Supplementary Material for *Adversarial Robustness of Discriminative Self-Supervised Learning in Vision*

Ömer Veysel Çağatan[1,2] [1]Ömer Faruk Tal [1]M. Emre Gürsoy
[1]Department of Computer Engineering, Koç University
[2]KUIS AI Center, Koç University
ocagatan19@ku.edu.tr, otal19@ku.edu.tr, emregursoy@ku.edu.tr

Abstract

This supplementary material provides a comprehensive examination of adversarial attacks on self-supervised learning (SSL) models for computer vision. The document details various adversarial attack methodologies, including instance-specific attacks (categorized into white-box and black-box approaches) and universal adversarial perturbations (both data-dependent and data-independent). It contains precise algorithmic descriptions of gradient-based methods like FGSM and PGD, optimization-based approaches, and specialized techniques for generating universal perturbations. The supplement presents extensive experimental results comparing the adversarial robustness of multiple SSL models across ImageNet classification, transfer learning, segmentation, and detection tasks. Additional visualizations include t-SNE plots, inter/intra-class distance analyses, and regression plots examining the relationship between direction ratio, magnitude ratio, and performance drop under attack. The document concludes with comprehensive result tables from experiments with different model architectures (ResNet vs. ViT), training durations, and SSL method variations, providing a thorough empirical foundation for the findings presented in the main paper.

1. Adversarial Attacks

1.1. Instance Adversarial Attacks

Instance adversarial methods, or per-instance generation, involve crafting distinct perturbations for each individual image within the dataset on which the model has been trained or fine-tuned. The generation of these perturbations relies on various techniques, which are determined by the specific goals of the attack, the level of access granted to the model—such as full access to model weights, predictions alone, or prediction scores (logits)—and the distance metrics employed. While multiple classification schemes for adversarial attacks exist, we adopt the widely accepted taxonomy for clarity and consistency.

White-box attacks, in this context, presume complete access to the model, including its architecture and parameters. The primary approach utilizes the gradients derived from the loss function to generate adversarial perturbations. These perturbations are then applied to the image within the constraints of specific distance metrics, such as l_0, l_1, l_2, or l_∞. Specifically, l_0 measures the number of altered pixels, l_1 quantifies the total absolute difference between images, l_2 computes the Euclidean distance, and l_∞ captures the magnitude of the largest perturbation applied to any pixel.

Gradient-based methods exploit the gradient of the neural network's loss function with respect to the input data, strategically altering the input to increase the loss and induce misclassification. The foundational work in this domain is attributed to the Fast Gradient Sign Method (FGSM) [10], which represents the first successful application of gradient-based adversarial perturbations. The Fast Gradient Sign Method (FGSM) computes the gradient of the cross-entropy loss with respect to the input image to determine the perturbation direction that maximizes the loss. The adversarial example is then generated by applying this perturbation in a single step. Due to this one-step update, FGSM is classified as a single-step adversarial attack.

Other FGSM methods apply gradeient update more than one time (iteratively) using much smaller step sizes to remain in predetermined lp ball. PI-FGSM modify the gradient update rule by focusing on patch-based rather than pixel-wise perturbations [8, 9]. DI-FGSM employs random padding and resizing operations to enhance data input for auxiliary models [25]. TAP also tries to increase cross-model transferability by introducing distance maximization between intermediate

feature maps of the adversarial and benign datapoints. It also regularize the images to reduce high frequency perturbations as they claim Convolution may act as a smoother, and it will increase the black-box transferability performance of perturbation [27]. TI-FGSM is also another iterative FGSM method, which uses translated version of benign input to enhance black-box transferability of adversarial attacks to models which are defended with various methods. TIFGSM suggest that these defended models uses different discriminative regions than the model on which adversarial examples are generated, which makes these adversarial examples less effective. TIFGSM uses diversified (shifted and padded) inputs which are obtained by approximating the gradients with convolution kernels [7].

On the other hand, Projected Gradient Descent (PGD) employs an iterative approach, projecting updates back onto the l_∞ ball of the original data point to generate adversarial perturbations [14]. The key distinction between PGD and FGSM variants, lies in the fact that PGD treats each iteration as a solution to the same optimization problem. PGD ensures that each iterative step remains within the neighborhood of the original data point, while iterative FGSM methods use the newly generated steps to continue further processing.

As with FGSM-based adversarial attacks, several improvements have been made to PGD to address specific needs [14]. For example, PGD-l_2 incorporates the l_2 norm instead of the l_∞ norm to better fool target models [14].

The Jitter attack introduces a novel objective function for adversarial perturbation generation, departing from the conventional Cross-Entropy objective. The study suggests that many adversarial attacks predominantly fool a limited set of classes rather than broadly deceiving the entire model. The proposed objective seeks to enhance the fooling rate across a wider range of classes, aiming for more generalized misclassification [21].

Improving the transferability of per-instance attacks can, however, lead to reduced effectiveness against auxiliary models, and vice-versa [8, 23]. Therefore, various strategies have been proposed to optimize attack performance based on the level of access to the target model.

In contrast, optimization-based attacks approach the generation of adversarial examples as an optimization problem, where a specific objective is minimized subject to given constraints. While gradient-based methods update images directly using gradient information and typically rely on the l_∞ norm as a boundary, optimization-based methods employ a more formal problem definition that allows for the use of advanced optimization techniques. Consequently, different lp norm is utilized in these methods alongside with l_∞ norms.

[3] constructs a minimization problem—focusing on minimizing the distance between adversarial examples and the original data points across several l norms—to develop the CW attack, one of the most prominent adversarial attack methods.

The EADEN attacks adopt a similar approach to the CW attack but introduce a modification to the loss function by incorporating an additional l_1 distance term in the minimization problem. The l_1 distance, which measures the total variation of the perturbation, promotes sparsity in the adversarial perturbation. While sparsity is not widely employed in adversarial example generation, it is commonly used in image denoising and restoration techniques. These methods utilize the Iterative Shrinkage-Thresholding Algorithm (ISTA) to solve the corresponding optimization problem [5].

