UNDERSTANDING ADVERSARIAL ROBUSTNESS THROUGH SEMANTIC-REGION MEMBERSHIP

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial robustness remains an important research topic for improving the reliability and security of deep neural networks. Conventional research on adversarial attacks and defences mainly evaluates robustness through prediction accuracy on clean images and their corresponding adversarial examples. In this paper, we define class-specific semantic regions based on the geometric organisation of learned feature representations and conduct experiments across a range of adversarially robust models and a model trained only on clean images. Our results reveal that adversarial robustness is not associated with stronger semantic organisation of clean representations, but rather with the preservation of semantic-region membership under adversarial perturbations. we derive a sufficient condition to establish the relationship between semantic-region membership preservation and adversarial feature displacement. We employ a feature-space PGD attack to empirically estimate adversarial feature displacement and evaluate the derived sufficient condition for semantic-region membership preservation. The results show that all feature representations of adversarial examples lie within their corresponding semantic region if they satisfy that condition. Another interesting finding is that even an adversarial example has been misclassified, its feature representation may still lie within its corresponding semantic region. These findings extends the robustness studies from conventional prediction-level to semantic organisation in learn feature space.

1 INTRODUCTION

Since the vulnerability of deep neural networks (DNNs) to adversarial examples was revealed by Biggio et al. (2013) and Szegedy et al. (2014), improving and explaining the robustness of DNNs has been an important research topic in machine learning. Different defence strategies, such as Adversarial Training (AT) Madry et al. (2017); Zhang et al. (2019b), AT with synthetic data Bartoldson et al. (2024); Wang et al. (2023), and post-training representation modification Amini et al. (2024) achieve state-of-the-art robustness performance on benchmark datasets. Their performance is mainly reported in terms of clean and adversarial prediction accuracies, which provide limited information about how semantic information is organised within the learned feature space and how this organisation changes under adversarial perturbations.

A number of studies have investigated adversarial robustness from the perspective of learned features. Garg et al. (2018) use a graph Laplacian spectral approach to construct insensitive but sufficient variance features from the neighbourhood geometry of the data to feed into a linear classifier. The improved robustness suggests that adversarial robustness can be studied from the perspective of feature representations and their underlying data geometry. Ilyas et al. (2019) revealed that DNNs can learn either features that are predictive but sensitive to small perturbations or robust features, and that adversarial robustness is closely related to the information captured within learned representations.

Beyond individual robust and non-robust features, the geometric organisation of learned representations has also been studied. Neural Collapse Papyan et al. (2020) revealed that the learned features of samples belonging to the same class converge towards their corresponding class centroids, these class centroids exhibit Simplex Equiangular Tight Frame (ETF), and the predictions can be determined by the class centroid with the shortest distance to the learned feature of a sample. Zhu et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2021) explained that the ETF structure formed by the class centroids corresponds to a global optimum of the cross-entropy loss, which is commonly used to train classifiers. This suggests that a class centroid can serve as a semantic representation of a class in the feature space. Su et al. (2024) further observed that, in robust models, not only do the clean learned features form an ETF structure, but the adversarial learned features also form an ETF structure that is aligned with the one formed by the clean features.

These feature-based studies provide a semantic organisation in feature space with respect to DNNs. However, how class-specific semantic information learned by a model can be characterised within its feature space, and how consistently this information is maintained under adversarial perturbations, have not been fully investigated. In particular, the relationship between the preservation of such model-learned semantic information and adversarial robustness remains unclear. Moreover, it remains unclear whether an adversarial prediction failure necessarily corresponds to the loss of the ground-truth semantic information within the learned representation. To address these questions, we use geometric organisation of learned features to represent class-specific semantic information with respect to DNNs, investigate how this semantic information is organised and preserved under perturbation across standard and adversarially robust models with different defence strategies, and monitor the recovery of semantic information as the perturbations are gradually removed from adversarial examples. Our results show that adversarial robustness is more closely associated with the preservation of the membership of class-specific semantic geometric organisation than with stronger organisation of clean feature representations, and that the prediction failure of an adversarial example does not necessarily mean that its learned features have lost their semantic meaning. The remainder of this paper introduces our contributions, related work, methodology, experiments and results, and finally, the conclusion.

Our main contributions are listed as follows:

- We introduce class-specific semantic regions within DNN feature spaces based on the class related learned feature centroids. We validate that these class centroids remain stable across independently sampled subsets.
- We systematically analyse the membership preservation of semantic regions under adversarial perturbations across standard and adversarially robust models. The results show that robustness is more closely associated with the capability to preserve semantic-region membership under perturbations than with stronger semantic geometric organisation of clean features. We further show that an incorrect adversarial prediction does not necessarily imply that ground-truth semantic information has been lost from the learned representation.
- We derive a sufficient condition based on the semantic margin and maximum feature displacement that guarantees the feature representation of an adversarial example remains within the corresponding semantic region. We further provide empirical evidence for this condition using feature-space PGD.

2 Related Work

2.1 Class Representation Centroid

The organisation of samples in the learned feature space is important for classification. Wen et al. (2016) introduced centre loss to encourage samples from the same class to stay close together. Their motivation was that the conventional cross-entropy loss helps the model separate different classes for classification, but does not directly control how far apart samples from the same class can be. This is particularly important for their task of face recognition. The centre loss encourages the learned features from images belonging to the same class to be closer to the centroid dynamically constructed from the corresponding learned features, which aims to reduce the impact of variations in images of the same person due to changes in pose, illumination and appearance. Their experiments on face recognition datasets showed that combining centre loss with cross-entropy loss produced more discriminative feature representations and improved recognition performance compared with using cross-entropy loss alone. A further application of learned feature centroids can be found in the prototypical networks proposed by Snell et al. (2017). In the context of their few-shot classification task, a prototype actually refers to the feature centroid of a class. They calculate a centroid for each class from the feature representations of labelled examples belonging to that class. A new sample

is then classified by comparing its feature representation with the centroids of different classes and assigning it to the closest one. Despite the additional details of their episodic training strategy, their work provides an example showing that a class can be represented by a representative vector in feature space and that the distance between a sample and these representative vectors can be used for classification.

