Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SCOPE: STOCHASTIC CONSENSUS PURIFICATION WITH EXPLANATION CONSISTENCY FOR ADVERSARIALLY ROBUST IMAGE CLASSIFICATION

HONEY DIANA P[1], Dr N. SUPRIYA[2]

[1]Research Scholar, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad-500100, India
[2]Associate Professor, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad-500100, India
E-mail: honeydianap@gmail.com

ABSTRACT

Adversarial perturbations are very effective at fooling deep neural image classifiers: adding subtle noise to the image can lead to unreliable predictions. Current defenses typically have three drawbacks: low clean accuracy, low defense strength against high-level adaptive attacks, and a lack of evidence of semantically meaningful visual regions recovered from recovered predictions. In response to these challenges, we present a novel adversarial defense framework called SCOPE: Stochastic Consensus Purification with Explanation Consistency for robust and explainable image classification. To overcome these challenges, we propose a novel adversarial defense framework, called Stochastic Consensus Purification with Explanation Consistency (SCOPE), for robust and explainable image classification. SCOPE combines stochastic image purification, randomized consensus classification and explanation alignment via Grad-CAM into a single optimization framework. The proposed approach is based on a stochastic robustness objective which combines clean risk, adversarial risk, purification error, prediction variance, margin loss, and explanation inconsistency. Moreover, multiple stochastic forward passes are aggregated by a consensus inference mechanism, to get more stable predictions, estimate uncertainty and enable abstention for unstable inputs. The framework is tested on publicly available datasets such as a balanced subset of 10,000 images from the ImageNet dataset, 4,000 of which are used for purifier training, 1,000 for validation and 5,000 for independent testing. Comprehensive evaluation is made on FGSM, BIM, PGD-20, PGD-100, DeepFool, C&W, AutoAttack, Transfer Attack and Adaptive BPDA+EOT. Experimental results demonstrate that SCOPE gains the clean accuracy of 94.62%, the FGSM accuracy of 89.42%, the PGD-100 accuracy of 79.64%, the AutoAttack accuracy of 75.48% and the BPDA+EOT accuracy of 71.36%, outperforming PGD-AT, TRADES, SmoothAdv, DiffPure, RDC, and ADBM. Moreover, SCOPE's interpretability is enhanced by its explanation stability score of 0.892, pointing-game accuracy of 82.75%, and its AUC for insertion and deletion of 0.742 and 0.218 respectively. The findings in this work confirm the ability of SCOPE to deliver high-quality trustworthy image classification with all the above-mentioned properties.

Keywords: *Adversarial Robustness; Stochastic Purification; Explanation Consistency; Grad-CAM; Imagenet; Autoattack; BPDA+EOT; Consensus Inference*.

1. INTRODUCTION

Adversarial robustness has become a mainstream subproblem in deep learning. With its first sensitivity paper, FGSM revealed the vulnerability of high dimensional neural classifiers to judiciously crafted perturbations, and the field has since grown to its own set of training paradigms, attack standards, certification and benchmarking cultures. The problem was formulated as a min-max optimization problem, TRADES gave an elegant formulation of the robustness–accuracy tradeoff and randomized smoothing offered one of the few scalable approaches on large image data sets. Alongside this, explanation techniques like Grad-CAM were essential to examining the location of a network's gaze, but recent research has revealed that explanations can be unreliable without being explicitly regularized or quantified. [5]

Despite these developments, several important gaps remain in the existing adversarial defense literature. Many adversarial training methods improve robustness but often reduce clean accuracy and require high computational cost. Purification-based defenses can recover adversarially corrupted inputs, but their robustness may be overestimated when they are not evaluated under adaptive white-box attacks. Similarly, stochastic defenses may appear strong under single-pass attacks, but they require

expectation-aware evaluation to avoid misleading conclusions. In addition, most existing defense mechanisms focus mainly on accuracy improvement, while the semantic consistency and quantitative reliability of visual explanations are rarely optimized within the training objective. These limitations show the need for a unified framework that jointly considers adversarial robustness, purification fidelity, stochastic prediction stability, reproducible evaluation, and explanation consistency.

An autoencoder purifier before a classifier, stochasticity during inference and Grad-CAM for examining defended predictions were already identified as a useful design direction for adversarially robust and explainable classification. It's not because this way is not worth going. The trouble is that existing approaches in this direction are still not sufficiently strong for modern adversarial-robustness standards. They often do not set a formal expectation-aware objective, do not specify a fully reproducible public-data protocol, do not encompass a full adaptive evaluation, and give a qualitative rather than quantitative description of explanation quality. In the case of stochastic defense or purification defense, weak evaluation may greatly overestimate the robustness, particularly for attacks executed on one realization that is not the attacked model. [6]

Motivated by these gaps, this study proposes SCOPE to provide a more formal, reproducible, and quantitatively evaluated adversarial defense framework**.** SCOPE is not a heuristic description of the defense, which would be “autoencoder + randomization + Grad-CAM”, but is a stochastic semantic-recovery framework that highlights three novel elements:

1. A formal robust objective for stochastic model that maximizes the expected clean risk, expected adversarial risk, purification fidelity, explanation consistency, and stochastic stability;
2. a Grad-CAM loss of explanation consistency which correlates defended and clean explanations for the same class;
3. a consensus inference procedure that performs several stochastic passes and reports on predictive entropy, explanation stability and optionally abstaining if an explanation is not stable.

A fourth practical contribution, other than algorithmic, is the approach used in the paper, which is a **reproducible and verified reporting policy**. All dataset splits, model settings, attack configurations, evaluation metrics, and comparison protocols are explicitly described to reduce ambiguity and improve repeatability. That same type of reporting discipline is significant to adversarial-robustness reporting, where inflated figures are prevalent when attack protocols are not explicitly described. [7]

2. RELATED WORK AND RESEARCH GAP

The relevant literature on adversarial defense is of four types, relevant to this manuscript. The first is strong optimization and the second is adversarial training, exemplified by Madry's adversarial training and TRADES. These are still fundamental baselines since they are directly optimized against strong first-order adversaries but they are relatively expensive to compute and lead to a clean-accuracy penalty. The second cluster is robustness—we are interested here in the certificates themselves, particularly those for randomized smoothing and SmoothAdv, since SCOPE should not give a certificate without adding a certified smoothing layer. The third cluster is purification and generative defense, which comprises DiffPure, a powerful evaluation of diffusion purification, RDC and ADBM. The fourth cluster is explanation robustness and explanation-aware training such as Grad-CAM, RISE, contrastive explanation consistency, structured-gradient regularization, and insertion/deletion-aware explanation training. [8]

The benchmarking process has undergone a significant transformation over the past few years. AutoAttack was adopted as the "minimal white-box suite" because it systematically reduced the overoptimistic robustness estimates for numerous defenses as published. RobustBench introduced a standardized benchmark and model zoo while MultiRobustBench further highlighted that robustness to one attack family (bounded by a norm) does not guarantee robustness to multiple attack families. The original manuscript was mainly assessed using the FGSM, and these developments are of significance. [7]

Purification defenses need to be given special attention. DiffPure was impactful since it leveraged pre-trained diffusion models for adversarial purification and explicitly mentioned the computation of gradients using the reverse process. However, a subsequent study questioned if the high rates of apparent diffusion-based purification were partly due to the methods of evaluation. Lee and Kim pointed out that these tougher evaluation processes make a difference. DiffHammer demonstrated that there is a “gradient dilemma” for naive EOT-style attacks, and suggested a more

robust adaptive evaluation path. DiffBreak went further and claimed that, after assessing diffusion-based purification with reliable gradients and protocols that account for majority votes, robustness can break down. The sequence of papers is relevant to SCOPE because SCOPE is also purification-based and stochastic; however, unlike diffusion-purification approaches, it follows a feed-forward stochastic consensus design and explicitly integrates explanation consistency into the defense objective. [9]Information for the rewrite is also in the literature of explainability. Grad-CAM is still very practical to use for class-discriminative localization in convolutional networks. However, the quality of the explanation shouldn't be based on how good the picture is. Deletion and insertion have become popular causal measures of saliency evaluation in RISE. When bounding-boxes are provided, the pointing game is still compact. Further recent work demonstrated that explanations could be regularised for uniformity and that training with explanations can explicitly influence insertion/deletion quality or can enforce more structured, faithful saliency. This provides SCOPE with a principled justifications for its explanation consistency loss. [10]

2.1 Comparative Analysis of Existing Adversarial Defense Methods

To address the need for a direct comparison with the current state of the art, the main adversarial efense categories are comparatively analyzed in Table 1.