While white-box attacks exploit full access to the model, this is often not a realistic scenario. In many cases, model weights are not shared, or gradient information is unavailable. Although efforts have been made to enhance cross-model transferability, as discussed previously, there are also specific attack schemes designed to target models in black-box settings. For example, the Square Attack leverages random search combined with model scores—probability distributions over class predictions—to generate perturbations. In essence, the algorithm makes random modifications to the input data and retains changes that yield progress toward the objective function [1].

These black-box attacks leverages gradient-free approaches remain relatively underexplored. For instance, the Simultaneous Perturbation Stochastic Approximation (SPSA) method estimates gradients by perturbing the input in random directions, enabling the approximation of gradients for objectives that cannot be differentiated analytically. This approach offers deeper insights into the model's behavior, with the paper also claiming that the stochastic perturbations introduced by sampling allow algorithms to converge toward a global minimum [24].

Among black-box attacks, some methods focus on l_0 norm-based perturbations. Pixle, for instance, is a black-box attack that utilizes random search and the l_0 norm, altering a small number of pixels to generate adversarial examples [20]. On a more constrained scale, the OnePixel attack modifies only a single pixel, maintaining an l_0 norm of 1, and despite its simplicity, it is capable of fooling models to some extent. However, it is less effective than other methods due to its significant restrictions. This raises important questions about our understanding of Deep Neural Networks and their vulnerability to minimal perturbations [22].

1.2. Universal Adversarial Perturbations

The Universal Adversary (UAP) represents a singular perturbation crafted for an entire image dataset. The rationale behind UAP is to identify a perturbation, subject to specified constraints, capable of deceiving the model across a majority of images in the dataset, as initially demonstrated by [15], which utilizes DeepFool to create an average perturbation for the entire dataset. It has been empirically observed that universal adversaries exhibit heightened transferability across diverse models and datasets compared to instance methods. UAP's are important as they are independent from the input - to some extend - they reveal intrinsic chracteristics of models of interest [4, 26].

Two primary techniques are employed for crafting UAPs: (1) generation with generative models, as evidenced by works such as [11, 18], and (2) learning a perturbation designed to disrupt the representations acquired by the models.

UAPs can be further categorized into two classes: data-dependent attacks, which require a comprehensive and general dataset that the attacker seeks to compromise (e.g., ImageNet), and data-independent attacks, which do not rely on any specific dataset.

The first example of UAP, referred to here as UAP-DeepFool (to avoid confusion with the broader class of UAP attacks), utilizes the DeepFool per-instance adversarial attack method which computes perturbations by manipulating the geometry of decision boundaries. UAP-DeepFool iteratively determines the worst-case direction for each data point, and aggregating the results into a universal perturbation - if it is succesfull -, which is then projected onto an l_∞ ball [15]. Following this work, UAPEPGD replaces the DeepFool approach with Projected Gradient Descent (PGD), an optimization-based adversarial attack method, to craft stronger adversarial examples [6].

ASV - to our best knowledge - is the first UAP that does not require label information, relying solely on images to generate UAPs. Adversarial Semantic Vectors (ASVs) represent one of the first UAP methods that do not require label information, relying solely on images to generate UAPs. The study suggests that since adversarial perturbations typically exhibit small magnitudes, perturbations in the non-linear maps computed by deep neural networks (DNNs) can be approximated using the Jacobian matrix [12]. Similarly, the STD (Dispersion Reduction) attack seeks to reduce the "contrast" of the internal feature map by targeting the lower layers of Convolutional Neural Networks (CNNs). These lower layers typically detect simple image features such as edges and textures, which are common across datasets and CNN models. By reducing the contrast (measured as the standard deviation of feature maps), the resulting images become indistinguishable to the model [13].

Self-Supervised Perturbation (SSP) takes a different approach, arguing that adversarial examples generated through gradients using labels fail to capture intrinsic properties of models. SSP aims to maximize "feature distortion," the changes in the network's internal representation caused by adversarial examples compared to the original image, in order to fool subsequent layers in the model [19].

FG-UAP builds upon this by exploiting a phenomenon referred to as "Neural Collapse," where, as noted, different class activations converge to class means, allowing a single common perturbation to fool the model across a wide range of images. This collapse happens primarily in the final layers of the model, and FG-UAP targets these regions to generate effective UAPs [26].

Another label-independent UAP method, L4A, focuses on the success of adversarial perturbations during cross-finetuning. L4A targets the lower layers of models, which remain more stable during finetuning (as they detect simple features), and utilizes the Frobenius norm for optimization, with variants such as L4A-base, L4A-fuse, and L4A-ugs. L4A-base attacks the lowest layer, L4A-fuse attacks lowest 2 layers and L4A-ugs uses samples from a Gaussian distribution where mean and standard deviation is in close range of downstream task [2].

Data-independent UAP methods do not utilize any dataset for adversarial perturbation generation, instead focusing on the intrinsic characteristics of models. Fast Feature Fool (FFF) was the first adversarial attack method that did not use a dataset. It aims to disrupt the features learned at individual CNN layers, proposing that non-discriminative activations can lead to eventual misclassification. FFF over-saturates the learned features at multiple layers, misleading subsequent layers in the network [16]. Following that work GD-UAP, changes the objective a little bit and add other variations such as "mean-std" and "sampled" versions to improve perturbation performance. The "mean-std" variant uses the mean and standard deviation of the test dataset to better align perturbations with dataset characteristics to prevent perturbation dataset mismatch, while the "sampled" version employs a small sample from the dataset to capture its statistics and semantics [17]. In our work, we have also integrated "mean-std" and "one-sample" versions of GD-UAP to FFF, since they are highlt similar as GD-UAP is a follow-up work FFF. PD-UAP, another data-independent method, focuses on predictive uncertainty rather than any specific image data, aligning perturbations with task-specific objectives [16].

To accommodate both Vision Transformers (ViTs) and ResNets, we have adapted some of these attacks, originally designed for CNNs, to work with ViTs. For low-level layer attacks, we applied them to the first few blocks of the ViT model, following methods like SSP and L4A. For FFF, which typically uses mean of ReLU activations and a logarithmic operation,

we modified the procedure to suit ViTs, which employ GeLU activations (capable of taking values below zero), by applying an absolute value operator between the mean and logarithmic functions. In conducting these experiments, we strove to maintain fair comparisons and minimized the introduction of tweaks to the original methodologies.

