q^* = \frac{D(0)}{D(0) - D(1)}, \tag{12}$$

the point where the segment crosses $H_{TT'}$.

Proof. U_w is linear in w, so D is affine in q and has at most one zero when its endpoint values have opposite signs. □

Corollary 2 concerns a pair of pipelines. The globally optimal pipeline along a segment changes where the segment leaves one winning region of Proposition 2 and enters another; such a switch can occur before the pairwise threshold when a third candidate becomes optimal in between.

2.6. Clean Inputs and Attack Probability

Preprocessing also changes predictions on unattacked inputs. For an attack probability $\pi \in [0, 1]$ and attack weights w, the expected recovery is

$$V_{\pi,w}(T) = (1 - \pi)\, A_{\text{clean}}(T) + \pi\, U_w(T). \tag{13}$$

Pipeline T is preferred to T' exactly when $\pi\,[U_w(T) - U_w(T')] > (1 - \pi)\,[A_{\text{clean}}(T') - A_{\text{clean}}(T)]$, which gives a threshold attack probability for each pair. Equation (13) is a special case of U_w with clean inputs as an additional scenario of weight $1 - \pi$, so Propositions 1 and 2 apply to it. We also report the three criteria with clean inputs as a fifth, equally weighted scenario.

2.7. Uncertainty Quantification

Sampling uncertainty was assessed with a paired nonparametric bootstrap over images: in each replicate, images were drawn with replacement from the evaluation sample, and the same draw was used for every pipeline and scenario. For stochastic pipelines, each image contributes its mean correctness over the R noise draws. We report 95% percentile intervals for the composition advantages in Equation (5) (4000 replicates), the frequency with which each criterion selects each pipeline (10,000 replicates; 2000 when clean inputs are added as a fifth scenario), and the winning margins and thresholds (2000 replicates). Within each scenario, S and $G + S$ were compared with the exact McNemar test on per-image correctness (majority vote over the noise draws for $G + S$), with the Holm correction across the five scenarios.

3. Results

3.1. Recovery Profiles

Table 3 gives recovery on the 150-image subsets. Without preprocessing, the gradient-based attacks leave at most 5.3% of the images correctly classified. Gaussian perturbation and feature squeezing recover most images under Boundary Attack, 88.47% and 87.33% with DenseNet169, but at most 23.33% under FGSM and at most 1.33% under PGD and BIM. Smoothing with Gaussian perturbation leads under PGD, FGSM, and BIM on both architectures, whereas Gaussian perturbation alone leads under Boundary Attack.

Table 3. Recovery (%) on the 150-image subsets. Numerical entries for pipelines containing G are means over $R = 20$ noise draws; the standard deviation across draws is at most 2.2 percentage points. Bold entries are the column maxima over the attacks, i.e., the regret benchmarks B_a in Equation (8). The last column is recovery on the same images without attack.

Pipeline	PGD	FGSM	BIM	Boundary	Clean
(a) DenseNet169					
I	0.00	5.33	0.00	13.33	100.00
G	0.67	19.67	0.67	**88.47**	92.03
F	0.67	23.33	1.33	87.33	88.67
S	24.67	42.00	22.00	75.33	76.67
$F + G$	4.20	37.27	4.57	84.63	87.40
$G + S$	**40.30**	**54.30**	**39.67**	66.37	67.13
$F + S$	32.67	38.67	31.33	46.67	42.67
(b) ResNet-50					
I	0.00	4.00	0.00	34.00	100.00
G	0.00	11.37	0.00	**82.27**	90.00
F	0.67	16.00	1.33	81.33	87.33
S	26.67	40.67	23.33	62.00	64.67
$F + G$	2.50	23.87	1.90	76.87	84.17
$G + S$	**33.37**	**44.67**	**33.60**	54.63	57.77
$F + S$	24.00	26.00	20.67	36.67	34.00

Preprocessing also costs accuracy on clean inputs. On the full DenseNet169 sample, recovery of clean images is 92.86% for G, 89.09% for F, 86.89% for $F + G$, 72.96% for S, 64.34% for $G + S$, and 43.49% for $F + S$. On all 1000 sampled images, clean top-1 accuracy falls from 61.40% without preprocessing to 48.20% with S and 43.22% with $G + S$ (ResNet-50: from 56.90% to 40.80% and 37.71%).

3.2. Composition Effects

Table 4 reports the advantage of each composition over its better constituent. Adding Gaussian noise after smoothing raises recovery under PGD, FGSM, and BIM by 14.50, 8.35, and 16.43 percentage points with DenseNet169 and by 9.24, 7.03, and 14.08 points with

ResNet-50; under Boundary Attack, it lowers recovery by 22.10 and 27.63 points relative to Gaussian perturbation alone, and on clean inputs by 28.52 and 31.09 points. All of the corresponding intervals exclude zero. The paired McNemar tests agree with the sign pattern. On the full DenseNet169 sample, $G + S$ is correct where S is not for 102, 82, and 117 images under PGD, FGSM, and BIM, against 13, 31, and 13 images in the opposite direction ($p < 10^{-5}$ after the Holm correction); under Boundary Attack, S is correct where $G + S$ is not for 19 images against seven ($p = 0.029$). For ResNet-50, the corresponding counts are 73, 71, and 94 against 25, 25, and 16 ($p < 10^{-5}$), and 24 against 14 under Boundary Attack ($p = 0.14$).