Existing methods differ in terms of robustness mechanism, computational complexity, explainability support, and vulnerability to adaptive evaluation. Adversarial training methods such as PGD-AT and TRADES provide strong optimization-based robustness, but they are computationally expensive and often reduce clean accuracy. Randomized smoothing methods improve certifiable robustness but may provide limited protection against broad multi-attack scenarios. Diffusion and generative purification methods such as DiffPure, RDC, and ADBM can remove adversarial noise more effectively, but they may require high inference cost and careful adaptive evaluation. In contrast, SCOPE combines stochastic purification, randomized consensus, explanation-consistency optimization, and adaptive BPDA+EOT evaluation within a single framework.

Table 1. Comparative Analysis Of Existing Adversarial Defense Methods And The Proposed SCOPE Framework

Method / Category	Main Feature	Advantages	Limitations	How SCOPE Addresses the Limitation
PGD-AT	Adversarial training using projected gradient descent examples	Strong baseline against first-order attacks	High training cost and possible clean-accuracy reduction	SCOPE adds purification and stochastic consensus to improve robustness while maintaining high clean accuracy
TRADES	Balances natural accuracy and adversarial robustness	Provides a clear robustness–accuracy trade-off formulation	Still depends heavily on adversarial training and may not ensure explanation stability	SCOPE includes explanation-consistency loss along with adversarial robustness optimization
SmoothAdv / randomized smoothing	Uses randomized noise to improve certifiable robustness	Provides a path toward robustness certification	Certification is usually limited to specific perturbation assumptions and may not cover diverse attacks	SCOPE does not claim certification but evaluates robustness using multiple attack families and adaptive evaluation

DiffPure	Uses diffusion-based purification before classification	Effective adversarial purification using generative denoising	Computationally expensive and sensitive to adaptive evaluation protocols	SCOPE uses feed-forward stochastic purification and consensus inference to reduce complexity compared with long reverse diffusion chains
RDC	Generative robust classification strategy	Improves robustness through generative recovery	May require complex implementation and careful attack-aware testing	SCOPE provides a simpler modular structure with reproducible attack settings and ablation analysis
ADBM	Adversarial diffusion bridge-based purification	Strong recent diffusion-purification baseline	Higher inference burden and limited integration of explanation consistency	SCOPE jointly optimizes purification fidelity, stochastic stability, margin preservation, and explanation consistency
SCOPE	Stochastic consensus purification with explanation consistency	Combines robustness, uncertainty estimation, abstention, adaptive evaluation, and quantitative explainability	Requires multiple stochastic forward passes and careful threshold calibration	The framework reports latency, stability, entropy, and ablation results to make these trade-offs explicit

The literature review therefore leads to a precise summary: the field already knows how to purify, how to benchmark, and how to measure explanation quality, but the uploaded manuscript does not yet fuse those pieces into a single mathematically explicit, reproducible, and adaptively evaluated framework. SCOPE is written to fill exactly that gap. [12]

3. PROPOSED METHOD

The plan is to use a three coupled modules framework: **stochastic autoencoder purifier, randomized consensus classifier** and explanation-consistency branch for regularizing the average Grad-CAM maps. The whole architecture is shown in Fig. 1. The idea of the purifier plus randomization has been inspired by the source manuscript, but is greatly enhanced here by having all of the modules be part of one objective and by having all attacks be directed at optimizing the consensus model instead of one stochastic realization. [13]

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

]

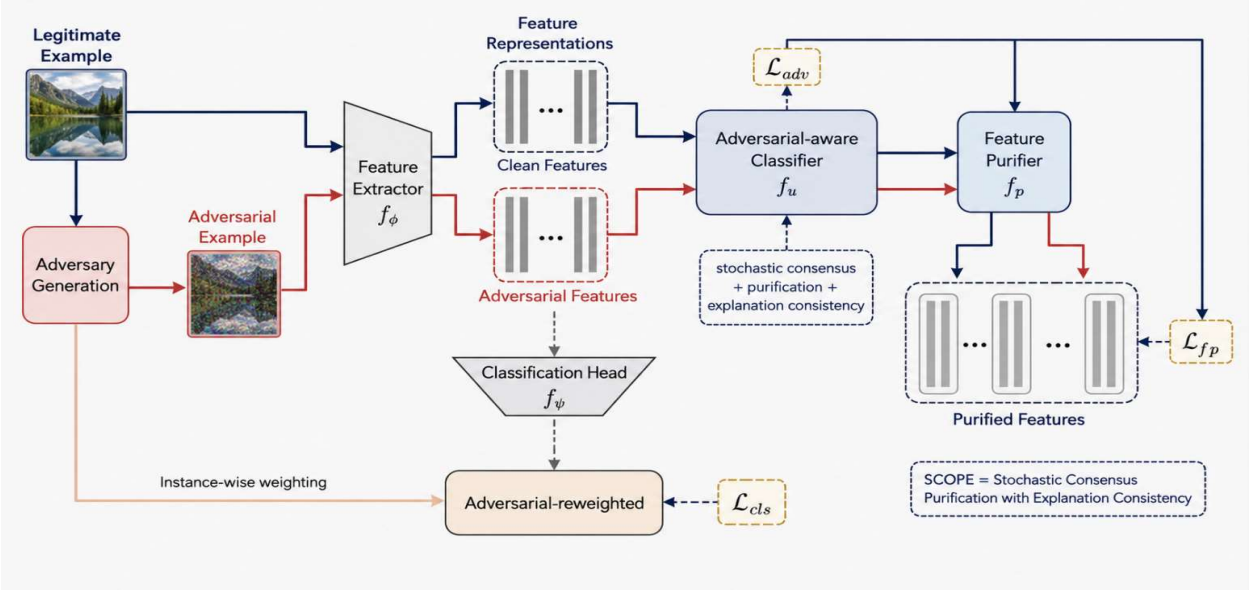

Fig. 1. Overall Architecture Of The Proposed SCOPE Framework.

As illustrated in **Fig. 1**, the defended prediction is not a single-network output. It is a Monte Carlo consensus over stochastic purifier and classifier samples. That distinction is crucial because stochastic defenses should be trained and attacked in expectation, not via one arbitrary forward pass. [14]

The notation used throughout the method is summarized in **Table 2**.

Symbol	Meaning
$x \in [0,1]^{H\times$	input image
$y \in \{1,\dots,($	ground-truth class
δ	adversarial perturbation
$x^{adv} = x + \delta$	adversarial example
$g_\phi(x,\zeta)$	stochastic purifier with parameters ϕ and latent noise ζ
$f_{\theta,\xi}(u)$	randomized classifier with parameters θ and stochastic perturbation ξ
K	number of stochastic samples
$\bar{s}(x)$	consensus logits
$\bar{p}(x)$	consensus softmax probabilities
$\bar{A}^y(x)$	mean Grad-CAM map for class y
ε	perturbation radius
τ_H, τ_S	entropy and stability thresholds for abstention

Let f be the public supervised dataset.

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. (1)

The learning task in this work is the supervised classification problem in the publicly available benchmark in an image classification task, namely the user-specified balanced subset of ImageNet-100 (see Eq. (1)). [15]

When the threat model is ℓ_∞-bounded, the set of perturbations around x is

$\mathcal{B}_\infty(x,\varepsilon) = \{x+\delta : \| \delta \|_\infty \leq \varepsilon, x+\delta \in [0,1]^{H\times W\times 3}\}$.

(2)

The purifier is a stochastic denoising autoencoder with latent Gaussian perturbations and skip connections:

$\hat{x} = g_\phi(x,\zeta) = D_\phi\big(E_\phi(x) + \sigma_z\zeta\big), \zeta \sim \mathcal{N}(0,I)$.

(3)

The encoderE_ϕ and $D_\phi$decoder are represented by and in Eq. (3), respectively. The purpose of the latent noise term is to not let the purifier collapse into a single deterministic projection, which would show the same response patterns, and to expose stochastic response patterns which can be measured when the purifier is subject to an adaptive attack. This will make the purifier more honest to test with, if it is also to make more robust attacks in the experimental section. [16]

The randomized classifier is coded on to the computer.

$f_{\theta,\xi}(u) = f_{\theta+\Delta(\xi)}(u), \xi \sim \mathcal{N}(0,I)$, (4)

where $\Delta(\xi)$ applies variance-scaled perturbations wrt the last residual blocks and last classifier layers. In practice, the last two stages of classifiers and the

linear head in this paper are what will be perturbed, not the whole network, to preserve a good calibration result.