1.3. FGSM and PGD versions

Attack Version	Attack Type	ε	Step Count	Norm
$FGSM_1$	FGSM	0.25	-	∞
$FGSM_2$	FGSM	1	-	∞
PGD_1	PGD	0.25	20	∞
PGD_2	PGD	1	20	∞
PGD_3	PGD	0.25	40	∞
PGD_4	PGD	1	40	∞
PGD_5	PGD	0.5	40	$\|\cdot\|_2$

Table 1. Hyperparameters of the different FGSM and PGD attacks that we use in ImageNet and transfer learning.

1.4. Categories

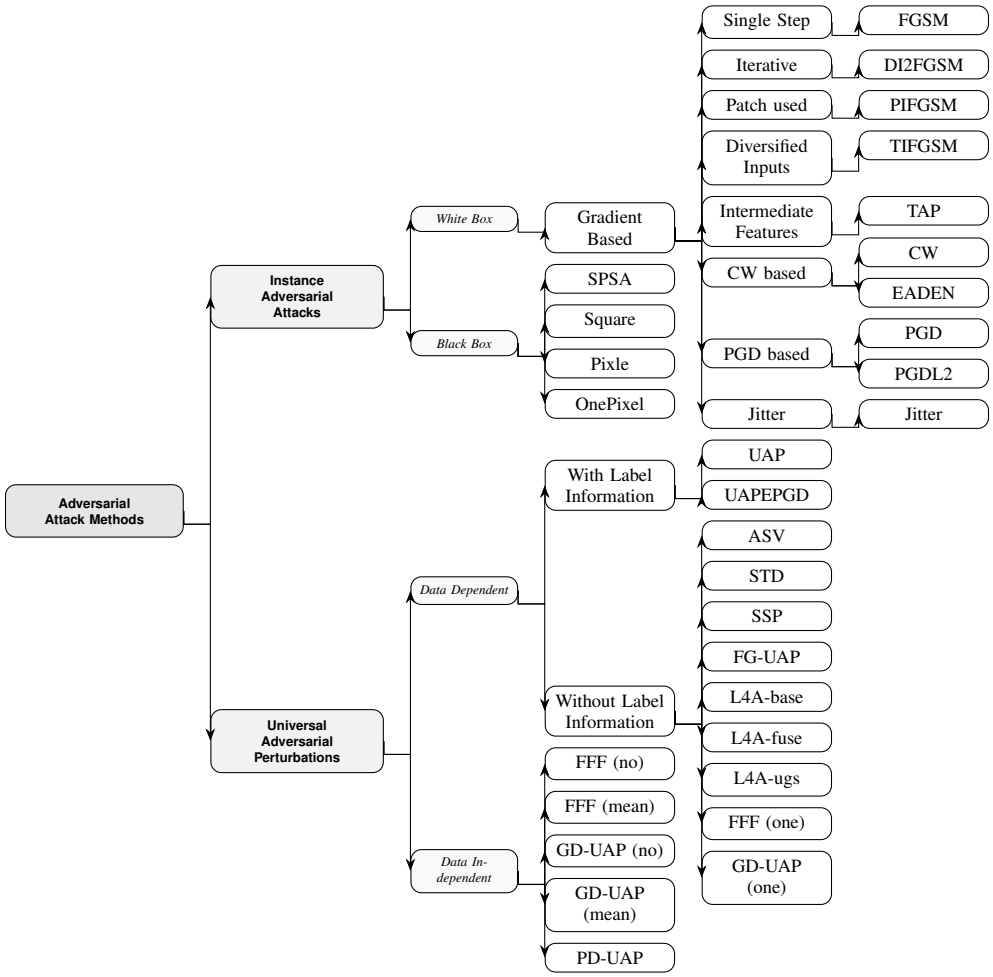


Figure 1. Classification of Selected Adversarial Attack Methods

2. DINOv2 and MAE

The primary challenge in comparing DINOv2 to other SSL models lies in its training data. DINOv2 is trained on a significantly larger dataset containing over 140 million images, which introduces a major discrepancy in scale compared to other models. Additionally, only the largest DINOv2 model is trained in a fully self-supervised manner; the smaller variants are obtained via distillation from this larger model. As a result, they are not inherently self-supervised and tend to perform worse than the original, potentially due to these non-uniform training procedures.

In contrast, MAE is entirely trained on ImageNet, but its performance under linear probing is relatively weak. To mitigate this, the authors apply partial fine-tuning to improve results. This complicates direct comparisons with models evaluated purely via linear probing, though MAE still achieves roughly average performance overall.

Given the diversity of training and evaluation setups, it is difficult to make definitive comparisons between these models. Nevertheless, we include them in our analysis to ensure broader coverage. The complete results can be found in Section 6.5.

3. Checkpoints and Repositories

The self-supervised learning checkpoints used in our experiments are from the following repositories: Barlow Twins, BYOL, DINO, DINOv2, MoCo v3, SimCLR, SwAV, VICReg, and MAE.

We have used Torchattacks library for IAAs and the repository of Ban and Dong [2] for UAPs

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer, 2020.

[2] Yuanhao Ban and Yinpeng Dong. Pre-trained adversarial perturbations. *Advances in Neural Information Processing Systems*, 35: 1196–1209, 2022.

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

[4] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat K Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020.

[5] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[6] Yingpeng Deng and Lina J Karam. Universal adversarial attack via enhanced projected gradient descent. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1241–1245. IEEE, 2020.

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[8] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 307–322. Springer, 2020.

[9] Lianli Gao, Qilong Zhang, Jingkuan Song, and Heng Tao Shen. Patch-wise++ perturbation for adversarial targeted attacks. *arXiv preprint arXiv:2012.15503*, 2020.

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

[11] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 43–49. IEEE, 2018.

[12] Valentin Khrulkov and Ivan Oseledets. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8562–8570, 2018.