Table 4. Composition advantage Δ_a over the better constituent (percentage points) with 95% bootstrap intervals. PGD, FGSM, BIM, and Clean use the full samples ($N = 614$ and $N = 569$); Boundary uses the 150-image subsets.

Composition	PGD	FGSM	BIM	Boundary	Clean
(a) DenseNet169					
$F + G$	+1.71 [+0.8, +2.8]	+10.04 [+7.8, +12.3]	+1.60 [+0.7, +2.6]	−3.83 [−7.7, −1.4]	−5.97 [−7.8, −4.3]
$G + S$	+14.50 [+11.6, +17.4]	+8.35 [+5.4, +11.4]	+16.43 [+13.4, +19.6]	−22.10 [−28.6, −15.9]	−28.52 [−31.5, −25.5]
$F + S$	+7.65 [+3.9, +11.4]	−3.91 [−7.8, +0.0]	+10.10 [+6.2, +13.8]	−40.67 [−49.3, −32.0]	−45.60 [−49.8, −41.4]
(b) ResNet-50					
$F + G$	+2.39 [+1.3, +3.6]	+7.78 [+5.7, +10.0]	+1.63 [+0.7, +2.7]	−5.40 [−9.9, −2.9]	−8.66 [−10.8, −6.6]
$G + S$	+9.24 [+6.2, +12.3]	+7.03 [+4.1, +9.9]	+14.08 [+10.9, +17.1]	−27.63 [−34.6, −20.8]	−31.09 [−34.8, −27.4]
$F + S$	−1.23 [−5.1, +2.5]	−11.95 [−15.8, −8.3]	+1.58 [−2.1, +5.3]	−44.67 [−53.3, −36.0]	−48.33 [−52.7, −43.9]

Feature squeezing with Gaussian perturbation gains most under FGSM (10.04 and 7.78 points) and slightly under the iterative attacks and loses under Boundary Attack and on clean inputs. Feature squeezing with smoothing gains under PGD and BIM with DenseNet169 but loses heavily under Boundary Attack and on clean inputs, and with ResNet-50, its advantage is negative under FGSM, and its intervals include zero under PGD and BIM. RQ1 therefore has the same answer on both architectures: the compositions with Gaussian noise help against the gradient-based attacks and hurt against the decision-based Boundary Attack.

3.3. Dominance and Winning Regions

On both subsets, I is weakly dominated by a single-component pipeline and $F + S$ by $G + S$, so Corollary 1 removes them without changing the optimum for any attack mixture, the maximin value, or the regret benchmarks. Table 5 lists the winning margins τ of Proposition 1. With DenseNet169, $G + S$, $F + G$, G, and S have positive margins, whereas F is strictly dominated by a mixture of the others ($\tau = -0.25$). $G + S$ wins the PGD, FGSM, and BIM columns, and G wins the Boundary column. $F + G$ wins no column, yet it is the unique optimum for mixtures of FGSM and Boundary Attack; its largest lead, 3.81 points, occurs at $w_{\text{FGSM}} = 0.39$ and $w_{\text{Boundary}} = 0.61$. S also wins no column and is optimal only with a margin of 0.27 points. The bootstrap shows which of these conclusions are stable: $\tau > 0$ holds in 100% of replicates for $G + S$, 95.7% for $F + G$, 68.1% for G, 56.9% for S, and 46.0% for F. With ResNet-50, $G + S$ again has a clear margin (10.27 points, positive in 100% of replicates), whereas G, F, S, and $F + G$ have margins between 0.75 and 1.49 points that are positive in 63–75% of replicates.

Table 5. Selection criteria on the 150-image subsets: mean and minimum recovery over the four attacks (%, to maximize), maximum regret (percentage points, to minimize), and the winning margin τ of Proposition 1 (percentage points). Bold entries indicate the selected pipelines.

	DenseNet169				ResNet-50			
Pipeline	**Mean**	**Min**	**Regret**	τ	**Mean**	**Min**	**Regret**	τ
I	4.67	0.00	75.13	−39.67	9.50	0.00	48.27	−26.45
G	27.37	0.67	39.63	+1.13	23.41	0.00	33.60	+0.93
F	28.17	0.67	39.63	−0.25	24.83	0.67	32.70	+0.75
S	41.00	22.00	**17.67**	+0.27	38.17	23.33	**20.27**	+1.49
$F+G$	32.67	4.20	36.10	+3.81	26.28	1.90	31.70	+0.95
$G+S$	**50.16**	**39.67**	22.10	+13.71	**41.57**	**33.37**	27.63	+10.27
$F+S$	37.33	31.33	41.80	−7.63	26.83	20.67	45.60	−9.37

Figure 1 shows the winning regions on the face of the attack simplex where PGD and BIM receive equal weight. $G+S$ occupies most of the face on both architectures, and the other candidates share the neighborhood of the FGSM–Boundary edge.