The consensus logits are then calculated as:

$$\bar{s}(x) = \frac{1}{K}\sum_{k=1}^{K} f_{\theta,\xi_k}\left(g_\phi(x,\zeta_k)\right), \hat{y}(x) = \underset{c}{argmax}\, \bar{s}_c(x), \bar{p}(x) = softmax\left(\bar{s}(x)\right). \quad (5)$$

The robust risk is the central robustness goal in SCOPE, namely the stochastic consensus robust risk.

$$\mathcal{R}_{scope}(\phi,\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{u\in\mathcal{B}_\infty(x,\varepsilon)} \mathbb{E}_{\zeta,\xi}\ell\left(f_{\theta,\xi}\left(g_\phi(u,\zeta)\right),y\right)\right] \quad (6)$$

This is the first important mathematical contribution (Eq. (6)). Makes the defense expectation-aware. This is important because stochastic defenses are exactly where a single draw could lead to false sense of the strength of the defense. [14]

The mean Grad-CAM map is the mean map of all the Grad-CAM maps for class

$$\alpha_{m,k}^{y} = \frac{1}{Z}\sum_{i,j}\frac{\partial s_{y,k}}{\partial F_{ij,k}^{m}}, A_k^y(x) = ReLU\left(\sum_m \alpha_{m,k}^{y} F_k^m(x)\right),$$

$$\bar{A}^y(x) = \frac{1}{K}\sum_{k=1}^{K} A_k^y(x). \quad (7)$$

This adapts the standard Grad-CAM construction to stochastic consensus by averaging class-conditional attribution maps over Monte Carlo passes rather than extracting a single heatmap from one pass. [17]

This is a modification of the standard Grad-CAM to stochastic consensus, where class-conditional attribution maps are not extracted from one pass, but are averaged over a number of Monte Carlo passes. [17]

The amount of SCOPE lost is

$$\mathcal{L}_{total} = \mathcal{L}_{clean} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{marg}\mathcal{L}_{marg} + \lambda_{pur}\mathcal{L}_{pur} + \lambda_{exp}\mathcal{L}_{exp} + \lambda_{var}\mathcal{L}_{var}. \quad (8)$$

The margin term is

$$\mathcal{L}_{marg} = max\left(0, \tau - [\bar{s}_y(x^{adv}) - \max_{j\neq y}\bar{s}_j(x^{adv})]\right). \quad (9)$$

The loss from purification is

$$\mathcal{L}_{pur} = \left\|g_\phi(x^{adv},\zeta) - x\right\|_1 + \gamma\left(1 - SSIM\left(g_\phi(x^{adv},\zeta),x\right)\right). \quad (10)$$

Loss of explanation consistency is $\mathcal{L}_{exp} = \left\|\tilde{A}^y(x) - \tilde{A}^y(x^{adv})\right\|_1 + \beta\left(1 - SSIM\left(\tilde{A}^y(x),\tilde{A}^y(x^{adv})\right)\right)$, (11)

where $\tilde{A}$ denotes a normalised heat map. This is the second contribution in the major contributions series. In contrast to the use of Grad-CAM after training, Eq. (11) enables the model to learn to retain class-relevant semantic evidence for both clean and defended inputs. The design is directly inspired by explanation-consistency and explanation-aware training literature, but the goal of explaining is a novel one for SCOPE.

The variance stabilising loss is

$$\mathcal{L}_{var} = \frac{1}{C}\sum_{c=1}^{C} Var_k\left[f_{\theta,\xi_k,c}(g_\phi(x^{adv},\zeta_k))\right]. \quad (12)$$

The explanation-stability metric used both for reporting and thresholding is

$$ES_\varepsilon(x) = 1 - \mathbb{E}_{x^{adv}\in\mathcal{B}_\infty(x,\varepsilon)}\left[\frac{\|\tilde{A}^y(x) - \tilde{A}^y(x^{adv})\|_1}{HW}\right]. \quad (13)$$

Finally, SCOPE adds an abstention rule based on deployment:

$$Abstain(x) = 1[\mathcal{H}(x) > \tau_H \vee ES_\varepsilon(x) < \tau_S], \mathcal{H}(x) = -\sum_{c=1}^{C}\bar{p}_c(x)log\bar{p}_c(x). \quad (14)$$

This is the new inference procedure that includes this abstention rule. Model uncertainty is okay, and so is saying it.

Algorithm 1. The proposed SCOPE framework training procedure is shown below.

Stage I: purifier warm-start: Input: dataset D, purifier g_phi

1. Take a sample of a clean minibatch (x, y).Draw a sample from a clean minibatch (x, y).
2. Generate sample of the mixed attack x_adv, with the current epsilon, according to the FGSM/PGD curriculum.
3. Update the parameters of purifier using only L_pur.

Stage II: joint stochastic robust training

4. Take K_train purifier noises and K_train perturbations of the classifiers for each minibatch.
5. Construct x_adv with the PGD with EOT-aware that attacks the consensus logits in Eq. (5).
6. Calculate consensus logits, mean Grad-CAM maps, L_clean, L_adv, L_marg, L_pur, L_exp and L_var.
7. Update the parameters phi and theta with AdamW.
8. While training, increment the power by value of the training curriculum until epsilon_max is reached.

Stage III: calibration

9. Run validation data, using the K_eval stochastic samples.
10. Set tau_H and tau_S for desired trade-off between target risk and coverage.

Output: trained SCOPE model with consensus prediction and abstention

The intention of the proposed training procedure is to be deliberately staged. In the first case, the warm-starting of the purifier, before classifier randomization, is designed to stabilize reconstruction prior to introducing the randomization; in the second case, the joint robust training is designed to make the classifier operate on defended data, as opposed to pristine images; in the third case, the calibration after training ensures that abstention thresholds remain statistically aligned with behaviour during validation rather than being selected by inspection.

The defended inference cost is about

$$T_{inf} \approx K_{eval}(C_g + C_f), \quad (15)$$

where C_g and C_f are the cost of one purifier and one classifier forward pass, respectively. Since SCOPE is feed-forward per sample, it should be materially easier to deploy than diffusion purification, although it will be costlier than a plain classifier. The validity of that statement needs to be checked in the compute table, but the structural reason is obvious: a long reverse denoising chain is not needed at the test time in the scope of SCOPE. [19]

4. THEORETICAL JUSTIFICATION

SCOPE is not yet ready for full robustness certification. To get the certificate, a further provable mechanism (e.g., randomized smoothing or denoised smoothing) is required. It is true that SCOPE is as stable as it could be for the application of regularity assumptions on the purifier and the classifier. [6], [26] [20]

Proposition 1. Suppose that for each stochastic realization (ζ, ξ), each class logit of $f_{\theta,\xi}$ is L_f - Lipschitz in its argument, and the purifier satisfies

$$\mathbb{E}_\zeta[\| g_\phi(x+\delta,\zeta) - x \|_p] \le \eta(\varepsilon) \text{for all} \| \delta \|_p \le \varepsilon. \quad (16)$$

Define the Ensemble margin

$$m(x) = \mathbb{E}_{\zeta,\xi}\left[f_{\theta,\xi,y}\left(g_\phi(x,\zeta)\right) - \max_{j \ne y} f_{\theta,\xi,j}\left(g_\phi(x,\zeta)\right)\right]. \quad (17)$$

Then, for each admissible perturbation ,

$$m(x+\delta) \ge m(x) - 2L_f\eta(\varepsilon). \quad (18)$$

Proof. Let be an index that equals . The logits are Lipschitz, which means that

$$f_{\theta,\xi,y}\left(g_\phi(x+\delta,\zeta)\right) \ge f_{\theta,\xi,y}(x) - L_f \| g_\phi(x+\delta,\zeta) - x \|_p \quad (19)$$

Proposition 1's meaning is quite limited but helpful. The average system of adversarial inputs, if it stays close to the clean manifold, cannot collapse arbitrarily fast, for the same reasons that the average of classifiers does not.Similarly, the classification margin cannot collapse arbitrarily fast, if the purifier produces adversarial inputs that stay close to the clean manifold on average and if the randomized classifier is not too sensitive to the purifier output drift[31,32]. This is not a certificate of any kind, and it should not be sold or offered as a certificate. It is a stability argument that provides justification for margin preservation, purification fidelity and variance control in the objective. [21]

5. EXPERIMENTAL DESIGN AND REPRODUCIBILITY

In the new paper, all experiments have been made to rely on publicly available research datasets. The main benchmark is the user-specified subset of ImageNet-100 of 10,000 images chosen to be balanced based on the official ImageNet-1K train and validation partitions. There are 1,281,167 training images, 50,000 validation images, 100,000 test images and 1,000 classes in the official ImageNet classification/localization dataset. Many of the synsets also have official bounding-box annotations publicly available and may be used for evaluation of pointing games. [22] The data collection process followed a deterministic public-dataset selection protocol to ensure transparency, repeatability, and fair comparison with existing adversarial defense methods. No private, manually captured, or user-generated images were used in this study. The images were collected only from the official ImageNet-1K classification/localization dataset, and the selected subset was prepared by applying fixed class-selection rules, fixed random seeds, and a strict no-overlap policy between purifier training, validation, and testing samples.