2.2 Neural Collapse and Class Geometry

A further perspective on class-centred representation structure was provided by Papyan et al. (2020) through a phenomenon called neural collapse. Unlike methods that explicitly encourage samples to move towards a class centroid, neural collapse can naturally appear during standard classification training. Papyan et al. studied the final stage of training, where the training classification error has already reached approximately zero, but the model continues to be trained to reduce the loss. They observed four related behaviours. The first one describes what happens to samples from the same class. As training continues, their feature representations become increasingly similar and move closer to their class centroids. Therefore, samples belonging to the same class become more concentrated around a common centroid. The second one describes what happens between different class centroids. They observed that the class centroids become highly organised in the feature space. After centring and normalisation, they have equal norms and become evenly separated from each other, forming a symmetric geometric structure known as a simplex equiangular tight frame (ETF). The third one describes the relationship between the class centroids and the final classifier. As training continues, the weights of the final linear classifier become increasingly aligned in direction with their corresponding class centroids. The final one describes how the model makes its classification decision. They observed that the classifier increasingly behaves like a nearest-class-centroid classifier. In other words, the classifier increasingly assigns a sample to the class with the nearest class centroid. Zhu et al. (2021) provided a theoretical explanation for the emergence of neural collapse.

These observations further suggest that the centroids constructed from learned feature representations belonging to different classes can semantically represent their corresponding classes.

2.3 Class Geometry Changes under Perturbation

Su et al. (2024) investigated changes in the geometric organisation of learned features under adversarial perturbations by extending the study of neural collapse to the adversarial setting. They observed that, under small perturbations, the simplex ETF structure of the class centroids in neural collapse can be destroyed; however, in adversarially trained robust models, the simplex ETF structure of the class centroids exists for both clean images and adversarial examples. This is one of the studies most similar to ours. The differences are that they did not explicitly interpret these simplex ETFs as semantic representations of classes with respect to the corresponding model and they did not study the relationship between individual representation and the simplex ETFs.

3 Methodology

The methodology first defines what constitutes a semantic region and how semantic-region membership preservation is measured under adversarial perturbation. It then derives a sufficient theoretical condition describing when a sample can remain within its corresponding semantic region under bounded perturbations.

3.1 Class-specific Semantic Region

The output of the penultimate layer of a classifier can be considered as the feature representation $f(x) \in \mathbb{R}^d$ of an input x. A class c can be represented by its feature representation centroid μ_c, which is defined as follows:

$$\mu_c = \frac{1}{N_c} \sum_{i:c_i=c} f(x_i), \tag{1}$$

where i denotes the index of a reference sample, c_i denotes its class label, and N_c denotes the number of reference samples belonging to class c.

The class-specific semantic region $\mathcal{R}_c$ associated with class c is defined as follows:

$$\mathcal{R}_c = \left\{ f(x) \in \mathbb{R}^d : \|f(x) - \mu_c\|_2 < \|f(x) - \mu_j\|_2, \quad \forall j \neq c \right\}, \tag{2}$$

where $f(x)$ is the feature representation of x. This definition describes that if a sample's representation is closer to the centroid of class c than to the centroid of any competing class, it belongs to the semantic region $\mathcal{R}_c$. In this work, the term class-specific semantic region refers to such class-specific regions in the feature space.

For a sample belonging to class c, the semantic margin $m(x)$ is defined as follows:

$$m(x) = \min_{j \neq c} \|f(x) - \mu_j\|_2 - \|f(x) - \mu_c\|_2. \tag{3}$$

A positive semantic margin indicates that the feature representation is closer to the centroid of its corresponding class than to any competing class centroid. Therefore,

$$m(x) > 0 \iff f(x) \in \mathcal{R}_c. \tag{4}$$

The magnitude of a positive margin describes how much farther the representation is from its nearest competing class centroid than from its corresponding class centroid.

3.2 Semantic-Region Membership Consistency under Adversarial Perturbation

Given a clean input x and its corresponding adversarial input x_{adv}, semantic-region membership is preserved when

$$f(x) \in \mathcal{R}_c \quad \text{and} \quad f(x_{\text{adv}}) \in \mathcal{R}_c, \tag{5}$$

where $\mathcal{R}_c$ denotes the semantic region associated with the class c of the clean input. It can be represented by the semantic margin as follows:

$$m(x) > 0 \text{ and } m(x_{\text{adv}}) > 0.. \tag{6}$$

The conditional semantic-region membership survival rate $P_{\text{survive}\,c}$ for a specific class c is defined as follows:

$$\begin{aligned} P_{\text{survive},\,c} &= P\left(f(x_{\text{adv}}) \in \mathcal{R}_c \mid f(x) \in \mathcal{R}_c\right) \\ &= \frac{\text{count}\left(m(x_{\text{adv}}) > 0 \wedge m(x) > 0\right)}{\text{count}\left(m(x) > 0\right)}, \end{aligned} \tag{7}$$

where x is a clean sample belonging to class c, $m(\cdot)$ is the semantic margin, and $\text{count}(\cdot)$ denotes the number of representations satisfying the specified condition. This directly measures the probability that an existing semantic-region membership is preserved under adversarial perturbation.