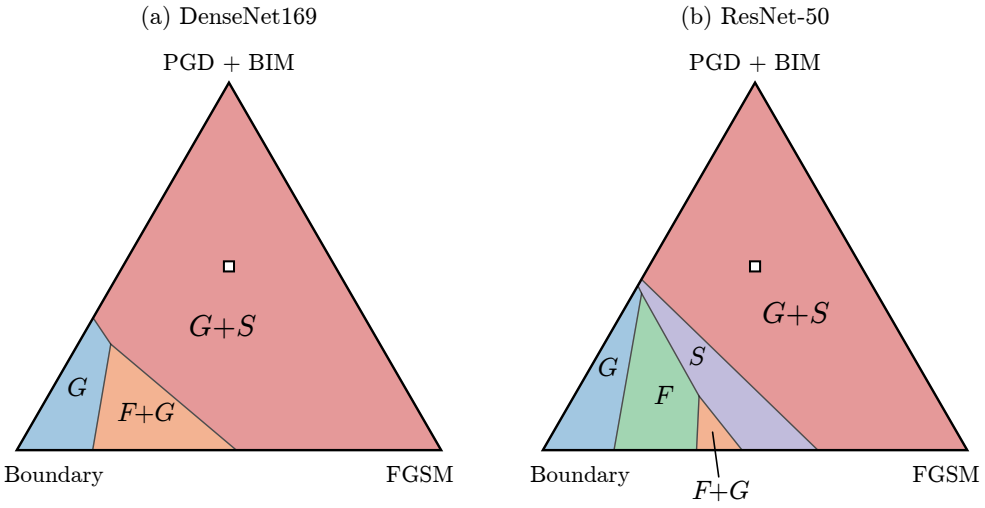


Figure 1. Winning regions $W(T)$ (Proposition 2) on the cross-section $w_{\mathrm{PGD}} = w_{\mathrm{BIM}}$ of the attack simplex, for the 150-image subsets: (**a**) DenseNet169; (**b**) ResNet-50. Vertices: All weight on the iterative pair (top), on FGSM (right), and on Boundary Attack (left). The square marks equal weights over the four attacks. Pipelines with empty regions on this face are not optimal for any weights in it. In panel (**b**), the label of the smallest region, $F+G$, is placed below the triangle.

3.4. Effect of the Decision Criterion

Table 5 gives the three criteria. With DenseNet169, $G+S$ has the highest mean recovery (50.16% against 41.00% for S) and the highest minimum (39.67% against 22.00%), whereas S has the smallest maximum regret (17.67 against 22.10 points). The largest regret of S comes from BIM, where it trails $G+S$ by 17.67 points; the regret of $G+S$ comes from Boundary Attack, where it trails G by 22.10 points. In bootstrap replicates, mean recovery selects $G+S$ in 100% of cases, maximin selection chooses $G+S$ in 99.2%, and minimax regret chooses S in 80.5%. ResNet-50 gives the same pattern: $G+S$ is selected by mean recovery (41.57% against 38.17%; 91.8% of replicates) and by maximin recovery (33.37% against 23.33%; 99.9%), and S by minimax regret (20.27 against 27.63 points; 94.2%); here, the largest regret of both pipelines comes from Boundary Attack.

The answer to RQ2 is therefore that the criterion matters. Mean and maximin recovery favor the composition, which is far ahead under the three gradient-based attacks; minimax regret favors smoothing alone, which stays closer to the best pipeline under every attack,

including Boundary Attack. S is selected by minimax regret even though its winning margin for weighted recovery is small on both architectures (0.27 and 1.49 points): Savage's criterion is not a weighted mean and can select a pipeline that is optimal for few or no attack mixtures. With clean inputs as a fifth, equally weighted scenario, DenseNet169 gives the same selections (mean 53.55% and minimum 39.67% for $G + S$; regret 23.33 points for S, selected in 98.2% of replicates). For ResNet-50, mean and maximin recovery again select $G + S$, whereas minimax regret selects $F + G$ (31.70 points; 63.2% of replicates).

3.5. Attack Weights and Attack Probability

Corollary 2 gives the boundary between $G + S$ and its strongest competitor under Boundary Attack, Gaussian perturbation alone. With DenseNet169, $U_w(G + S) - U_w(G) = 39.63\, w_{\text{PGD}} + 34.63\, w_{\text{FGSM}} + 39.00\, w_{\text{BIM}} - 22.10\, w_{\text{Boundary}}$. Along the family that assigns weight b to Boundary Attack and splits $1 - b$ equally among the other three attacks, the pairwise threshold is $b^* = 0.631$. The global optimum switches slightly earlier: $G + S$ is optimal for $b < 0.617$ (95% interval 0.446–0.700), $F + G$ for $0.617 < b < 0.685$, and G beyond. With ResNet-50, $G + S$ is optimal for $b < 0.487$ (0.155–0.618), followed by S, F, and finally G for $b > 0.703$. Hence, the composition remains the best choice unless Boundary Attack receives about half or more of the total weight.

Figure 2 adds the cost on clean inputs through Equation (13), with equal weights over the four attacks. With DenseNet169, no preprocessing is optimal for attack probabilities $\pi < 0.26$ (95% interval 0.17–0.35), Gaussian perturbation and then $F + G$ are optimal for intermediate values, and $G + S$ is optimal for $\pi > 0.54$ (0.42–0.74). With ResNet-50, no preprocessing is optimal for $\pi < 0.42$ (0.29–0.52), and $G + S$ becomes optimal for $\pi > 0.67$; in 9.5% of bootstrap replicates, it is never optimal along this family. A defense chosen for attacked inputs alone therefore pays a substantial price when attacks are rare.