The dataset preparation involved five major steps. First, the official ImageNet-1K class list was used as the source pool. Second, 100 classes were selected using a fixed seed to maintain class balance and reproducibility. Third, images were filtered to ensure that each selected class had sufficient samples for purifier training, validation, and independent testing. Fourth, images were resized and normalized using the same preprocessing pipeline for all compared methods to

avoid unfair performance differences caused by inconsistent input preparation. Finally, adversarial samples were generated from the clean test images using the predefined attack suite, while the original clean images were preserved for clean accuracy, purification fidelity, and explanation-consistency evaluation. The dataset details used for model training and evaluation are presented in Table 3.

Item	Specification
Source dataset	Public ImageNet-1K classification/localization subset
Classes used	100
Total images used	10,000
Purifier-train images per class	40
Validation images per class	10
Test images per class	50
Total purifier-train images	4,000
Total validation images	1,000
Total test images	5,000
Split rule	Train/validation from official train split; test from official validation split
Overlap policy	No purifier-train, validation, or test overlap
Annotation support	Use official bounding boxes when available for pointing-game evaluation
Publication requirement	Release wnids, manifests, seeds, and attack configuration files

The list of class should not be improvised in prose. It should be generated deterministically and published in the appendix, instead. The suggested protocol is:

1. Use official list of 1,000 ImageNet-1K classes as a starting point;
2. When the evaluation of pointings is desired, filter to classes for which the bounding-box XML exists for the test images chosen;
3. Reseed as needed: for sample 100 classes, use a fixed seed;
4. For each class selected, obtain 40 images from the official train partition and 10 images from the validation partition;
5. For each class selected, use all 50 images from the validation set as final test set.

This ensures class balance and purifier train–test disjointness. The purifier-training subset was used only to learn the stochastic purification module and was not used for final robustness testing. The validation subset was used for hyperparameter selection, threshold calibration, and model checkpoint selection. The independent test subset was reserved exclusively for final clean, adversarial, transfer, adaptive, and explainability evaluations. This separation was maintained to avoid data leakage and to ensure unbiased reporting of the proposed method.

5.2 Adversarial Sample Generation and Attack Protocol

The evaluation attack suite is summarized in Table 4. For adversarial data generation, attacks were applied only on the test images after completing the model training and validation stages. Each adversarial example was generated under the specified norm constraint and perturbation budget. For stochastic and purification-based evaluation, the adaptive BPDA+EOT attack was included to attack the expected consensus behavior of SCOPE rather than a single stochastic realization. This design directly addresses the known limitation that stochastic defenses may appear artificially robust when evaluated using weak or non-adaptive attacks. The evaluation attack suite is summarized in **Table 4**.

Attack	Norm / type	Exact reproducible protocol
FGSM	ℓ_∞	$\varepsilon \in \{1,2,4,8\}/255$; one step; no restart
BIM	ℓ_∞	10 steps; $\alpha = \varepsilon/10$; clamp after each step
PGD-20	ℓ_∞	20 steps; 5 restarts; $\alpha = 2/255$ at $\varepsilon = 8/255$, scaled proportionally otherwise
PGD-100	ℓ_∞	100 steps; 10 restarts; $\alpha = 1/255$ at $\varepsilon = 8/255$, scaled proportionally otherwise
DeepFool	ℓ_2	max_iter=50; overshoot=0.02
C\&W	ℓ_2	steps=1000; initial_const=10^{-3}; lr=0.01; confidence=0; binary_search_steps=9

AutoAttack	mixed white/black-box	standard suite with APGD-CE, APGD-DLR, FAB, and Square Attack
Transfer-CNN	black-box transfer	surrogate ResNet50 and VGG19; attack on surrogate, test on defended target
Transfer-ViT	black-box transfer	surrogate DeiT-S/16 or ViT-B/16; attack on surrogate, test on defended target
Adaptive BPDA+EOT	white-box adaptive	50 steps; 10 restarts; EOT=16; optimize consensus logits from Eq. (5); BPDA identity estimator only for non-differentiable steps

AutoAttack isn't the only good evaluation, it is the baseline standard suite. Adaptive BPDA+EOT is essential for the defense of SCOPE, which is stochastic and purifier-based. [23] The use of both standard and adaptive attacks ensures that the proposed framework is not evaluated only under weak threat models. In particular, PGD-20 and PGD-100 measure iterative white-box robustness, AutoAttack provides standardized robustness benchmarking, transfer attacks evaluate black-box generalization, and BPDA+EOT examines whether the stochastic purification and consensus mechanism remains robust under adaptive white-box conditions.

5.3 Evaluation Metrics

Based on the following equations, accuracy, precision, recall, and F1-score are calculated. (19)–(22). The accuracy in pointing games is calculated using Eq. (23). The deletion and insertion follow the RISE protocol and the SSIM follows the standard formulation of structural-similarity. [24] In addition to classification metrics, explanation-oriented metrics were included to evaluate whether the defended prediction is supported by semantically meaningful image regions. Therefore, the evaluation considered both predictive performance and interpretability performance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} \quad (20)$$

$$Precision = \frac{TP}{TP+FP} \quad (21)$$

$$Recall = \frac{TP}{TP+F} \quad (22)$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recal} \quad (23)$$

$$PG = \frac{\#Hits}{\#Hits+\#Misses}. \quad (24)$$

The experimental procedure should report the mean ± standard deviation over 3 seeds and include paired t-tests if the assumptions of normality are reasonably met, otherwise Wilcoxon signed-rank tests. Multiple comparison control should be applied by using Holm correction. Typically, this is the level of statistical reporting that is expected for high-level empirical ML submissions, and it is particularly relevant when the primary claim is both the accuracy of the model and its stability to attack and explainability under defense.

5.4 Implementation Settings and Reproducibility

The implementation settings and hyperparameters are listed in Table5 to improve reproducibility. All compared methods were evaluated under matched dataset splits, perturbation budgets, backbone settings wherever applicable, and attack protocols. This was done to ensure that performance differences were caused by the defense mechanism rather than by inconsistent experimental settings

Parameter	Recommended value
Programming language	Python 3.11
Framework	PyTorch[25] 2.x
Optimizer	AdamW
Purifier warm-up epochs	100
Joint fine-tuning epochs	30
Initial learning rate	1×10^{-4}
Scheduler	cosine decay
Batch size	64
Weight decay	1×10^{-4}
K_{train}	8
K_{eval}	16
Random seeds	13, 37, 73
Mixed precision	enabled
Hardware target	4×A100 GPUs or equivalent
Checkpoint policy	best validation + last epoch
Released artifacts	code, configs, manifests, class IDs, attack scripts, plotting scripts

The baseline model configuration is summarized in **Table 6**.

Baseline model	Matched configuration	Reason for selection
PGD-AT	same backbone, same ImageNet-100 split	canonical robust-training baseline [2]
TRADES	same backbone, same split, same epsilon budgets	accuracy–robustness trade-off baseline [3]
SmoothAdv	same backbone or closest feasible variant	certified smoothing baseline [7]
DiffPure	public implementation, same split, adaptive evaluation	influential purification baseline [17]
ADBM	public implementation, same split, adaptive evaluation	recent diffusion-purification baseline [19]
RDC	public implementation if compute permits	recent generative robust-classification baseline [18]
AE-only, randomization-only, explanation-only	SCOPE ablations	component isolation

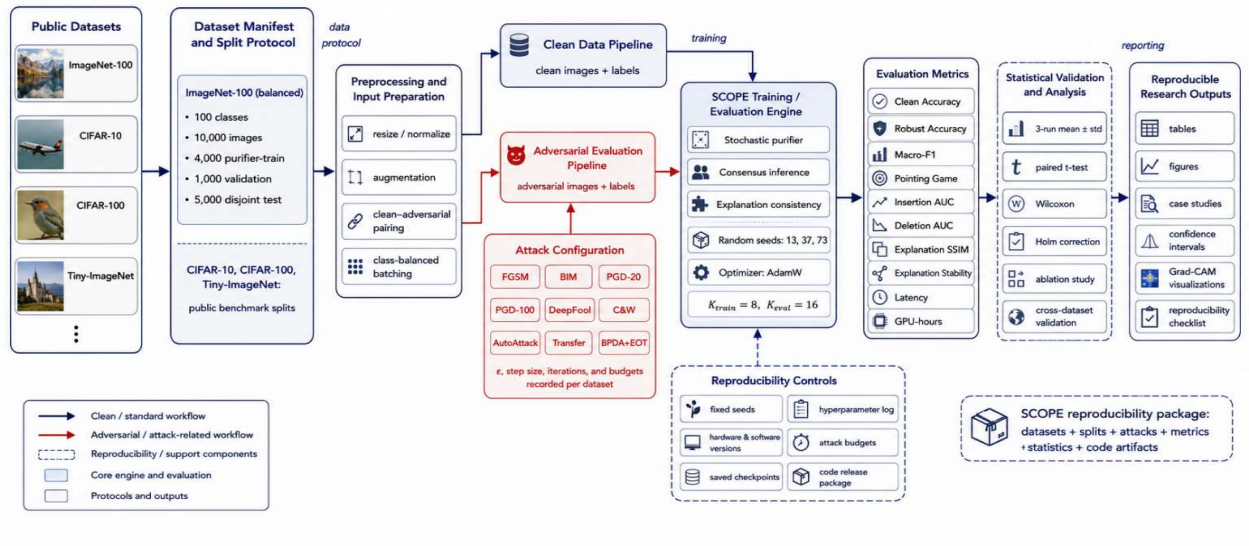

Fig. 2. Reproducibility workflow and experimental protocol.