[13] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 940–949, 2020.

[14] 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.

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

[16] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. *arXiv preprint arXiv:1707.05572*, 2017.

[17] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2452–2465, 2018.

[18] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R Venkatesh Babu. Nag: Network for adversary generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 742–751, 2018.

[19] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020.

[20] Jary Pomponi, Simone Scardapane, and Aurelio Uncini. Pixle: a fast and effective black-box attack based on rearranging pixels. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2022.

[21] Leo Schwinn, René Raab, An Nguyen, Dario Zanca, and Bjoern Eskofier. Exploring misclassifications of robust neural networks to enhance adversarial attacks. *Applied Intelligence*, 53(17):19843–19859, 2023.

[22] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[23] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[24] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International conference on machine learning*, pages 5025–5034. PMLR, 2018.

[25] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.

[26] Zhixing Ye, Xinwen Cheng, and Xiaolin Huang. Fg-uap: Feature-gathering universal adversarial perturbation. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023.

[27] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.

4. Regression Visualization

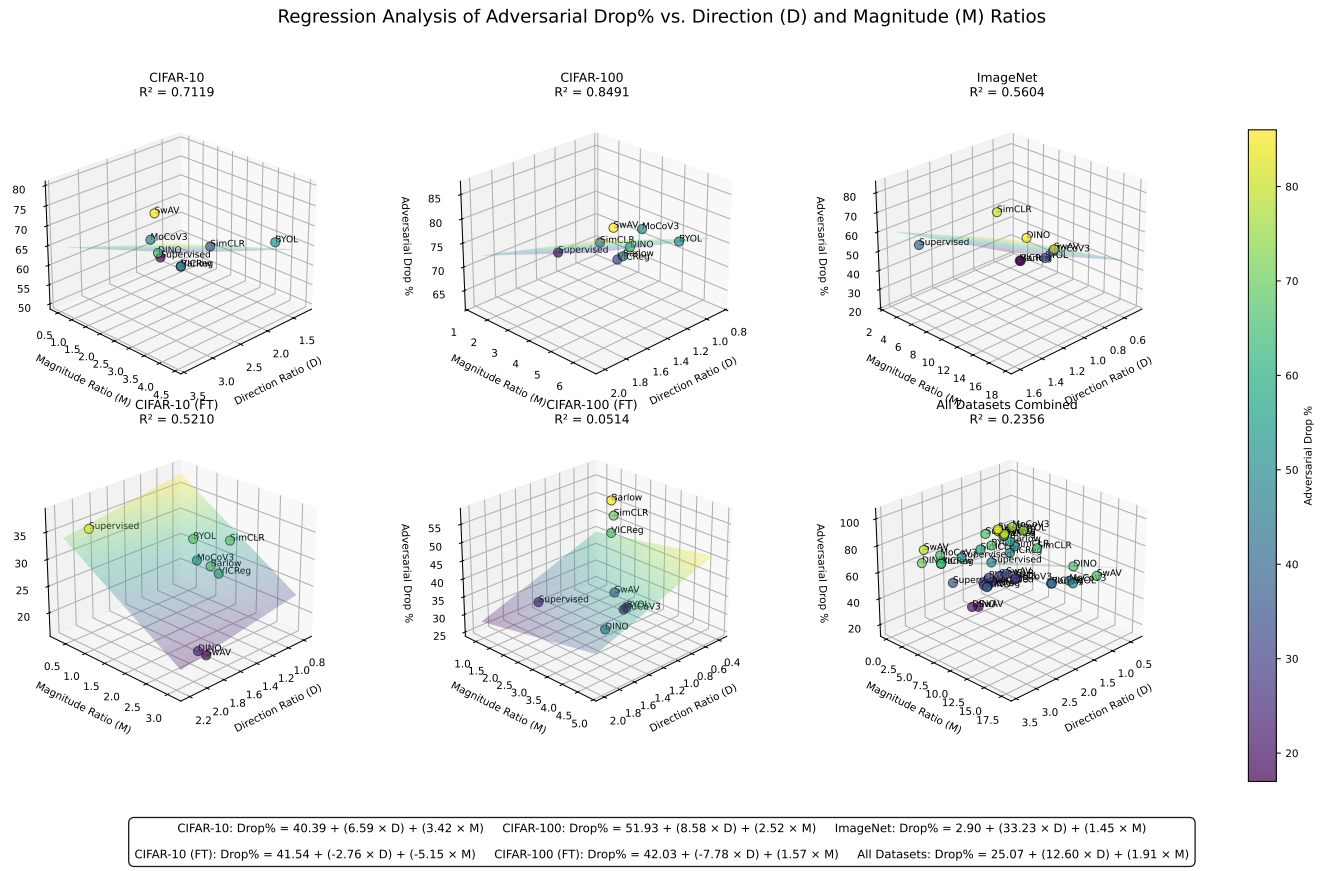
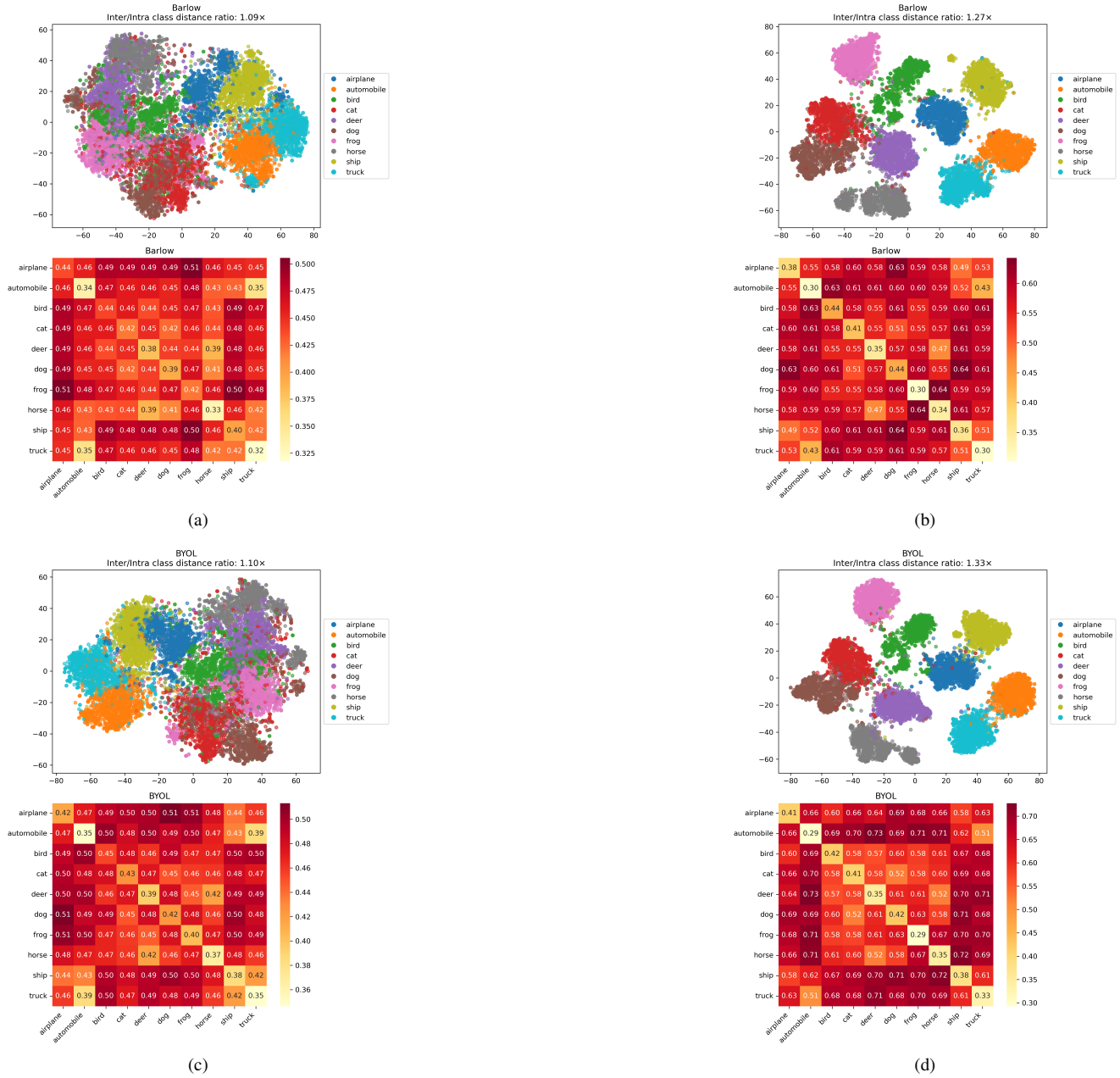