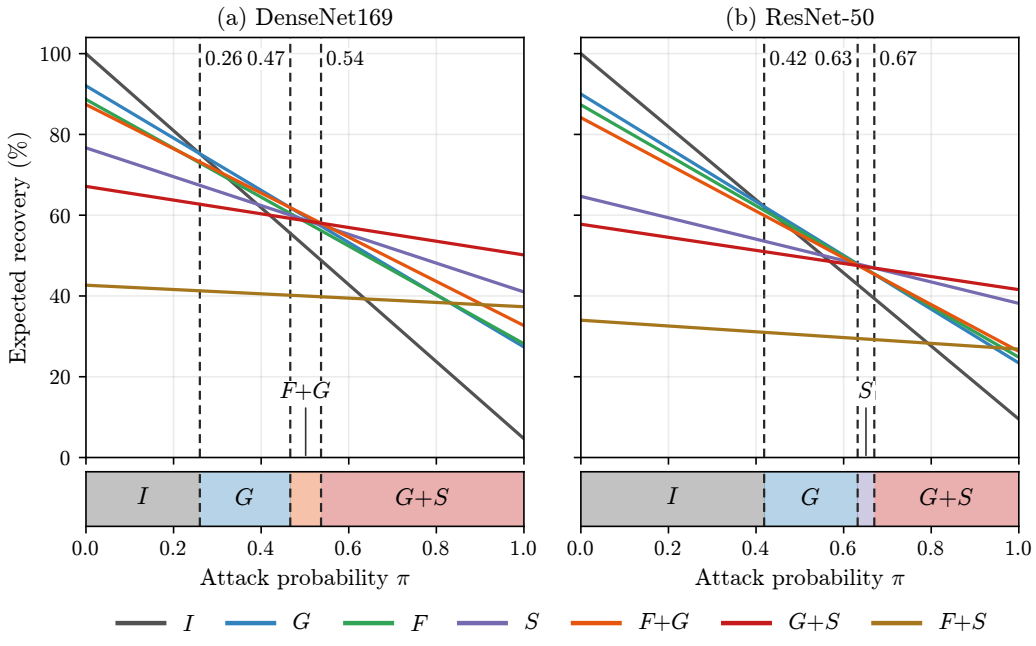


Figure 2. Expected recovery $V_{\pi,w}$ (Equation (13)) as a function of the attack probability π, with equal weights over the four attacks, for the 150-image subsets: (**a**) DenseNet169; (**b**) ResNet-50. Dashed lines mark the probabilities at which the optimal pipeline changes, with their values at the top of each panel. The bar below each panel shows the optimal pipeline in each interval; the narrowest interval in each panel is labeled above the bar.

3.6. Sensitivity Analyses and Adaptive Attack

Order of operations. With the reverse order $S \circ G$ (noise and then smoothing), DenseNet169 recovers 36.31%, 49.37%, and 34.14% under PGD, FGSM, and BIM on the full sample (against 39.10%, 51.02%, and 38.26% for $G \circ S$), 69.07% under Boundary Attack on the subset (against 66.37%), and 69.41% of clean images (against 64.34%). The order therefore matters: smoothing the noise improves recovery on clean inputs and under Boundary Attack but weakens the gains under the gradient-based attacks. The sign pattern of the composition advantage is unchanged, with $+11.7$, $+6.7$, $+12.3$, and -19.4 points for $S \circ G$ under PGD, FGSM, BIM, and Boundary Attack.

The remaining checks use the original 76-image sample with DenseNet169 (Section 2.1); there, PGD had no random start, and the main Boundary column reports results obtained with 100 iterations.

Noise level. Table 6 reports $\Delta_a(G, S)$ for $\sigma \in \{4, 9, 16\}$. The sign pattern (gains under the gradient-based attacks and a loss under Boundary Attack) holds for all three noise levels.

FGSM budget. With $\epsilon \in \{2.4, 4, 8, 16\}$, $\Delta_{\text{FGSM}}(G, S)$ equals $+17.04$, $+19.87$, $+12.30$, and $+4.28$ points, respectively, so the gain under FGSM does not depend on the smaller main budget.

Boundary Attack budget. Increasing the search from 100 to 1000 iterations reduced the mean L_2 distortion from 5879 to 1533 and the mean L_∞ distortion from 68.3 to 18.1, while recovery of G and F rose rather than fell and $\Delta_{\text{Boundary}}(G, S)$ changed from -22.89 to -25.59 points. The high recovery under Boundary Attack therefore reflects the proximity of its outputs to the decision boundary rather than large distortion.

Table 6. Composition advantage $\Delta_a(G, S)$ (percentage points) for three noise levels on the original 76-image sample (DenseNet169; Boundary Attack with 100 iterations).

Noise Level	PGD	FGSM	BIM	Boundary
$\sigma = 4$	$+14.01$	$+17.43$	$+13.36$	-16.64
$\sigma = 9$	$+28.03$	$+17.04$	$+27.89$	-22.89
$\sigma = 16$	$+31.32$	$+12.89$	$+29.54$	-24.21

Adaptive attack. We attacked F, S, $F + G$, and $G + S$ with PGD under the same budget, using backward-pass differentiable approximation for the bit-depth reduction and the median filter [42] and expectation over transformation with ten noise draws for the Gaussian component [65]. Recovery fell to 0 of 76 images for all four pipelines. The selection analysis above therefore applies to an attacker who does not adapt to the preprocessing; against an adaptive attacker, none of these pipelines provides protection, consistent with [42,45,46].