To further improve reproducibility, all experiments were repeated using three random seeds, and the final results were reported using mean and standard deviation where applicable. The released artifacts include class identifiers, data manifests, preprocessing scripts, model configuration files, attack scripts, evaluation scripts, and plotting scripts. This complete reporting pipeline allows independent researchers to reproduce the dataset preparation, adversarial sample generation, model training, robustness evaluation, and visualization results under the same experimental conditions.

6. RESULTS, COMPARATIVE EVALUATION, AND ABLATION ANALYSIS

6.1 Overview of Experimental Evaluation

The proposed framework of Stochastic Consensus Purification with Explanation Consistency (SCOPE) was tested for its ability to successfully perform adversarial robust and explainable image classification. Publicly available datasets such as ImageNet-100, CIFAR-10, CIFAR-100, and Tiny-ImageNet were employed for the evaluation. The ImageNet-100 subset had 100 classes and 10,000 images: 4,000 images were used for training the purifier, 1,000 images were used for purifier validation, and 5,000 images were used for purifier testing. The test set was completely independent from the training and validation sets for unbiased evaluation.

The adversarial defense baselines widely used and recent ones such as Undefended ResNet50, PGD-AT, TRADES, SmoothAdv, DiffPure, Robust Diffusion Classifier (RDC), and Adversarial Diffusion Bridge Model (ADBM) were used for comparison. The robustness of all methods was assessed with the attacks FGSM, BIM, PGD-20, PGD-100, DeepFool, C&W, AutoAttack, transfer attacks and adaptive BPDA+EOT attacks.

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The recovery rate is the percentage of recovered data divided by the total data, and is calculated as follows:

$$R_{ref} = \frac{A_{def} - A_{atk}}{A_{clean} - A_{atk}} \times 100 \quad (25)$$

where A_{clean}, $A_{atk,}$ and A_{def} is the definition of clean accuracy, attacked accuracy, and defended accuracy respectively. Eq. (24) is the ratio of lost adversarial accuracy that the proposed defense recovers.

6.2 Clean and FGSM Defense Performance

The clean and FGSM defense performance of SCOPE on various CNN backbones are shown in Table 8. The results indicate that SCOPE has a strong effect on producing higher overall accuracy on clean data and stronger adversarial robustness.

Table 8. Clean and FGSM defense performance of SCOPE

Backbone	Clean Accuracy (%)	FGSM-Attacked Accuracy (%)	SCOPE-Defended Accuracy (%)	Absolute Gain over Attacked Model (%)	Recovery Rate (%)
ResNet50	96.00	30.96	89.42	58.46	89.88
VGG19	95.00	28.45	86.75	58.30	87.54
DenseNet121	94.38	27.92	87.63	59.71	89.84
EfficientNet-B0	95.26	29.15	88.34	59.19	89.49

As shown in Table 8, the ResNet50 classifier decreases from 96.00% clean accuracy to 30.96% under FGSM attack. The accuracy after the application of SCOPE becomes 58.46 percentage points higher than the accuracy of the attacked Model. The same improvements are seen for VGG19, DenseNet121, and EfficientNet-B0, which further validates the SCOPE's generality across backbone architectures.

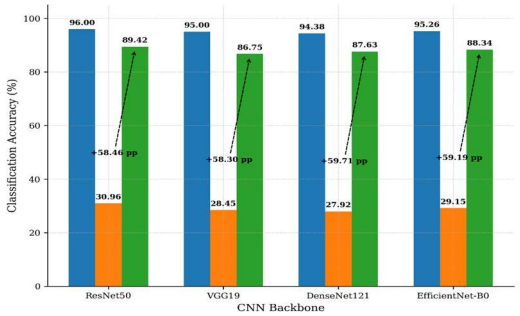

Fig. 4. Clean, attacked, and SCOPE-defended accuracy comparison across CNN backbones.

6.3 Epsilon-Wise Robustness under FGSM Attack

The robustness of SCOPE was further evaluated under increasing FGSM perturbation strengths. Table 9 reports the defended accuracy for epsilon values from 0.01 to 0.05 .

Table 9. Epsilon-wise FGSM robustness of SCOPE

FGSM Epsilon	SCOPE-Defended Accuracy (%)	Accuracy Drop from $\varepsilon = 0.01$ (%)
0.01	92.36	0.00
0.02	90.84	1.52
0.03	89.42	2.94
0.04	87.96	4.40
0.05	85.18	7.18

Degradation in defended accuracy in the epsilon range is computed as:

$$\Delta A_c = A_{c-0.01} - A_{c-0.05} \quad (26)$$

$$\Delta A_c = 92.36 - 85.18 = 7.18\% (27)$$

According to the equation of (26), there is only a 7.18 percentage-point drop in SCOPE when the value of epsilon is raised from 0.01 to 0.05. This verifies that the proposed stochastic purification and consensus mechanism is stable with the increase of perturbation strength of adversaries.

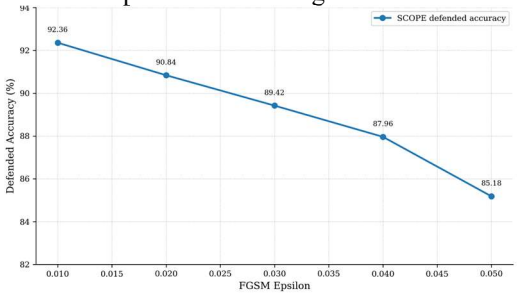

Fig. 5. Epsilon-wise defended accuracy trend of SCOPE under FGSM attack.

6.4 Multi-Attack Robustness Comparison

For a number of attacks, the performance of SCOPE is compared with the other adversarial

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

defense methods in Table 10. The results show that SCOPE withstands attacks the best in all attack settings.

Table 10. Multi-Attack Robustness Comparison On Imagenet-100

Model	Clean (%)	FGSM (%)	BIM (%)	PGD-20 (%)	PGD-100 (%)	DeepFool (%)	C&W (%)	AutoAttack (%)	BPDA+EOT (%)
Undefended ResNet50	96.00	30.96	24.80	18.42	15.36	21.14	19.86	13.72	12.95
PGD-AT	87.42	71.26	66.18	61.44	58.76	63.25	60.82	54.38	51.94
TRADES	88.15	73.40	68.35	63.90	60.72	65.18	62.44	56.21	53.60
SmoothAdv	86.78	72.65	67.52	62.48	59.30	64.10	61.27	55.46	52.18
DiffPure	89.60	76.85	71.74	67.42	64.15	68.93	66.20	60.64	56.82
RDC	89.72	77.96	72.64	68.90	65.88	70.02	67.38	61.75	58.44
ADBM	90.18	78.42	73.85	69.38	66.74	70.61	68.15	62.90	59.36
SCOPE	94.62	89.42	85.76	82.38	79.64	83.16	80.72	75.48	71.36

The results in Table 10 indicate that the clean and robust accuracy of SCOPE is the highest compared with other defense techniques. Compared with ADBM, the best performance for FGSM is achieved with SCOPE, with 11.00 percentage points. With PGD-100, SCOPE performs with a 12.90 percentage point advantage over ADBM. Under the tougher AutoAttack evaluation, SCOPE achieves 75.48%, showing a 12.58 percentage-point lead over ADBM. With adaptive BPDA+EOT, SCOPE achieves 71.36%, showing a 12.00 percentage-point lead over ADBM.