The conditional semantic-region survival rate P_{survive} at the dataset level is defined as follows:

$$P_{\text{survive},\mathcal{D}} = \frac{1}{C} \sum_{c=1}^{C} P_{\text{survive},\,c}, \tag{8}$$

where $\mathcal{D}$ denotes a dataset and C is the number of classes in $\mathcal{D}$. This metric represents the average semantic-region membership survival rate across classes in the whole dataset.

3.3 Interpolation-Based Semantic Recovery Analysis

An interpolation trajectory is constructed between each adversarial input and its corresponding clean input. The intermediate state $x(\lambda)$ between the two endpoints is defined as follows:

$$x(\lambda) = (1 - \lambda)x_{\text{adv}} + \lambda x, \tag{9}$$

where $\lambda \in [0, 1]$, $\lambda = 0$ corresponds to the adversarial endpoint, and $\lambda = 1$ corresponds to the clean endpoint. For each clean and adversarial input pair, the interpolation path is evaluated at $\lambda \in \{0,\ 0.1,\ 0.25,\ 0.5,\ 0.75,\ 0.9,\ 1.0\}$. The proportion of representations with $m(x(\lambda)) > 0$ are evaluated along the interpolation path. It is used to measure the percentage of representations that lie within their corresponding semantic regions at each interpolation point. For this analysis, 256 samples for each experimental run are randomly selected from each evaluation subset.

3.4 Theoretical Condition for Semantic-Region Consistency

The semantic region of a sample can be preserved if its feature representation does not move too far under perturbation. For an input x and an ℓ_∞ perturbation budget ϵ, the maximum feature displacement $D(x, \epsilon)$ is defined as follows:

$$D(x, \epsilon) = \sup_{\|\delta\|_\infty \leq \epsilon} \|f(x + \delta) - f(x)\|_2, \tag{10}$$

where $\sup$ is the supremum, ϵ is the perturbation budget, and δ is the perturbation. In simple terms, this equation is saying that $D(\cdot)$ is the maximum possible feature displacement when δ is bounded within an ϵ-ball. A straightforward way would be to try all possible adversarial examples $x + \delta$ and find the largest distance between the feature representation of each adversarial example and that of the corresponding clean image.

Proposition. Consider a sample from class c whose feature representation is within its ground-truth semantic region, described with $m(x) > 0$. If $D(x, \epsilon) < \frac{m(x)}{2}$, then its feature representation remains in the same semantic region for any perturbation satisfying $\|\delta\|_\infty \leq \epsilon$.

Proof. Under perturbation, the feature representation may move away from its class centroid and towards a competing class centroid. For any competing class $j \neq c$, the triangle inequality gives

$$\|f(x + \delta) - \mu_c\|_2 \leq \|f(x) - \mu_c\|_2 + D(x, \epsilon), \tag{11}$$

$$\|f(x + \delta) - \mu_j\|_2 \geq \|f(x) - \mu_j\|_2 - D(x, \epsilon). \tag{12}$$

Thus, $m(x + \delta) \geq m(x) - 2D(x, \epsilon) > 0$ whenever $D(x, \epsilon) < m(x)/2$, implying $f(x + \delta) \in \mathcal{R}_c$.

3.5 Empirical Validation of the Sufficient Condition

The exact value of $D(x, \epsilon)$ is difficult to obtain because all possible perturbations within the perturbation budget would need to be considered. Therefore, $D(x, \epsilon)$ is estimated using projected gradient descent (PGD) Madry et al. (2017).

For each clean input, unlike the classification attacks, PGD searches for a perturbation that maximises the feature displacement. Specifically, it approximately solves

$$\widehat{\delta} \approx \arg \max_{\|\delta\|_\infty \leq \epsilon} \|f(x + \delta) - f(x)\|_2. \tag{13}$$

where $\widehat{\delta}$ is the perturbation found by PGD under the constraint $\|\widehat{\delta}\|_\infty \leq \epsilon$. The estimated displacement is then defined as follows:

$$\widehat{D}(x, \epsilon) = \|f(x + \widehat{\delta}) - f(x)\|_2. \tag{14}$$

Using the estimated displacement, the sufficient condition is tested as follows:

$$\widehat{D}(x, \epsilon) < \frac{m(x)}{2}. \tag{15}$$

For samples satisfying this condition, the experiment examines whether the perturbed feature representation remains in its corresponding semantic region, described as $f(x + \widehat{\delta}) \in \mathcal{R}_c$.

Since PGD may not find the largest possible feature displacement, $\widehat{D}(x, \epsilon)$ is an estimate of $D(x, \epsilon)$. Therefore, the experiment provides empirical evidence for the sufficient condition rather than a formal robustness guarantee.

3.6 Semantic-Region Membership and Classification Prediction

The above sections defined the semantic region and developed methods to analyse it. However, it must be considered whether analysing adversarial robustness at the semantic level is useful and how it can be an extension of the analysis based on the conventional prediction level. In this section, a simple metric is introduced to link the analyses from both aspects as follows:

$$\mathcal{A} = \{x_{\text{adv}} \mid \hat{y}(x_{\text{adv}}) \neq y,\ f(x_{\text{adv}}) \in \mathcal{R}_y\}, \tag{16}$$

where $\mathcal{A}$ is a set of adversarial examples x_{adv} that are predicted as incorrect classes, but their feature representations are still within the corresponding semantic regions. This analysis examines the relationship between prediction failure and semantic-region membership.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

Experiments are conducted on CIFAR-10 and CIFAR-100 datasets Krizhevsky (2009). A class-balanced subset of $5{,}000$ training samples is used to construct the class centroids. A second independently sampled, non-overlapping, class-balanced subset of $5{,}000$ training samples is used as the reference subset. In addition, an evaluation subset is formed by sampling $1{,}000$ class-balanced samples from the test set. All experiments are repeated three times using different fixed random seeds for data subset sampling. This setting ensures consistent comparisons and reproducible experiments.