4. Discussion

4.1. Attack-Dependent Composition Effects

Adding Gaussian noise after median smoothing changes the recovery profile in opposite directions across attacks, and this sign pattern is reproduced on two architectures, on two samples, for three noise levels, for four FGSM budgets, for two Boundary Attack budgets, and for both orders of the two operations. The gradient-based attacks place a structured perturbation within a fixed budget; smoothing removes part of it, and the added noise further disrupts what remains. Boundary Attack instead returns inputs just across the decision boundary. Any preprocessing moves such inputs, which explains the high recovery of G and F, but the median filter also removes class information, and the noise added after it is not filtered. The losses under Boundary Attack closely track the losses on clean inputs (Table 4), which suggests that the cost of $G + S$ on nearly clean inputs drives

its loss under this attack. This remains a hypothesis: aggregate recovery shows the net change in correct predictions but does not identify which image features produce it.

Stalder et al. study Gaussian noise with repeated bilateral filtering and emphasize complementary mechanisms and integration with training [2]. Our comparison concerns three operations and their pairs at inference, followed by selection according to recovery, minimum performance, regret, and clean accuracy. The gain of $F + G$ under FGSM illustrates why these are separate questions: the pair improves on both of its constituents but remains below $G + S$ under that attack, and its value appears only for mixtures of FGSM and Boundary Attack.

4.2. Implications for Defense Selection

The choice of objective changes the selected pipeline. Mean and maximin recovery favor $G + S$; minimax regret favors S; and when clean inputs carry weight through the attack probability, no preprocessing or a single Gaussian perturbation is preferable for low probabilities. These are not contradictions to be resolved by the data alone: they reflect different deployment priorities, and the thresholds in Sections 3.4 and 3.5 state where each priority leads.

Dominance analysis separates robust from fragile conclusions. Removing I and $F + S$ is safe for every mixture by Corollary 1. The need for $G + S$ is beyond doubt on both architectures, and the need for $F + G$ is stable for DenseNet169 (95.7% of replicates) but not for ResNet-50. The remaining candidates have margins of about one percentage point, so whether they are needed depends on the sample. The winning margin τ and its bootstrap distribution make this distinction explicit, which a list of column winners would not: two of the four DenseNet169 candidates with positive margins win no individual attack.

The weights must be chosen to represent the intended use. Equal weights assign one-quarter of the objective to each attack; PGD and BIM differ only in the random start, so equal weights give the projected-gradient iteration half of the total weight. Other priorities lead to other weights, and Figure 1 shows where they lead.

4.3. Scope and Limitations

Recovery was measured against attacks generated for the undefended classifiers. The adaptive evaluation shows that this setting is optimistic: an attacker who differentiates through the preprocessing defeats all tested pipelines. The selection procedure itself does not depend on how the recovery table was obtained, and it can be applied unchanged to recovery under adaptive attacks when some pipelines retain non-zero recovery.

The samples are moderate in size. The full ImageNetV2 samples contain 614 and 569 images from as many classes, but the selection analyses use the 150-image subsets on which Boundary Attack was run, so differences of a few percentage points are uncertain; the bootstrap intervals quantify this. Boundary Attack used 1000 iterations and did not find an adversarial input for 13.3% (DenseNet169) and 34.0% (ResNet-50) of the subset images. We studied two convolutional architectures, three simple operations, and their pairs in a fixed order (the reverse order was checked only for G and S); other architectures, longer compositions, learned purification, and certified defenses may produce different profiles. The analysis of Section 2.5 applies unchanged to any finite set of candidate pipelines.

5. Conclusions

Adding Gaussian noise after median smoothing helps against the gradient-based attacks and hurts against Boundary Attack, on both DenseNet169 and ResNet-50 and across all sensitivity checks. Because the effect changes sign across attacks, the preferred pipeline

depends on the decision criterion: mean and maximin recovery select the composition, whereas minimax regret selects smoothing alone.

We characterized mixture-optimal selection among a finite set of pipelines: a pipeline is optimal for some attack weights exactly when no mixture of the others strictly dominates it, its winning region is a convex polytope, and preference thresholds are hyperplane crossings. Dominance removes the identity and feature squeezing with smoothing; for DenseNet169, feature squeezing with Gaussian perturbation is needed even though it wins no individual attack.

The composition remains optimal unless Boundary Attack receives about half or more of the attack weight, and when clean accuracy is included through an attack probability, no preprocessing is preferable below a probability of about 0.26 (DenseNet169) or 0.42 (ResNet-50). All empirical conclusions concern a non-adaptive attacker; under an adaptive attack, none of the tested pipelines retains any recovery.

Author Contributions: Conceptualization, N.K.; methodology, N.K. and A.S.; software, B.O.; validation, A.S.; formal analysis, N.K.; investigation, B.O.; data curation, B.O.; writing—original draft preparation, B.O.; writing—review and editing, N.K. and A.S.; visualization, B.O.; and supervision, N.K. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Ministry of Education and Science of Ukraine under the project "Next-Generation Copernicus Services: An Intelligent Digital Twin for Forest Cover Analysis in Ukraine" (State Registration No. 0126U001302) and by the Horizon Europe project FUTUREFOR "Copernicus Applications for Next-Generation Forest Monitoring" (Grant Agreement No. 101180278).