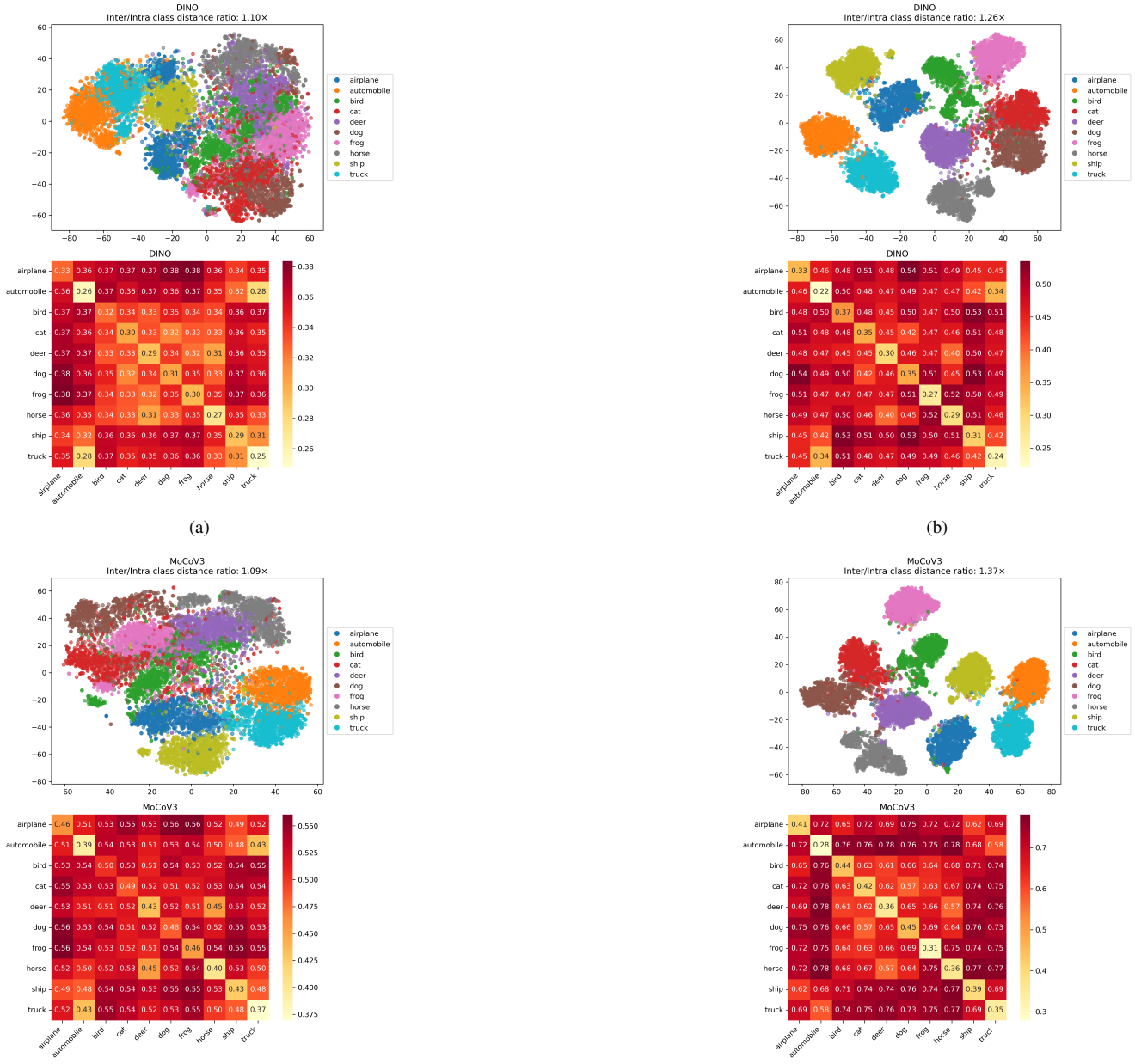
Figure 2. Regression analysis of adversarial performance drop percentage as a function of Direction Ratio (D) and Magnitude Ratio (M) across different datasets. The 3D plots show the fitted regression planes for CIFAR-10, CIFAR-100, and ImageNet datasets (top row), their fine-tuned counterparts (bottom left and center), and the combined analysis (bottom right). Each data point represents a different self-supervised learning method, color-coded by drop percentage. The R^2 values highlight the strong explanatory power for non-fine-tuned datasets (0.56-0.85) compared to fine-tuned ones (0.05-0.52). Note the positive coefficients for D and M in non-fine-tuned scenarios versus negative coefficients in fine-tuned contexts, suggesting fundamentally different robustness mechanisms. Regression equations are displayed below each corresponding plot.