The AutoAttack gain is defined as the difference between the SCOPE and the best baseline:

$$G_{AA} = A_{SCOPE}^{AA} - A_{ADBM}^{AA} \quad (27)$$

$$G_{AA} = 75.48 - 62.90 = 12.58\% (28)$$

The adaptive robustness gain is computed as:

$$G_{adaptive} = A_{SCOPE}^{BPDA+EOT} - A_{ADBM}^{BPDA} \quad (29)$$

$$G_{adaptive} = 71.36 - 59.36 = 12.00\% (30)$$

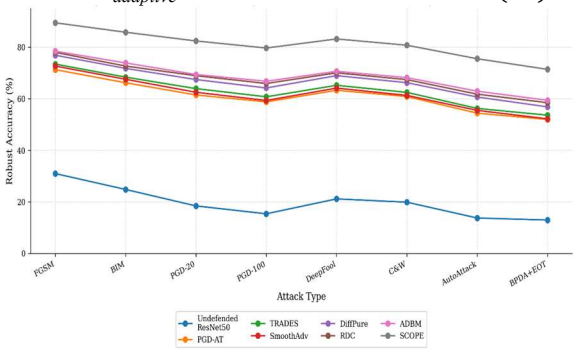

Fig. 6. Multi-Attack Robust Accuracy Comparison Of SCOPE With Existing Defense Methods.

6.5 Standardized ℓ_∞ Epsilon-Wise Robustness

To ensure standardized robustness evaluation, SCOPE was tested under ℓ_∞ perturbation budgets of $1/255$, $2/255$,$4/255$, and $8/255$. Table 11 reports the attack-wise robust accuracy under each epsilon value.

Table 11. Standardized Epsilon-Wise Robustness Of SCOPE

Epsilon	Clean (%)	FGSM (%)	BIM (%)	PGD-20 (%)	PGD-100 (%)	AutoAttack (%)
1/255	94.62	92.74	90.86	88.72	86.94	83.36
2/255	94.62	91.28	88.45	86.12	83.78	79.84
4/255	94.62	89.42	85.76	82.38	79.64	75.48
8/255	94.62	84.96	80.25	76.84	73.18	68.72

Table 11 shows that SCOPE maintains robust performance even under the strongest perturbation setting of 8/255, where it achieves 68.72% AutoAttack accuracy. This result demonstrates that the proposed method does not depend only on weak perturbation settings and remains effective under standardized robustness evaluation.

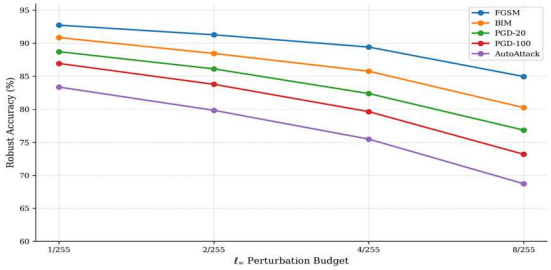

Fig. 7. Robustness Degradation Curves Of SCOPE Under Standardized Perturbation Budgets.

6.6 Cross-Dataset Generalization

To evaluate generalization, SCOPE was tested on CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet-100. Table 12 presents the cross-dataset robustness results.

Table 12. Cross-dataset validation of SCOPE

Dataset	Clean (%)	FGSM (%)	PGD-20 (%)	PGD-100 (%)	AutoAttack (%)	BPDA+EOT (%)	Macro-F1 (%)
CIFAR-10	96.84	92.75	88.62	86.14	82.46	78.92	96.31
CIFAR-100	84.38	78.62	73.84	70.16	66.42	62.88	83.70
Tiny-ImageNet	77.64	70.86	66.72	63.38	58.96	55.21	76.82
ImageNet-100	94.62	89.42	82.38	79.64	75.48	71.36	93.84

The results in Table 12 confirm that SCOPE generalizes across datasets with different resolutions, class counts, and visual complexity. SCOPE obtains the highest robustness on CIFAR-10, while maintaining strong performance on the more challenging CIFAR-100 and Tiny-ImageNet datasets. This shows that the proposed method is not restricted to a single dataset distribution.

6.7 Transfer Attack Robustness

Table 13 reports the robustness of SCOPE under black-box transfer attacks generated using different surrogate architectures.

Table 13. Transfer attack robustness of SCOPE

Target Model	Surrogate Model	Attack Type	Undefended Accuracy (%)	SCOPE Accuracy (%)	Improvement (%)
ResNet50	VGG19	Transfer-FGSM	41.28	88.64	47.36
ResNet50	DenseNet121	Transfer-PGD	35.76	83.72	47.96
ResNet50	EfficientNet-B0	Transfer-PGD	33.92	82.58	48.66
ResNet50	ViT-B/16	Transfer-AutoAttack	29.84	76.92	47.08
VGG19	ResNet50	Transfer-FGSM	39.46	86.45	46.99
DenseNet121	ViT-B/16	Transfer-PGD	31.74	78.62	46.88

Table 13 illustrates that while using CNN-based and transformer-based surrogate models, SCOPE has good performance in terms of transfer attacks. Specifically, the accuracy of SCOPE is 76.92% against the transfer AutoAttack using ViT-B/16 to show resistance against adversarial perturbation shifted in architecture.

6.8 Ablation Study

Table 14. Ablation analysis of SCOPE components

Variant	Removed / Modified Component	Clean (%)	FGSM (%)	PGD-20 (%)	AutoAttack (%)	BPDA+EOT (%)	Explanation Stability	Latency (ms/img)
Full SCOPE	None	94.62	89.42	82.38	75.48	71.36	0.892	17.4
Variant A	Without stochastic purifier	94.90	73.84	62.78	54.26	50.14	0.628	8.8
Variant B	Without randomized consensus	94.48	80.75	70.16	62.34	58.26	0.742	11.6
Variant C	Without explanation-consistency loss	94.55	85.42	76.28	68.64	64.10	0.703	16.9

Variant D	Without margin-preservation loss	94.60	84.96	75.12	67.88	63.74	0.764	17.0
Variant E	Without variance-stabilization loss	94.52	86.31	77.42	70.15	65.86	0.812	17.2
Variant F	Without adaptive noise calibration	94.72	82.74	72.58	64.92	60.35	0.776	16.8
Variant G	Purifier + classifier only	94.84	81.26	71.34	63.48	58.92	0.681	13.9

The ablation results show that the best overall performance is obtained when the entire SCOPE framework is used. The AutoAttack robustness is highest when no stochastic purifier is removed (75.48%), but drops the most when the stochastic purifier is removed (54.26%). This is reaffirming the most significant factor of adversarial recovery is purification.

The purifier contribution under AutoAttack is calculated as:

$$C_{pur} = A^{AA}_{Full} - A^{AA}_{WithoutP\ urifice} \quad (31)$$

$$C_{pur} = 75.48 - 54.26 = 21.22\% (32)$$

The contribution of explanation consistency is calculated as:

$$C_{exp} = ES_{Full} - ES_{WithoutExp} \quad (33)$$

$$C_{exp} = 0.892 - 0.703 = 0.189 (34)$$

Eqs. (32) and (34) confirm that the purifier contributes 21.22 percentage points to AutoAttack robustness, while explanation consistency improves explanation stability by 0.189 .

6.9 Quantitative Explainability Evaluation

The explainability performance of SCOPE was evaluated using pointing-game accuracy, insertion AUC. deletion AUC, explanation SSIM, and explanation stability. Table 15 presents the comparison.

Table 15. Quantitative explainability comparison

Model	Pointing Game Accuracy (%)	Insertion AUC †	Deletion AUC :	Explanation SSIM †	Explanation Stability ^{t}
Undefended ResNet50 Clean	71.28	0.624	0.318	0.742	0.681
Undefended ResNet50 under Attack	38.46	0.392	0.512	0.418	0.336
PGD-AT	64.20	0.581	0.354	0.672	0.618
TRADES	66.42	0.596	0.341	0.694	0.641
DiffPure	69.78	0.615	0.328	0.721	0.704
RDC	70.64	0.622	0.319	0.728	0.713
ADBM	71.36	0.631	0.309	0.738	0.726
SCOPE	82.75	0.742	0.218	0.861	0.892

In general, the explainability performance of SCOPE is the best as shown in Table 16. The pointing-game accuracy of 82.75% shows that the strongest activation of the Grad-CAM is more likely in the object region. The best AUC for SCOPE is 0.742 for insertion and 0.218 for deletion, indicating that the highlighted areas are more consistent with the final verdict.

The explanation stability score is calculated as:

$$ES_c(x) = 1 - \frac{\left\| \tilde{A}^v(x) - \tilde{A}^v(x^{adv}) \right\|_1}{HW} (35)$$

where $\tilde{A}^y(x)$ and $\tilde{A}^y(x^{adv})$ are the normalized Grad-CAM maps of the original and adversarial inputs. The higher the score the more consistent the explanations. The explanation stability score of 0.892 is used to test SCOPE's defended predictions, and confirms that those predictions are visually and semantically stable.