Data Availability Statement: The images used in this study are openly available on GitHub at https://github.com/modestyachts/ImageNetV2 (accessed on 15 September 2026) and https://github.com/nottombrown/imagenet_stubs (accessed on 15 September 2026). The code, the sampled image identifiers, and the per-image predictions are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Guo, C.; Rana, M.; Cisse, M.; van der Maaten, L. Countering Adversarial Images using Input Transformations. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://openreview.net/forum?id=SyJ7ClWCb (accessed on 12 November 2026).
2. Stalder, N.; Grewe, B.F.; Saponati, M.; Vilimelis Aceituno, P. A combination of noise and bilateral filters achieve supralinear and scalable adversarial robustness in CNNs. *arXiv* **2026**, arXiv:2606.02267. https://doi.org/10.48550/arXiv.2606.02267.
3. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the ICLR, Banff, AB, Canada, 14–16 April 2014.
4. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015.
5. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. *arXiv* **2016**, arXiv:1607.02533. https://doi.org/10.48550/arXiv.1607.02533.
6. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://openreview.net/forum?id=rJzIBfZAb (accessed on 12 November 2026).
7. Moosavi-Dezfooli, S.-M.; Fawzi, A.; Frossard, P. DeepFool: A simple and accurate method to fool deep neural networks. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582. https://doi.org/10.1109/CVPR.2016.282.
8. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2017; pp. 39–57. https://doi.org/10.1109/SP.2017.49.
9. Brendel, W.; Rauber, J.; Bethge, M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. *arXiv* **2018**. https://doi.org/10.48550/arXiv.1712.04248.

10. Andriushchenko, M.; Croce, F.; Flammarion, N.; Hein, M. Square Attack: A query-efficient black-box adversarial attack via random search. In Proceedings of ECCV, Glasgow, UK, 23–28 August 2020; pp. 484–501. https://doi.org/10.1007/978-3-030-58592-1_29.
11. Omelchenko, B.R.; Shelestov, A.Y. Assessment of the Resilience of Pattern Recognition Systems to the Influence of Malignant Interferences. *Int. Sci. Tech. J. Probl. Control Inform.* **2025**, *70*, 96–108. https://doi.org/10.34229/1028-0979-2025-4-6. (In Ukrainian)
12. Dziugaite, G.K.; Ghahramani, Z.; Roy, D.M. A study of the effect of JPG compression on adversarial images. *arXiv* **2016**, arXiv:1608.00853. https://doi.org/10.48550/arXiv.1608.00853.
13. Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A. Mitigating Adversarial Effects Through Randomization. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://openreview.net/forum?id=Sk9yuql0Z (accessed on 12 November 2026).
14. Das, N.; Shanbhogue, M.; Chen, S.-T.; Hohman, F.; Li, S.; Chen, L.; Kounavis, M.E.; Chau, D.H. Shield: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, UK, 19–23 August 2018. https://doi.org/10.1145/3219819.3219910.
15. Liu, L.; Chen, S.; Wu, J.; Feng, W.; Cheng, Z.; Yin, X.; Yang, W.; Zhang, T. Adversarial Attacks Already Tell the Answer: Directional Bias-Guided Test-Time Defense for Vision-Language Models. In Proceedings of the ICLR, Rio de Janeiro, Brazil, 23–27 April 2026. Available online: https://proceedings.iclr.cc/paper_files/paper/2026/file/1033880329f87ab8302b4fb39a52aebc-Paper-Conference.pdf (accessed on 12 November 2026).
16. Xu, W.; Evans, D.; Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In Proceedings of NDSS, San Diego, CA, USA, 18–21 February 2018. https://doi.org/10.14722/ndss.2018.23198.
17. Zantedeschi, V.; Nicolae, M.-I.; Rawat, A. Efficient Defenses Against Adversarial Attacks. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 39–49. https://doi.org/10.1145/3128572.3140449.
18. Qin, Z.; Fan, Y.; Zha, H.; Wu, B. Random Noise Defense against Query-Based Black-Box Attacks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34. Available online: https://proceedings.neurips.cc/paper/2021/hash/3eb414bf1c2a66a09c185d60553417b8-Abstract.html (accessed on 12 November 2026).
19. Zhang, Y.; Liang, P. Defending against Whitebox Adversarial Attacks via Randomized Discretization. In Proceedings of the AISTATS, PMLR, Naha, Japan, 16–18 April 2019; Volume 89, pp. 684–693. Available online: https://proceedings.mlr.press/v89/zhang19b.html (accessed on 12 November 2026).
20. Cohen, J.M.; Rosenfeld, E.; Kolter, J.Z. Certified Adversarial Robustness via Randomized Smoothing. In Proceedings of the ICML, PMLR, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 1310–1320. Available online: https://proceedings.mlr.press/v97/cohen19c.html (accessed on 12 November 2026).
21. Salman, H.; Yang, G.; Li, J.; Zhang, P.; Zhang, H.; Razenshteyn, I.; Bubeck, S. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32. Available online: https://proceedings.neurips.cc/paper/2019/hash/3a24b25a7b092a252166a1641ae953e7-Abstract.html (accessed on 12 November 2026).
22. Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; Jana, S. Certified Robustness to Adversarial Examples with Differential Privacy. In Proceedings of the IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 20–22 May 2019. https://doi.org/10.1109/SP.2019.00044.
23. Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; Kushman, N. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://iclr.cc/virtual/2018/poster/163 (accessed on 12 November 2026).
24. Samangouei, P.; Kabkab, M.; Chellappa, R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://openreview.net/forum?id=BkJ3ibb0- (accessed on 12 November 2026).
25. Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; Zhu, J. Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1778–1787. https://doi.org/10.1109/CVPR.2018.00191.
26. Xie, C.; Wu, Y.; van der Maaten, L.; Yuille, A.; He, K. Feature Denoising for Improving Adversarial Robustness. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019; pp. 501–509. https://doi.org/10.48550/arXiv.1812.03411.
27. Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; Anandkumar, A. Diffusion Models for Adversarial Purification. In Proceedings of the ICML, PMLR, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 16805–16827. Available online: https://proceedings.mlr.press/v162/nie22a.html (accessed on 12 November 2026).
28. Zhang, H.; Yao, Z.; Sakurai, K. Versatile Defense Against Adversarial Attacks on Image Recognition. *arXiv* **2024**, arXiv:2403.08170. https://doi.org/10.48550/arXiv.2403.08170.