5. t-SNE and Inter/Intra Class Distance Visualization



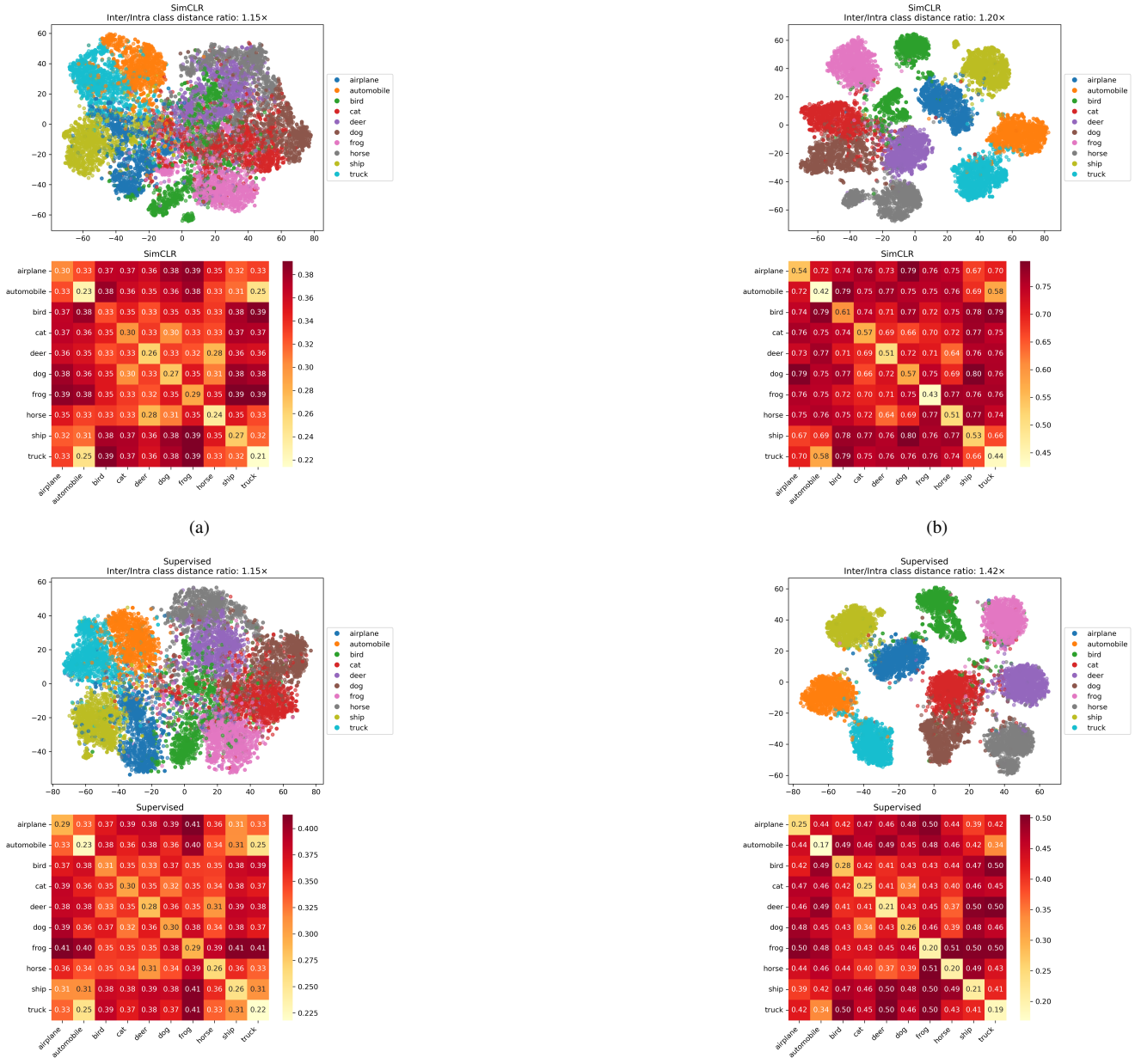
(a) (b) (c) (d)

Figure 3. Comparison of feature representations for CIFAR-10 images using ResNet50 with different self-supervised learning (SSL) methods. **Top row:** Results from Barlow. **Bottom row:** Results from BYOL. For each SSL method, **left** panels show results from **probed** models and **right** panels show results from **fine-tuned** models. Within each panel, the **upper** plots display t-SNE visualizations of the 2048-dimensional feature vectors using Euclidean distance, with points colored by class and Inter/Intra class distance ratios indicated. The **lower** plots show the corresponding class-wise distance matrices computed using cosine similarity, with the average distances between samples from each pair of classes. Higher Inter/Intra class distance ratios indicate better class separation in the feature space.