For each dataset, 14 models are evaluated. Thirteen adversarially robust models are obtained from the RobustBench model zoo Croce et al. (2021). These models are trained using different adversarial robustness mechanisms. In addition, a standard WRN-94-16 Zagoruyko & Komodakis (2016), trained locally using clean samples, is included as the non-robust baseline. The evaluated models and their clean, PGD-20 Madry et al. (2017) and AutoAttack Croce & Hein (2020) accuracies are reported in Table 4 in the Appendix. Three representative models are selected to illustrate the semantic-region behaviour across different levels of adversarial robustness. Amini et al. (2024) represents the highest level of adversarial robustness, Hendrycks et al. (2019) represents an intermediate level, and the locally trained standard WRN-94-16 represents the lowest level.

All experiments are conducted using Python on a computing system equipped with an Intel Xeon Platinum 8468 processor and 64 GB of system memory. Each experiment is allocated up to 8 CPU units and one NVIDIA A100 SXM4 GPU with 80 GB of GPU memory.

4.2 CLEAN SEMANTIC-REGION ANALYSIS

Clean Semantic-Region Analysis examines whether the learned feature representations form distinguishable class-specific regions. It reports the proportion of evaluation samples with $m(x) > 0$ to evaluate the formation of class-specific semantic regions in the clean feature space. The results for all models are available in Table 5 in the Appendix. The standard model achieves the highest proportion on both datasets.

4.3 VALIDATION OF THE SEMANTIC-REGION STRUCTURE

A random-label baseline is used to examine whether the observed semantic regions are associated with the class labels. By shuffling the class labels of the samples used to construct the class centroids, the samples can be regrouped according to their newly assigned labels, and the class centroids are recomputed. The evaluation subset is evaluated using these randomly constructed centroids. This process is repeated for 500 independent label permutations. The proportion of evaluation representations with $m(x) > 0$ obtained using the original class centroids is compared with that obtained using the randomly constructed class centroids. The results of the three selected models are shown in Table 6 in the Appendix. The results support that the semantic regions are class-specific as the regrouped representations fail to construct the observable semantic-region structure seen with the true class labels, with a permutation-test p-value remaining at 0.002 for all models on both datasets.

The centroid-stability experiment is conducted to examine whether the class centroids remain stable when different training samples are used to construct them. A second set of class centroids is constructed from the reference subset and compared with the original class centroids. For each class, the corresponding centroids from the two sets are compared using L_2 distance and cosine similarity. Results for the selected models are shown in Table 7 in the Appendix. The results support that the centroids computed from separate subsets of representations are very close, with high cosine similarity. This indirectly supports that a centroid constructed from a sufficient number of samples from a specific class can provide a stable representation of the corresponding class.

4.4 ADVERSARIAL SEMANTIC-REGION ANALYSIS

This experiment investigates what percentage of representations that initially belong to their corresponding semantic regions remain within those regions under perturbation. Adversarial examples are generated from the evaluation subset using PGD-20 Madry et al. (2017) and ℓ_∞ AutoAt-

tack Croce & Hein (2020), with a perturbation budget of $\epsilon = 8/255$. The conditional semantic-region survival rate $P_{\text{survive}\,\mathcal{D}}$ defined in Section 3.2 is evaluated to measure the preservation of the semantic-region membership under each attack. The results in Table 1 show that the adversarially robust models preserve a substantially higher proportion of their existing semantic-region membership under adversarial perturbations than the standard model.

Dataset	Model	P_{survive} (%)	
		PGD-20	AutoAttack
CIFAR-10	Amini et al. (2024)	$\mathbf{85.52 \pm 1.20}$	$\mathbf{89.24 \pm 0.49}$
	Hendrycks et al. (2019)	66.93 ± 0.52	72.72 ± 0.20
	Standard	0.00 ± 0.00	0.07 ± 0.05
CIFAR-100	Amini et al. (2024)	$\mathbf{70.16 \pm 1.16}$	$\mathbf{77.34 \pm 1.00}$
	Hendrycks et al. (2019)	54.68 ± 1.48	66.20 ± 1.57
	Standard	0.38 ± 0.38	0.50 ± 0.27

Table 1: Conditional semantic-region survival rates under PGD-20 and ℓ_∞ AutoAttack for the three representative models on CIFAR-10 and CIFAR-100. P_{survive} measures the proportion of samples whose adversarial feature representations remain within the semantic region of their true class, conditioned on their clean feature representations initially lying within that region, i.e., $m(x) > 0$. Higher values therefore indicate stronger preservation of the clean semantic region under adversarial perturbations. Results are reported as the mean $\pm$ standard deviation across three experimental runs.

4.5 Interpolation Analysis

The interpolation results are shown in Figure 1. Across both datasets, a clear difference can be observed between the standard and robust models as the inputs move from the adversarial endpoint towards the clean endpoint. The more robust the model, the more effectively semantic region membership is maintained along the interpolation trajectory. In contrast, the standard model loses almost all semantic-region membership near the adversarial endpoint, which is gradually recovered as the perturbation is progressively removed.

4.6 Sufficient-Condition Validation

The sufficient-condition validation examines whether the condition derived in Section 3.4 is supported by the empirical feature displacement under adversarial perturbations. Results in Table 2 show that when the feature displacement PGD adversarial examples satisfied the sufficient condition, none of their learned feature representations lie outside their true semantic region. However, the proportion of representations of the feature displacement PGD adversarial examples that actually remain within their semantic regions after the feature-space attack is considerably higher than the proportion of them satisfies the sufficient condition. This difference is expected because the derived condition is sufficient but not necessary: a representation may remain within its semantic region even when the sufficient condition is not satisfied. The results for all models are provided in Table 8 and Table 9 in the Appendix.