29. Xiao, C.; Chen, Z.; Jin, K.; Wang, J.; Nie, W.; Liu, M.; Anandkumar, A.; Li, B.; Song, D. DensePure: Understanding Diffusion Models for Adversarial Robustness. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=p7hvOJ6Gq0i (accessed on 12 November 2026).
30. Liu, Y.; Liu, K.; Xiao, Y.; Dong, Z.; Xu, X.; Wei, P.; Lin, L. Towards Understanding the Robustness of Diffusion-Based Purification: A Stochastic Perspective. In Proceedings of the ICLR, Singapore, 24–28 April 2025. Available online: https://iclr.cc/virtual/2025/poster/28104 (accessed on 12 November 2026).
31. Song, K.; Lai, H.; Pan, Y.; Yin, J. MimicDiffusion: Purifying Adversarial Perturbation via Mimicking Clean Diffusion Model. In Proceedings of the CVPR, Seattle, WA, USA, 16–22 June 2024; pp. 24665–24674. Available online: https://openaccess.thecvf.com/content/CVPR2024/html/Song_MimicDiffusion_Purifying_Adversarial_Perturbation_via_Mimicking_Clean_Diffusion_Model_CVPR_2024_paper.html (accessed on 12 November 2026).
32. Li, X.; Sun, W.; Chen, H.; Li, Q.; He, Y.; Shi, J.; Hu, X. ADBM: Adversarial Diffusion Bridge Model for Reliable Adversarial Purification. In Proceedings of the ICLR, Singapore, 24–28 April 2025. Available online: https://proceedings.iclr.cc/paper_files/paper/2025/hash/c26719edf1e6fba8f1ca7d3938e27785-Abstract-Conference.html (accessed on 12 November 2026).
33. Lei, C.T.; Yam, H.M.; Guo, Z.; Qian, Y.; Lau, C.P. Instant Adversarial Purification with Adversarial Consistency Distillation. In Proceedings of the CVPR, Nashville, TN, USA, 11–15 June 2025; pp. 24331–24340. https://doi.org/10.48550/arXiv.2408.17064.
34. Chen, H.; Dong, Y.; Wang, Z.; Yang, X.; Duan, C.; Su, H.; Zhu, J. Robust Classification via a Single Diffusion Model. In Proceedings of the ICML, PMLR, Vienna, Austria, 21–27 July 2024; Volume 235, pp. 6643–6665. Available online: https://proceedings.mlr.press/v235/chen24k.html (accessed on 12 November 2026).
35. Lee, M.; Kim, D. Robust Evaluation of Diffusion-Based Adversarial Purification. In Proceedings of ICCV, Paris, France, 2–6 October 2023; pp. 134–144. Available online: https://openaccess.thecvf.com/content/ICCV2023/html/Lee_Robust_Evaluation_of_Diffusion-Based_Adversarial_Purification_ICCV_2023_paper.html (accessed on 12 November 2026).
36. Kang, M.; Song, D.; Li, B. DiffAttack: Evasion Attacks Against Diffusion-Based Adversarial Purification. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/ea0b28cbbd0cbc45ec4ac38e92da9cb2-Abstract-Conference.html (accessed on 12 November 2026).
37. Wang, K.; Fu, X.; Han, Y.; Xiang, Y. DiffHammer: Rethinking the Robustness of Diffusion-Based Adversarial Purification. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 89535–89562. Available online: https://proceedings.neurips.cc/paper_files/paper/2024/hash/a2fac827e992d55dcfdd4263e98528f4-Abstract.html (accessed on 12 November 2026).
38. Kassis, A.; Hengartner, U.; Yu, Y. DiffBreak: Is Diffusion-Based Purification Robust? In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2025. Available online: https://neurips.cc/virtual/2025/poster/117374 (accessed on 12 November 2026).
39. Wu, B.; Wei, S.; Zhu, M.; Zheng, M.; Zhu, Z.; Zhang, M.; Chen, H.; Yuan, D.; Liu, L.; Liu, Q. Defenses in Adversarial Machine Learning: A Survey. *arXiv* **2023**, arXiv:2312.08890. https://doi.org/10.48550/arXiv.2312.08890.
40. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble Adversarial Training: Attacks and Defenses. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://iclr.cc/virtual/2018/poster/157 (accessed on 12 November 2026).
41. Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.P.; El Ghaoui, L.; Jordan, M.I. Theoretically Principled Trade-off between Robustness and Accuracy. In Proceedings of the ICML, PMLR, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 7472–7482. https://doi.org/10.48550/arXiv.1901.08573.
42. Athalye, A.; Carlini, N.; Wagner, D. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In Proceedings of the ICML, PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 274–283. Available online: https://proceedings.mlr.press/v80/athalye18a.html (accessed on 12 November 2026).
43. Uesato, J.; O'Donoghue, B.; Kohli, P.; van den Oord, A. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. In Proceedings of the ICML, PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 5025–5034. Available online: https://proceedings.mlr.press/v80/uesato18a.html (accessed on 12 November 2026).
44. Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; Kurakin, A. On Evaluating Adversarial Robustness. *arXiv* **2019**, arXiv:1902.06705. https://doi.org/10.48550/arXiv.1902.06705.
45. Tramèr, F.; Carlini, N.; Brendel, W.; Madry, A. On Adaptive Attacks to Adversarial Example Defenses. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33. Available online: https://proceedings.neurips.cc/paper/2020/hash/11f38f8ecd71867b42433548d1078e38-Abstract.html (accessed on 12 November 2026).
46. Croce, F.; Gowal, S.; Brunner, T.; Shelhamer, E.; Hein, M.; Cemgil, T. Evaluating the Adversarial Robustness of Adaptive Test-time Defenses. In Proceedings of the ICML, PMLR, Baltimore, MD, USA, 17–23 July 2022; Volume162, pp. 4421–4435. Available online: https://proceedings.mlr.press/v162/croce22a.html (accessed on 12 November 2026).