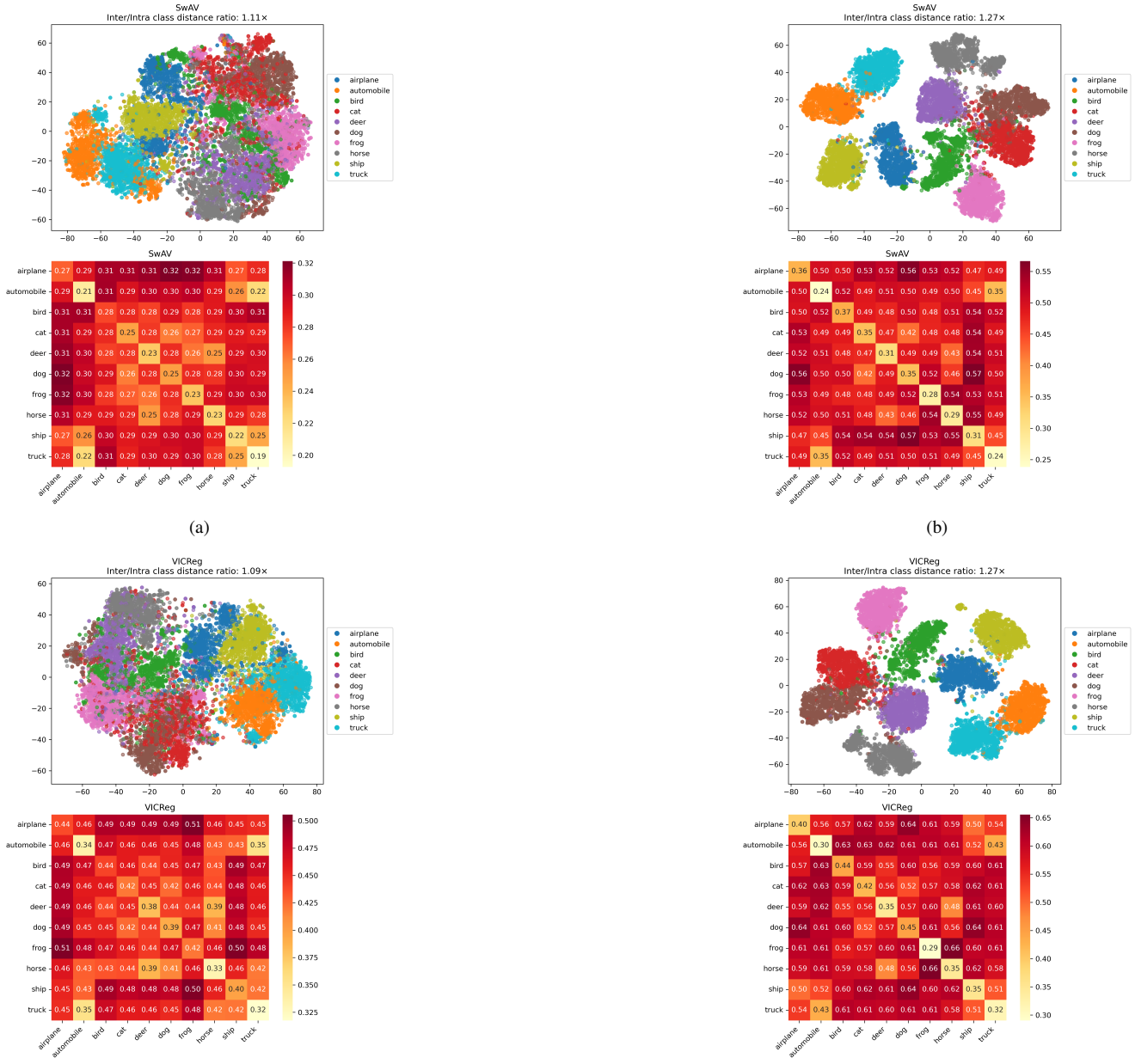
(a) (b) (c) (d)

Figure 4. Comparison of feature representations for CIFAR-10 images using ResNet50 with different self-supervised learning (SSL) methods. **Top row:** Results from DINO. **Bottom row:** Results from MoCoV3. For each SSL method, **left** panels show results from **probed** models and **right** panels show results from **fine-tuned** models. Within each panel, the **upper** plots display t-SNE visualizations of the 2048-dimensional feature vectors using Euclidean distance, with points colored by class and Inter/Intra class distance ratios indicated. The **lower** plots show the corresponding class-wise distance matrices computed using cosine similarity, with the average distances between samples from each pair of classes. Higher Inter/Intra class distance ratios indicate better class separation in the feature space.



(a) (b) (c) (d)

Figure 5. Comparison of feature representations for CIFAR-10 images using ResNet50 with different self-supervised learning (SSL) methods. **Top row:** Results from SimCLR. **Bottom row:** Results from Supervised. For each SSL method, **left** panels show results from **probed** models and **right** panels show results from **fine-tuned** models. Within each panel, the **upper** plots display t-SNE visualizations of the 2048-dimensional feature vectors using Euclidean distance, with points colored by class and Inter/Intra class distance ratios indicated. The **lower** plots show the corresponding class-wise distance matrices computed using cosine similarity, with the average distances between samples from each pair of classes. Higher Inter/Intra class distance ratios indicate better class separation in the feature space.



(a) (b) (c) (d)

Figure 6. Comparison of feature representations for CIFAR-10 images using ResNet50 with different self-supervised learning (SSL) methods. **Top row:** Results from SwAV. **Bottom row:** Results from VICReg. For each SSL method, **left** panels show results from **probed** models and **right** panels show results from **fine-tuned** models. Within each panel, the **upper** plots display t-SNE visualizations of the 2048-dimensional feature vectors using Euclidean distance, with points colored by class and Inter/Intra class distance ratios indicated. The **lower** plots show the corresponding class-wise distance matrices computed using cosine similarity, with the average distances between samples from each pair of classes. Higher Inter/Intra class distance ratios indicate better class separation in the feature space.