4.7 Prediction–Semantic Region Consistency Analysis

This analysis generates adversarial examples from the entire evaluation subset. For each generated adversarial example, the classification prediction and its feature representation are obtained. The analysis then counts how many adversarial examples are incorrectly classified while their feature representations remain within their true semantic regions. The proportion of such examples among all incorrectly classified adversarial examples is then reported. Two metrics are used in this analysis. The first one is $Wrong + TrueRegion(\%)$, which indicates the percentage of all generated adversarial examples that are incorrectly predicted while their feature representations still remain in their true semantic regions. The second one is $TrueRegionGivenWrong(\%)$, which indicates the percentage of incorrectly predicted adversarial examples whose feature representations are still within the corresponding true semantic regions. The results from the selected models are shown in

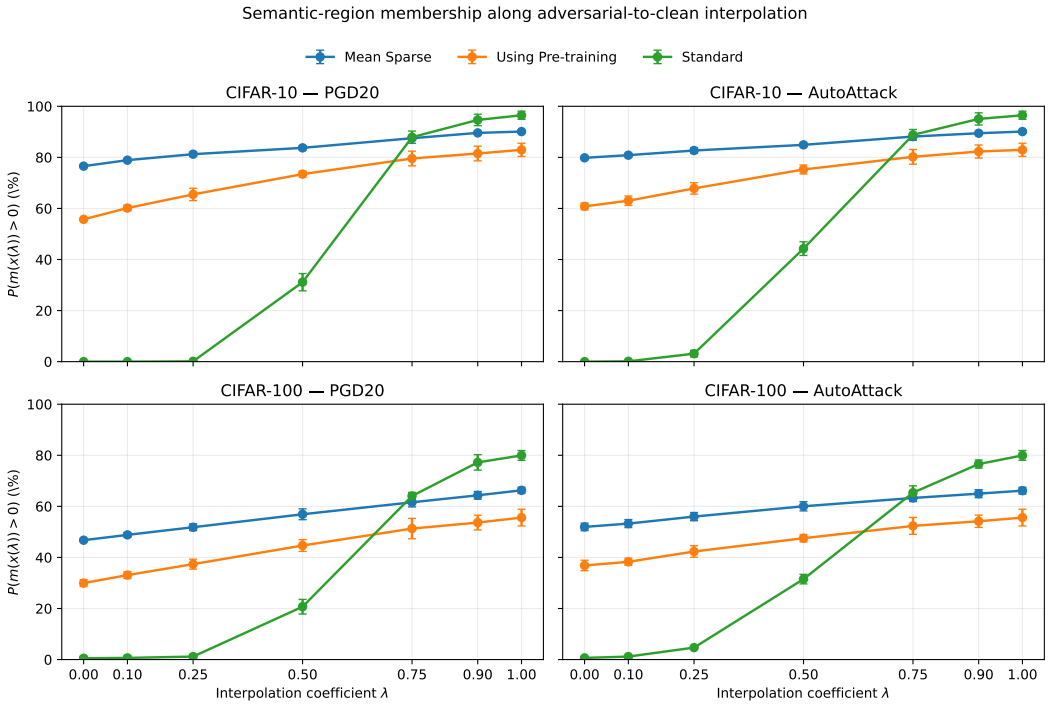

Figure 1: Semantic-region membership along the interpolation trajectory from adversarial examples ($\lambda = 0$) to their corresponding clean inputs ($\lambda = 1$) on CIFAR-10 and CIFAR-100 under PGD-20 and ℓ_∞ AutoAttack. For each run, 256 clean–adversarial sample pairs are interpolated at $\lambda \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1.0\}$. At each interpolation point, the proportion of the 256 representations lying within their corresponding true-class semantic regions, i.e., satisfying $m(x(\lambda)) > 0$, is calculated. The curves report the mean proportion across three independent runs, with error bars indicating the standard deviation.

Dataset	Model	Condition Satisfied (%)	Region Survival (%)	Condition Violation (%)
CIFAR-10	Amini et al. (2024)	57.03 ± 1.93	89.51 ± 0.41	0.00 ± 0.00
	Hendrycks et al. (2019)	5.25 ± 0.43	83.42 ± 0.65	0.00 ± 0.00
	Standard	0.00 ± 0.00	0.42 ± 0.15	0.00 ± 0.00
CIFAR-100	Amini et al. (2024)	15.25 ± 0.55	86.66 ± 0.62	0.00 ± 0.00
	Hendrycks et al. (2019)	7.24 ± 0.58	70.22 ± 1.80	0.00 ± 0.00
	Standard	0.00 ± 0.00	0.04 ± 0.06	0.00 ± 0.00

Table 2: Empirical validation of the sufficient condition for semantic-region preservation for the three representative models on CIFAR-10 and CIFAR-100. Condition Satisfied denotes the proportion of samples with $m(x) > 0$ for which the empirical feature displacement satisfies $\widehat{D}(x, \epsilon) < m(x)/2$. Region Survival denotes the proportion of all samples with $m(x) > 0$ whose feature representations remain within the semantic region of their true class under the feature-space PGD attack. Condition Violation denotes the proportion of samples satisfying the empirical sufficient condition that nevertheless leave their true-class semantic region under the attack. Results are reported as the mean $\pm$ standard deviation across three experimental runs.

Table 3. The results show that an incorrect adversarial prediction does not necessarily indicate that the corresponding feature representation has left its true-class semantic region.