47. Croce, F.; Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of the ICML, PMLR, Virtually, 13–18 July 2020; Volume 119, pp. 2206–2216. Available online: https://proceedings.mlr.press/v119/croce20b.html (accessed on 12 November 2026).
48. Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; Chiang, M.; Mittal, P.; Hein, M. RobustBench: A standardized adversarial robustness benchmark. *arXiv* **2020**, arXiv:2010.09670. https://doi.org/10.48550/arXiv.2010.09670.
49. Maini, P.; Wong, E.; Kolter, Z. Adversarial Robustness Against the Union of Multiple Perturbation Models. In Proceedings of the ICML, PMLR, Virtual, 13–18 July 2020; Volume 119, pp. 6640–6650. Available online: https://proceedings.mlr.press/v119/maini20a.html (accessed on 12 November 2026).
50. Jiang, E.; Singh, G. RAMP: Boosting Adversarial Robustness Against Multiple l_p Perturbations for Universal Robustness. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 43759–43787. https://doi.org/10.52202/079017-1388.
51. Dai, S.; Mahloujifar, S.; Xiang, C.; Sehwag, V.; Chen, P.-Y.; Mittal, P. MultiRobustBench: Benchmarking Robustness Against Multiple Attacks. In Proceedings of the ICML, PMLR, Honolulu, HI, USA, 23–29 July 2023; Volume 202, pp. 6760–6785. Available online: https://proceedings.mlr.press/v202/dai23c.html (accessed on 12 November 2026).
52. Wald, A. Statistical Decision Functions Which Minimize the Maximum Risk. *Ann. Math.* **1945**, *46*, 265–280. https://doi.org/10.2307/1969022.
53. Savage, L.J. The Theory of Statistical Decision. *J. Am. Stat. Assoc.* **1951**, *46*, 55–67. https://doi.org/10.1080/01621459.1951.10500768.
54. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. https://doi.org/10.1109/CVPR.2017.243.
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
56. Recht, B.; Roelofs, R.; Schmidt, L.; Shankar, V. Do ImageNet Classifiers Generalize to ImageNet? In Proceedings of the ICML, PMLR, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 5389–5400. Available online: https://proceedings.mlr.press/v97/recht19a.html (accessed on 12 November 2026).
57. Nicolae, M.-I.; Sinn, M.; Tran, M.N.; Buesser, B.; Rawat, A.; Wistuba, M.; Zantedeschi, V.; Baracaldo, N.; Chen, B.; Ludwig, H.; et al. Adversarial Robustness Toolbox v1.0.0. *arXiv* **2018**, arXiv:1807.01069. https://doi.org/10.48550/arXiv.1807.01069.
58. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the CVPR, Miami, FL, USA, 20–25 June 2009; pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.
59. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.
60. Brown, T. Imagenet_stubs: A Small Collection of ImageNet Images. Available online: https://github.com/nottombrown/imagenet_stubs (accessed on 24 October 2026).
61. Adversarial Robustness Toolbox Contributors. Evasion Attacks: Projected Gradient Descent and Basic Iterative Method. Available online: https://adversarial-robustness-toolbox.readthedocs.io/en/latest/modules/attacks/evasion.html (accessed on 24 October 2026).
62. Chernoff, H. Rational Selection of Decision Functions. *Econometrica* **1954**, *22*, 422–443.
63. von Neumann, J. Zur Theorie der Gesellschaftsspiele. *Math. Ann.* **1928**, *100*, 295–320.
64. Pearce, D.G. Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica* **1984**, *52*, 1029–1050.
65. Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing Robust Adversarial Examples. In Proceedings of the ICML, PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 284–293. Available online: https://proceedings.mlr.press/v80/athalye18b.html (accessed on 12 November 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.