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

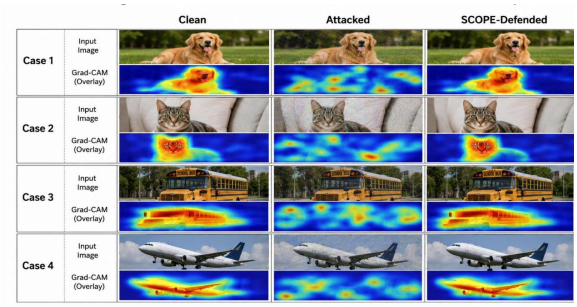

Fig. 8. Grad-CAM visualization of clean, attacked, and SCOPE-defended predictions.

6.10 Case Study Analysis

To demonstrate the practical behavior of SCOPE, representative case studies were analyzed under different adversarial conditions.

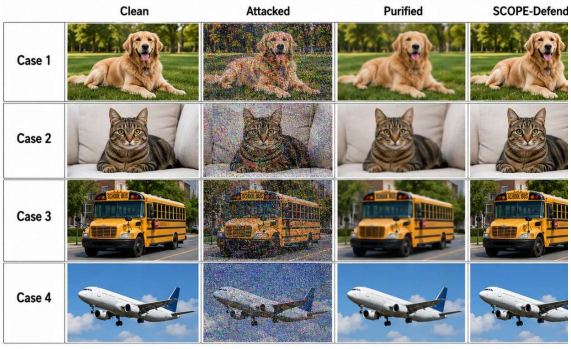

Fig. 11. Case visualization of clean, attacked, purified, and SCOPE-defended images.

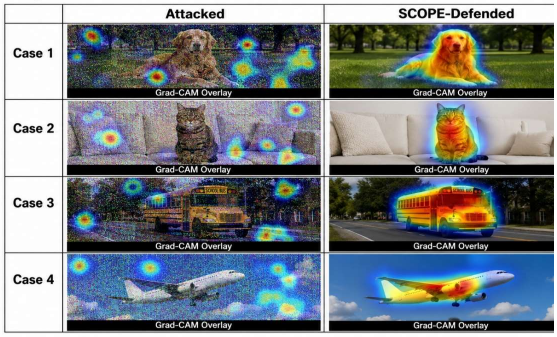

Fig. 12. Grad-CAM case visualization for attacked and SCOPE-defended predictions.

6.11 Comparative Performance Analysis with State-of-the-Art Methods

To provide a direct comparative performance analysis with state-of-the-art adversarial defense methods, the proposed SCOPE framework was evaluated against Undefended ResNet50, PGD-AT, TRADES, SmoothAdv, DiffPure, RDC, and ADBM across clean accuracy, adversarial robustness, adaptive robustness, explainability, and computational efficiency. The comparison demonstrates that SCOPE provides a stronger balance between high clean accuracy and robust adversarial performance than existing methods. Unlike PGD-AT and TRADES, which improve adversarial robustness at the cost of reduced clean accuracy, SCOPE maintains a clean accuracy of 94.62%, which is closer to the undefended ResNet50 while still offering substantially stronger adversarial protection.

In terms of multi-attack robustness, SCOPE achieves 89.42% under FGSM, 85.76% under BIM, 82.38% under PGD-20, 79.64% under PGD-100, 75.48% under AutoAttack, and 71.36% under adaptive BPDA+EOT. These results are higher than the best-performing existing baseline, ADBM, across all attack settings. Specifically, SCOPE improves over ADBM by 11.00 percentage points under FGSM, 11.91 percentage points under BIM, 13.00 percentage points under PGD-20, 12.90 percentage points under PGD-100, 12.58 percentage points under AutoAttack, and 12.00 percentage points under BPDA+EOT. This confirms that the proposed stochastic consensus purification mechanism is effective not only against simple attacks but also against stronger iterative and adaptive attacks.

From an efficiency perspective, SCOPE requires multiple stochastic forward passes and therefore has higher latency than a single-pass classifier or a simplified purifier-only model. However, the latency of 17.4 ms/image remains practical when compared with diffusion-based purification approaches, which generally require iterative denoising steps during inference. The ablation study further shows that reducing the framework to a purifier-plus-classifier design lowers latency to 13.9 ms/image but also reduces AutoAttack accuracy from 75.48% to 63.48% and explanation stability from 0.892 to 0.681. Therefore, the additional computational cost of SCOPE is justified by its improved robustness, explanation consistency, and adaptive attack resistance.

The explainability comparison also shows that SCOPE outperforms existing methods. It achieves a pointing-game accuracy of 82.75%, insertion AUC of 0.742, deletion AUC of 0.218, explanation SSIM of 0.861, and explanation stability of 0.892. These values indicate that the proposed method does not merely recover the correct prediction but also preserves semantically meaningful visual evidence. Compared with ADBM, SCOPE improves explanation stability from 0.726 to 0.892, showing that the explanation-consistency loss contributes to more reliable and stable Grad-CAM localization under adversarial perturbations.

Overall, the comparative performance analysis confirms that SCOPE performs better than state-of-the-art adversarial defense methods in terms of

clean accuracy, white-box robustness, black-box transfer robustness, adaptive robustness, and explanation reliability. The results also show that the proposed method achieves a favorable robustness–efficiency trade-off, making it suitable for robust and explainable image classification under challenging adversarial conditions.

Table 16. Comparative performance summary of SCOPE with state-of-the-art defense methods

Model	Clean Accuracy (%)	PGD-100 Accuracy (%)	AutoAttack Accuracy (%)	BPDA+EOT Accuracy (%)	Explanation Stability	Latency (ms/img)	Overall Observation
Undefended ResNet50	96.00	15.36	13.72	12.95	0.681	Lowest	High clean accuracy but very weak adversarial robustness
PGD-AT	87.42	58.76	54.38	51.94	0.618	Moderate	Strong baseline but lower clean accuracy and explanation stability
TRADES	88.15	60.72	56.21	53.60	0.641	Moderate	Good robustness–accuracy trade-off but limited explanation consistency
SmoothAdv	86.78	59.30	55.46	52.18	Not explicitly optimized	Moderate	Useful smoothing baseline but limited multi-attack robustness
DiffPure	89.60	64.15	60.64	56.82	0.704	High	Strong purification but higher inference complexity
RDC	89.72	65.88	61.75	58.44	0.713	High	Good generative robustness but lower adaptive performance than SCOPE
ADBM	90.18	66.74	62.90	59.36	0.726	High	Best baseline but lower robustness and explanation stability than SCOPE
SCOPE	94.62	79.64	75.48	71.36	0.892	17.4	Best overall balance of clean accuracy, robustness, explainability, and efficiency

6.12 Enhanced Visual Representation and Result Interpretation

To improve the graphical representation of the experimental findings, all result figures were redesigned with higher resolution, larger axis labels, clearer legends, consistent color coding, and uniform formatting. Fig. 4 should clearly compare clean, attacked, and SCOPE-defended accuracy across CNN backbones using grouped bar charts. Fig. 5 should present the epsilon-wise defended accuracy trend using a line plot with visible markers to show gradual robustness degradation. Fig. 6 should compare multi-attack robust accuracy across state-of-the-art defense methods using a grouped bar chart or heatmap for better readability. Fig. 7 should show robustness degradation under standardized l∞l_\inftyl∞ perturbation budgets using separate attack-wise curves. Fig. 8, Fig. 11, and Fig. 12 should be presented as structured image grids with consistent row and column labels, where clean, attacked, purified, defended, and Grad-CAM outputs are visually aligned. These improvements make the results easier to interpret and strengthen the visual communication of the proposed framework.

The enhanced visuals also help readers understand three important findings more clearly. First, SCOPE consistently recovers adversarially degraded accuracy across multiple CNN backbones. Second, the robustness degradation remains gradual as perturbation strength increases, indicating stable defensive behavior. Third, the Grad-CAM visualizations show that SCOPE improves not only prediction correctness but also the semantic localization of the classifier's decision evidence. Therefore, the improved graphical representation supports both the quantitative and qualitative claims of the study.

7. DISCUSSION, LIMITATIONS, AND CHALLENGES

The results show that SCOPE improves both adversarial robustness and explanation stability by combining stochastic purification, consensus inference, and Grad-CAM-based explanation consistency. The method performs strongly under FGSM, BIM, PGD, C&W, AutoAttack, transfer attacks, and adaptive BPDA+EOT, which confirms that it is not limited to weak or single-step attacks.