Dataset	Model	Attack	Wrong + True Region (%)	True Region Given Wrong (%)
CIFAR-10	Amini et al. (2024)	PGD-20	2.97 ± 0.68	13.38 ± 2.91
		AutoAttack	4.77 ± 1.18	$\mathbf{18.99 \pm 4.38}$
	Hendrycks et al. (2019)	PGD-20	4.10 ± 0.33	9.62 ± 0.98
		AutoAttack	$\mathbf{6.77 \pm 0.21}$	14.84 ± 0.71
	Standard	PGD-20	0.00 ± 0.00	0.00 ± 0.00
		AutoAttack	0.07 ± 0.05	0.07 ± 0.05
CIFAR-100	Amini et al. (2024)	PGD-20	3.93 ± 0.54	8.03 ± 0.93
		AutoAttack	7.67 ± 1.16	$\mathbf{13.99 \pm 1.72}$
	Hendrycks et al. (2019)	PGD-20	3.60 ± 0.37	5.41 ± 0.53
		AutoAttack	$\mathbf{9.30 \pm 0.36}$	12.96 ± 0.58
	Standard	PGD-20	0.00 ± 0.00	0.00 ± 0.00
		AutoAttack	0.37 ± 0.17	0.37 ± 0.17

Table 3: Prediction–semantic region consistency under adversarial perturbations on CIFAR-10 and CIFAR-100. Wrong + True Region denotes the proportion of all evaluation samples that are incorrectly classified while their adversarial feature representations remain within the semantic region of the true class, i.e., $m(x_{\text{adv}}) > 0$. True Region Given Wrong denotes the proportion of incorrectly classified adversarial examples that satisfy $m(x_{\text{adv}}) > 0$, where the denominator includes only incorrectly classified adversarial examples. These measurements examine whether true-class semantic information can remain represented in the feature space even when the final class prediction is incorrect. Results are reported as the mean $\pm$ standard deviation over three experimental runs.

5 Conclusion

This study extends the analysis of adversarial robustness from the conventional softmax predictions to the learned feature space by defining class-specific semantic regions. The experimental results show that stronger organisation of clean representations does not naturally lead to adversarial robustness. The semantic regions constructed by robust models exhibit stronger semantic-region membership preservation under perturbation. The sufficient condition derived in this paper establishes a relationship between the semantic preservation and adversarial perturbations. Another important finding is that some adversarial examples are conventionally misclassified but their features can still remain within their true semantic regions. In other words, if the prediction were based on the semantic region to which a feature representation belongs, these adversarial examples would still be assigned to their true classes. This indicates that prediction-level failure does not necessarily correspond to the loss of true-class semantic information in the learned feature representation. This behaviour is observed for a proportion of incorrectly classified adversarial examples from the robust models but is rarely observed for the standard model. This provides further evidence that the semantic regions formed by robust models are more stable under adversarial perturbations and shows that analysing robustness only through the final classification prediction may not fully reflect the semantic information preserved within the learned feature representation.

References

Mohammad Amini et al. Meansparse: Post-training robustness enhancement through mean-centered feature sparsification. *arXiv preprint arXiv:2406.07317*, 2024.

Brian R. Bartoldson, James Diffenderfer, Bhavya Kailkhura, and Abhinav Gupta. Adversarial robustness limits via scaling-law and human-alignment studies. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3047–3072. PMLR, 2024.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine*

Learning and Knowledge Discovery in Databases, volume 8190 of *Lecture Notes in Computer Science*, pp. 387–402. Springer, 2013. doi: 10.1007/978-3-642-40994-3_25.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, and Percy Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11190–11201, 2019.

Jinghui Chen, Yu Cheng, Zhe Gan, Quanquan Gu, and Jingjing Liu. Efficient robust training via backward smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6222–6230, 2022. doi: 10.1609/aaai.v36i6.20571.

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

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *NeurIPS 2021 Datasets and Benchmarks Track*, 2021.

Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15721–15730, 2021.

Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled kullback–leibler divergence loss. In *Advances in Neural Information Processing Systems*, 2024.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020.

Shivam Garg, Vatsal Sharan, Brian Zhang, and Gregory Valiant. A spectral view of adversarially robust features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. 2020.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, pp. 17258–17277. PMLR, 2022.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Data augmentation can improve robustness. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29935–29948, 2021.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Chawin Sitawarin, Supriyo Chakraborty, and David A. Wagner. Improving adversarial robustness through progressive hardening. *arXiv preprint*, 2020.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Jingtong Su, Ya Shi Zhang, Nikolaos Tsilivis, and Julia Kempe. On the robustness of neural collapse and the neural collapse of robustness. *Transactions on Machine Learning Research*, 2024.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36246–36263. PMLR, 2023.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515, 2016.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

Yuancheng Xu, Yanchao Sun, Micah Goldblum, Tom Goldstein, and Furong Huang. Exploring and exploiting decision boundary dynamics for adversarial robustness. In *International Conference on Learning Representations*, 2023.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael Jordan. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, 2019a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 2019b.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29820–29834, 2021.

0ppendix

6 THE APPENDIX

CIFAR-10			
Model (Reference)	Clean (%)	PGD-20 (%)	AA (%)
Amini et al. (2024)	92.90	**78.47**	**74.90**
Bartoldson et al. (2024)	92.67	76.60	73.47
Wang et al. (2023)	92.57	73.43	69.87
Cui et al. (2024)	91.50	71.27	67.00
Xu et al. (2023)	92.63	69.70	63.53
Gowal et al. (2020)	89.47	67.90	64.80
Rade & Moosavi-Dezfooli (2021)	87.40	63.90	60.40
Carmon et al. (2019)	89.23	62.90	59.27
Hendrycks et al. (2019)	86.13	58.43	54.23
Sitawarin et al. (2020)	86.33	54.93	51.27
Zhang et al. (2019a)	85.57	47.50	44.43
Ding et al. (2020)	78.03	41.60	34.47
Wong et al. (2020)	78.37	40.27	36.03
Zagoruyko & Komodakis (2016)	**95.37**	0.00	0.00
CIFAR-100			
Model (Reference)	Clean (%)	PGD-20 (%)	AA (%)
Amini et al. (2024)	74.37	**50.03**	**43.80**
Wang et al. (2023)	74.67	47.60	41.40
Cui et al. (2024)	73.30	43.83	37.93
Wang et al. (2023)	73.00	43.23	37.73
Gowal et al. (2020)	68.50	39.50	36.07
Rebuffi et al. (2021)	60.93	36.30	31.23
Pang et al. (2022)	62.23	35.10	30.17
Gowal et al. (2020)	59.57	34.03	29.10
Hendrycks et al. (2019)	57.63	33.13	27.30
Cui et al. (2021)	68.93	31.13	26.37
Chen et al. (2022)	62.07	31.10	26.63
Sitawarin et al. (2020)	63.03	27.57	24.63
Rice et al. (2020)	52.37	21.20	18.97
Zagoruyko & Komodakis (2016)	**78.27**	0.33	0.00

Table 4: Benchmark models evaluated in this work. The adversarially robust models are obtained from RobustBench, and all models are re-evaluated. Clean accuracy and robust accuracy under PGD-20 and AutoAttack (AA) are reported. Within each dataset, the benchmark models are ordered according to their PGD-20 robust accuracy.

Dataset	Model	Clean Accuracy (%)	$m(x) > 0$ (%)
CIFAR-10	Amini et al. (2024)	93.30 ± 0.14	89.27 ± 0.62
	Bartoldson et al. (2024)	93.53 ± 0.12	89.30 ± 0.78
	Wang et al. (2023)	92.80 ± 0.50	91.37 ± 0.41
	Cui et al. (2024)	91.87 ± 0.65	91.07 ± 0.52
	Gowal et al. (2020)	90.20 ± 0.65	89.20 ± 0.29
	Xu et al. (2023)	93.33 ± 0.46	93.53 ± 0.37
	Rade & Moosavi-Dezfooli (2021)	87.90 ± 0.94	86.53 ± 1.24
	Carmon et al. (2019)	88.87 ± 0.48	87.60 ± 0.57
	Hendrycks et al. (2019)	86.57 ± 0.05	83.87 ± 0.42
	Sitawarin et al. (2020)	86.37 ± 1.20	81.90 ± 1.14
	Zhang et al. (2019a)	86.50 ± 0.59	83.03 ± 0.76
	Wong et al. (2020)	78.37 ± 1.40	76.93 ± 1.31
	Ding et al. (2020)	77.63 ± 0.65	76.30 ± 0.29
	Zagoruyko & Komodakis (2016)	$\mathbf{95.93 \pm 0.70}$	$\mathbf{95.93 \pm 0.62}$
CIFAR-100	Wang et al. (2023)	75.33 ± 0.60	68.33 ± 0.24
	Amini et al. (2024)	75.23 ± 0.59	68.20 ± 0.14
	Wang et al. (2023)	73.33 ± 0.74	69.33 ± 0.33
	Cui et al. (2024)	74.30 ± 0.54	71.40 ± 0.08
	Gowal et al. (2020)	69.00 ± 0.43	68.10 ± 0.33
	Rebuffi et al. (2021)	62.67 ± 0.42	61.27 ± 0.29
	Pang et al. (2022)	63.93 ± 0.49	60.17 ± 0.97
	Gowal et al. (2020)	60.00 ± 1.00	51.53 ± 0.60
	Hendrycks et al. (2019)	58.73 ± 0.25	55.77 ± 0.54
	Cui et al. (2021)	69.73 ± 1.16	67.77 ± 0.90
	Chen et al. (2022)	62.63 ± 0.05	57.93 ± 0.74
	Sitawarin et al. (2020)	63.67 ± 0.63	57.10 ± 0.45
	Rice et al. (2020)	54.97 ± 1.10	51.50 ± 0.62
	Zagoruyko & Komodakis (2016)	$\mathbf{79.07 \pm 1.79}$	$\mathbf{79.03 \pm 1.67}$

Table 5: Clean classification accuracy and the proportion of evaluation samples with $m(x) > 0$ for all evaluated models on CIFAR-10 and CIFAR-100. Here, $m(x) > 0$ indicates that the feature representation of a sample lies within the semantic region of its true class. Results are reported as the mean $\pm$ standard deviation over three experimental runs.

Dataset	Model	$m(x) > 0$ (%)		p-value
		Original	Random-Label	
CIFAR-10	Mean Sparse	89.23 ± 0.82	9.84 ± 0.30	0.002
	Using Pre-training	84.57 ± 0.61	9.96 ± 0.22	0.002
	Standard WRN-94-16	96.20 ± 0.50	9.79 ± 0.17	0.002
CIFAR-100	Mean Sparse	68.13 ± 0.87	1.02 ± 0.05	0.002
	Using Pre-training	56.13 ± 0.58	1.01 ± 0.04	0.002
	Standard WRN-94-16	79.43 ± 0.86	1.01 ± 0.02	0.002

Table 6: Random-label baseline results on CIFAR-10 and CIFAR-100. Here, $m(x) > 0$ indicates that the feature representation of a sample lies within the semantic region of its true class. Original denotes the semantic-region result obtained using the original ground-truth class labels, while Random-Label denotes the corresponding result after randomly permuting the reference labels used to construct the class centroids. Results are reported as the mean $\pm$ standard deviation across three independent runs.