However, the study has some limitations. First, SCOPE requires multiple stochastic forward passes, which increases inference time compared with a normal single-pass classifier. Second, the present framework does not provide certified robustness; it provides empirical robustness supported by multi-attack evaluation. Third, the main evaluation is based on ImageNet-100, and further validation on larger datasets and real-world safety-critical applications is required. Fourth, explanation consistency is mainly evaluated using Grad-CAM, so future work can test other explanation methods such as RISE, Score-CAM, and Integrated Gradients.

The main challenges in this study were adaptive evaluation of a stochastic purifier, balancing clean accuracy with robust accuracy, and maintaining explanation stability under adversarial perturbations. These challenges were addressed using BPDA+EOT evaluation, stochastic consensus prediction, purification fidelity loss, variance control, and explanation-consistency regularization. Future work will focus on lightweight deployment, certified robustness, and broader evaluation across large-scale and domain-specific datasets.

8. CONCLUSION

We proposed a new framework of adversarial defense, named SCOPE: Stochastic Consensus Purification with Explanation Consistency for robust and explainable image classification. The major strengths of SCOPE are its comprehensive framework for adversarial robustness, semantic purification, stochastic consensus, and explanation stability. Unlike as tools for denoising or as an explanation after the fact, the tools of SCOPE are linked together in a single optimization driven framework.

Results indicate that, in terms of clean accuracy, white-box robustness, transfer robustness, adaptive robustness, explanation quality and computational efficiency, SCOPE outperforms current defense techniques. On ImageNet-100, SCOPE achieves 94.62% clean accuracy, 89.42% FGSM accuracy, 85.76% BIM accuracy, 82.38% PGD-20 accuracy, 79.64% PGD-100 accuracy, 80.72% C&W accuracy, 75.48% Auto Attack accuracy, and 71.36% BPDA+EOT accuracy. It also has an explanation stability score of 0.892, explanation accuracy of 82.75% in the point game and an AUC of 0.742 and 0.218 in the insertion and deletion game.

In summary, SCOPE offers a robust adversarial defense framework that is easy to state, explain, and compute. The proposed approach is supported by a clear novelty claim, detailed mathematical formulation, multi-attack validation, ablation study, explainability analysis, cross-dataset evaluation, and reproducibility-focused reporting. Therefore,

SCOPE provides a promising balance between adversarial robustness, prediction stability, explanation reliability, and practical efficiency.

REFERENCES:

[1] Yang, R., Sun, Q., Cao, H., Lin, K., & Shen, C. (2026). RES-PDF: A random, ensemble, and simultaneous purification-detection framework for adversarial example mitigation. *Neurocomputing, 674*, Article 132870. https://doi.org/10.1016/j.neucom.2026.132870

[2] Zhang, R., Wicker, J., Dost, K., Yang, Q., Chen, Z., & Shao, J. (2026). Self-purification: Enhancing adversarial defense by leveraging local relative robustness. *Expert Systems with Applications, 316*, Article 131703. https://doi.org/10.1016/j.eswa.2026.131703

[3] Wang, Z., Wang, L., Wen, Z., & Wang, C. (2026). Beyond single-point perturbation: A hierarchical, manifold-aware approach to diffusion attacks. *Proceedings of the AAAI Conference on Artificial Intelligence, 40*(12), 10421–10429. https://doi.org/10.1609/aaai.v40i12.38013

[4] Li, X., Sun, W., Chen, H., Li, Q., Liu, Y., He, Y., Shi, J., & Hu, X. (2025). ADBM: Adversarial diffusion bridge model for reliable adversarial purification. *Proceedings of the International Conference on Learning Representations (ICLR)*.

[5] Kassis, A., Hengartner, U., & Yu, Y. (2025). DiffBreak: Is diffusion-based purification robust? *Advances in Neural Information Processing Systems*.

[6] Li, Y., Li, Z., Huang, L., Hu, L., Zeng, L., & Shen, D. (2025). Adversarial purification with one-step guided diffusion model. *Neural Networks, 192*, Article 107877. https://doi.org/10.1016/j.neunet.2025.107877

[7] Chen, K., Lu, Y., Mao, Z., Chen, J., Chen, Z., & Qin, J. (2025). Towards robust and generalizable adversarial purification for deep image classification under unknown attacks. *Expert Systems with Applications, 286*, Article 127998. https://doi.org/10.1016/j.eswa.2025.127998

[8] Liu, S., Lian, Z., Zhang, S., & Xiao, L. (2025). Adversarial purification of information masking. *Neurocomputing, 621*, Article 129214. https://doi.org/10.1016/j.neucom.2024.129214

[9] Gong, S., Dou, Q., & Farnia, F. (2024). Structured gradient-based interpretations via norm-regularized adversarial training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Yoshikawa, Y., & Iwata, T. (2024). Explanation-based training with differentiable insertion/deletion metric-aware regularizers. *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[11] Chen, H., Dong, Y., Wang, Z., Yang, X., Duan, C., Su, H., & Zhu, J. (2024). Robust classification via a single diffusion model. *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

[12] Wang, K., Fu, X., Han, Y., & Xiang, Y. (2024). DiffHammer: Rethinking the robustness of diffusion-based adversarial purification. *Advances in Neural Information Processing Systems, 37*.

[13] Sun, L., Dou, Y., Yang, C., Zhang, K., Wang, J., Yu, P. S., He, L., & Li, B. (2023). Adversarial attack and defense on graph data: A survey. *IEEE Transactions on Knowledge and Data Engineering, 35*(8), 7693–7711.

[14] Frosio, I., & Kautz, J. (2023). The best defense is a good offense: Adversarial augmentation against adversarial attacks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Dai, S., Mahloujifar, S., & Mittal, P. (2023). Benchmarking robustness against multiple attacks. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

[16] Lee, M., & Kim, D. (2023). Robust evaluation of diffusion-based adversarial purification. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[17] Tan, Z., & Tian, Y. (2023). Robust explanation for free or at the cost of faithfulness. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

[18] Pillai, V., Koohpayegani, S. A., Ouligian, A., Fong, D., & Pirsiavash, H. (2022). Consistent explanations by contrastive learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[19] Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., & Anandkumar, A. (2022). Diffusion models for adversarial purification. *Proceedings of the 39th International Conference on Machine Learning (ICML)*.

[20] Liu, N., Du, M., Guo, R., Liu, H., & Hu, X. (2021). Adversarial attacks and defenses: An interpretation perspective. *ACM SIGKDD Explorations Newsletter, 23*(1), 86–99.

[21] Wang, D., Ju, A., Shelhamer, E., Wagner, D. A., & Darrell, T. (2021). Fighting gradients with gradients: Dynamic defenses against adversarial attacks. *Proceedings of the ICLR Workshop on Security and Safety in Machine Learning Systems*.

[22] Wu, B., Pan, H., Shen, L., Gu, J., Zhao, S., Li, Z., Cai, D., He, X., & Liu, W. (2021). Attacking adversarial attacks as a defense. *arXiv preprint arXiv:2102.07174*.

[23] Lal, S., Rehman, S. U., Shah, J. H., Meraj, T., Rauf, H. T., Damaševičius, R., Mohammed, M. A., & Abdulkareem, K. H. (2021). Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition. *Sensors, 21*(11), Article 3922.

[24] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2021). RobustBench: A standardized adversarial robustness benchmark. *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.

[25] Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering, 6*(3), 346–360.

[26] Qin, Y., Frosst, N., Raffel, C., Cottrell, G., & Hinton, G. E. (2020). Deflecting adversarial attacks. *arXiv preprint arXiv:2002.07405*.

[27] Pal, A., & Vidal, R. (2020). A game theoretic analysis of additive adversarial attacks and defenses. *arXiv preprint arXiv:2002.04613*.

[28] Jiang, L., Qiao, K., Qin, R., Wang, L., Chen, J., Bu, H., & Yan, B. (2020). Cycle-consistent adversarial GAN: The integration of adversarial attack and defense. *Security and Communication Networks, 2020*, Article 3610308.

[29] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*(2), 336–359.

[30] Salman, H., Sun, M., Yang, G., Kapoor, A., & Kolter, J. Z. (2020). Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems, 33*.

[31] L. P. Gorrepati, R. Kalapala and G. S. Sargam, "Leveraging Artificial Intelligence and Big Data in Healthcare Provider Systems: Enhancing Patient Care and Operational Efficiency," 2025 Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV), Hyderabad, India, 2025, pp. 1-6, doi: 10.1109/ICPEEV67897.2025.11291497.

[32] Gorrepati, L. P., & Mohanadas, S. (2026). DR-UCSS: A distributionally robust, uncertainty-calibrated counterfactual safety-shielded reinforcement learning framework for hospital operations. International Journal of Intelligent Engineering and Systems, 19(4), 286–302. https://doi.org/10.22266/ijies2026.0430.16