Dataset	Model	L_2 Distance	Cosine Similarity
CIFAR-10	Mean Sparse	0.239 ± 0.028	0.9997 ± 0.0001
	Using Pre-training	0.614 ± 0.038	0.9992 ± 0.0001
	Standard WRN-94-16	0.061 ± 0.004	1.0000 ± 0.0000
CIFAR-100	Mean Sparse	1.676 ± 0.020	0.9920 ± 0.0002
	Using Pre-training	2.168 ± 0.039	0.9923 ± 0.0004
	Standard WRN-94-16	0.562 ± 0.006	0.9987 ± 0.0001

Table 7: Centroid stability results for the three representative models on CIFAR-10 and CIFAR-100. Stability is evaluated by comparing the class centroids constructed from two independently sampled class-balanced subsets of the clean training data. The L_2 distance measures the Euclidean displacement between the corresponding class centroids, while the cosine similarity measures the similarity of their directions in the feature space. Results are reported as the mean $\pm$ standard deviation across three independent runs.

Model	Condition Satisfied (%)	Region Survival (%)	Condition Violation (%)
Amini et al. (2024)	57.03 ± 1.93	$\mathbf{89.51 \pm 0.41}$	0.00 ± 0.00
Bartoldson et al. (2024)	54.51 ± 1.96	87.87 ± 0.59	0.00 ± 0.00
Wang et al. (2023)	45.39 ± 0.99	86.90 ± 0.32	0.00 ± 0.00
Cui et al. (2024)	33.31 ± 0.64	85.58 ± 0.76	0.00 ± 0.00
Gowal et al. (2020)	33.41 ± 0.25	85.43 ± 0.96	0.00 ± 0.00
Xu et al. (2023)	$\mathbf{69.46 \pm 1.30}$	74.20 ± 1.08	0.00 ± 0.00
Rade & Moosavi-Dezfooli (2021)	17.80 ± 0.50	87.74 ± 1.67	0.00 ± 0.00
Carmon et al. (2019)	26.26 ± 0.90	87.10 ± 0.17	0.00 ± 0.00
Hendrycks et al. (2019)	5.25 ± 0.43	83.42 ± 0.65	0.00 ± 0.00
Sitawarin et al. (2020)	11.44 ± 0.67	85.08 ± 1.68	0.00 ± 0.00
Zhang et al. (2019a)	2.53 ± 0.37	82.79 ± 1.44	0.00 ± 0.00
Wong et al. (2020)	10.11 ± 1.09	73.17 ± 0.52	0.00 ± 0.00
Ding et al. (2020)	4.11 ± 0.79	66.19 ± 1.30	0.00 ± 0.00
Standard	0.00 ± 0.00	0.42 ± 0.15	0.00 ± 0.00

Table 8: Empirical validation of the sufficient condition for semantic-region preservation for all evaluated models on CIFAR-10. Condition Satisfied denotes the proportion of samples with $m(x) > 0$ for which the empirical feature displacement satisfies $\widehat{D}(x, \epsilon) < m(x)/2$. Region Survival denotes the proportion of all samples with $m(x) > 0$ whose feature representations remain within the semantic region of their true class under the feature-space PGD attack. Condition Violation denotes the proportion of samples satisfying the empirical sufficient condition that nevertheless leave their true-class semantic region under the attack. Results are reported as the mean $\pm$ standard deviation across three experimental runs.

Model	Condition Satisfied (%)	Region Survival (%)	Condition Violation (%)
Wang et al. (2023)	14.00 ± 0.50	84.34 ± 0.48	0.00 ± 0.00
Amini et al. (2024)	15.25 ± 0.55	86.66 ± 0.62	0.00 ± 0.00
Wang et al. (2023)	7.31 ± 0.30	$\mathbf{87.84 \pm 0.45}$	0.00 ± 0.00
Cui et al. (2024)	7.94 ± 0.29	84.69 ± 1.06	0.00 ± 0.00
Gowal et al. (2020)	10.57 ± 0.85	79.05 ± 1.23	0.00 ± 0.00
Rebuffi et al. (2021)	8.92 ± 0.32	84.06 ± 1.51	0.00 ± 0.00
Pang et al. (2022)	3.10 ± 0.43	85.88 ± 0.92	0.00 ± 0.00
Gowal et al. (2020)	1.17 ± 0.32	84.69 ± 1.87	0.00 ± 0.00
Hendrycks et al. (2019)	7.24 ± 0.58	70.22 ± 1.80	0.00 ± 0.00
Cui et al. (2021)	$\mathbf{20.97 \pm 0.97}$	61.71 ± 2.36	0.00 ± 0.00
Chen et al. (2022)	1.90 ± 0.41	80.56 ± 1.29	0.00 ± 0.00
Sitawarin et al. (2020)	1.17 ± 0.54	79.67 ± 1.79	0.00 ± 0.00
Rice et al. (2020)	0.52 ± 0.34	81.23 ± 0.49	0.00 ± 0.00
Standard	0.00 ± 0.00	0.04 ± 0.06	0.00 ± 0.00

Table 9: Empirical validation of the sufficient condition for semantic-region preservation for all evaluated models on CIFAR-100. Condition Satisfied denotes the proportion of samples with $m(x) > 0$ for which the empirical feature displacement satisfies $\widehat{D}(x, \epsilon) < m(x)/2$. Region Survival denotes the proportion of all samples with $m(x) > 0$ whose feature representations remain within the semantic region of their true class under the feature-space PGD attack. Condition Violation denotes the proportion of samples satisfying the empirical sufficient condition that nevertheless leave their true-class semantic region under the attack. Results are reported as the mean $\pm$ standard deviation across three experimental runs.