6. Full Results

6.1. ImageNet

Table 2. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Imagenet-1k dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	42.41	39.41	24.68	42.67	24.29	38.83	24.71	42.42
$FGSM_2$	18.11	13.47	5.66	15.53	8.84	12.18	6.35	18.11
PGD_1	42.38	39.63	25.65	42.39	26.6	35.26	26.48	42.41
PGD_2	1.48	0.65	0.18	1.06	0.25	0.37	0.18	1.5
PGD_3	42.6	39.82	25.85	42.56	26.79	35.39	26.73	42.6
PGD_4	1.19	0.5	0.14	0.82	0.2	0.28	0.14	1.2
PGD_5	5.18	3.44	0.67	4.79	0.9	1.9	0.69	5.15
DIFGSM	52	52.71	41.12	54.09	42.57	51.43	45.65	52.49
CW	0.18	0.02	0	0.02	0.02	0.02	0	0.19
Jitter	59.83	61.92	60.26	62.47	56.4	62.75	61.16	59.84
TIFGSM	61.04	62.27	56.98	61.47	55.63	62.16	60.07	59.91
PIFGSM	34.38	29.83	14.54	34.1	13.34	28.64	14.12	34.43
EADEN	0	0	0	0	0	0	0	0
OnePixel	69.34	72.5	72.83	72.64	66.47	73.27	72.73	69.38
Pixle	25.22	28.67	19.41	31.45	21.75	23.21	16.95	25.23
SPSA	66.59	69.59	68.11	69.93	63.01	69.48	68.61	66.63
Square	4.44	2.62	1.3	3.15	4.22	0.87	1.99	4.49
TAP	70.31	74.36	73.78	73.72	68.1	68.98	75.05	70.33
ASV	62.17	68.46	68.69	67.59	61.84	64.28	67.66	62.51
FFF (mean-std)	53.72	62.38	59.45	46.78	56.79	53.45	61.14	54.76
FFF (no-data)	39.64	31.94	40.73	57.39	53.52	42.49	32.15	38.68
FFF (one-sample)	30.53	54.41	16.15	57.18	28.63	40.57	29.96	30.14
FG-UAP	4.34	1.83	2.3	3.08	1.03	1.89	2.17	1.95
GD-UAP (mean-std)	56.92	61.1	49.85	53.59	55.51	59.6	57.38	58.07
GD-UAP (no-data)	33.41	37.96	26.59	43.66	37.05	36.01	36.07	31.29
GD-UAP (one-sample)	42.17	37.54	25.9	35.1	14.11	25.27	36.37	40.08
L4A-base	42.17	36.35	51.13	41.09	7.83	26.73	10.9	28.36
L4A-fuse	28.33	36.01	50.69	40.39	8.34	26.01	10.94	28.55
L4A-ugs	27.56	60.03	56.83	65.27	53.08	34.27	49.91	58.23
PD-UAP	65.06	54.06	45.06	56.248	60.96	46.79	51.75	65.07
SSP	27.7	41.05	38.55	48.56	35.06	9.43	23.27	33.26
STD	49.19	58.08	52.44	53.44	50.3	56.16	56.08	48.88
UAP (DeepFool)	13.23	20.25	15.18	24.25	23.72	13.67	8.59	13.22
UAPEPGD	67.74	67.21	70.6	71.57	64.36	71.26	70.98	67.88
Clean Accuracy	71.20	74.57	75.28	74.57	68.90	76.13	75.27	71.26
IAA Avg.	33.14 ↓54%	32.86↓56%	27.28↓64%	34.04↓54%	26.63↓61%	31.39↓59%	27.87↓63%	33.12↓54%
UAP Avg.	40.24 ↓43%	45.79↓39%	41.88↓44%	47.82↓36%	38.25↓44%	37.99↓50%	37.83↓50%	41.30↓42%
Adv Avg.	36.48 ↓49%	38.94↓48%	34.16↓55%	40.53↓46%	32.10↓53%	34.50↓55%	32.56↓57%	36.98↓48%

6.2. Segmentation

6.2.1. Pascal VOC 2012

Metric	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
				Alma				
mIOU (↑)	0.35	0.33	0.34	0.4	0.31	0.26	0.38	0.39
APSR (↓)	99.02	99.01	99.02	98.91	99	99.01	99.01	98.99
				Asma				
mIOU (↑)	49.4	63.39	61.36	61.57	32.06	77.3	62.12	50.38
APSR (↓)	15.39	10.95	11.38	12.18	22.78	5.29	11.56	14.48
				DAG				
mIOU (↑)	0.02	0.02	0.02	0.02	0.03	0.05	0.02	0.02
APSR (↓)	99.87	99.91	99.89	99.88	99.83	99.74	99.89	99.89
				DDN				
mIOU (↑)	5.62	4.64	5.11	7.16	1.67	1.52	6.91	4.94
APSR (↓)	89.66	92.6	92.75	88.01	97.24	88.56	90.77	87.23
				FGSM				
mIOU (↑)	30.35	29.28	30.41	29.43	32.15	38.31	29.4	29.84
APSR (↓)	35.91	45.62	39.66	41.71	33.55	21.36	42.94	39.31
				FMN				
mIOU (↑)	5.4	5.29	4.86	5.19	5.07	2.74	4.9	6.2
APSR (↓)	91.18	92.25	91.02	91.42	89.88	93.53	91.94	89.99
				PGD				
mIOU (↑)	12.67	13.16	12.75	13.06	12.88	10.92	12.98	13.04
APSR (↓)	70.07	82	77	79.27	71.15	67.4	77.31	72.43
Clean mIOU (↑)	72.63	70.37	71.65	71.25	71.96	77.35	70.8	70.33
Clean APSR (↓)	7.18	8.29	7.64	7.83	7.2	5.27	8.21	8.01
Adversarial mIOU (↑)	14.83↓80%	16.59↓78%	16.41↓77%	16.69↓77%	12.02↓83%	18.73↓76%	16.67↓77%	14.97↓79%
Adversarial APSR (↓)	71.59↑64%	74.62 ↑66%	72.96 ↑65%	73.05↑65%	73.35↑66%	67.84↑64%	73.35↑65%	71.76↑64%

Table 3. Performance metrics (mIOU and APSR) for various self-supervised and supervised models under different adversarial attacks, using unfrozen backbones. Clean and adversarial scores are reported, with percentage changes in adversarial performance noted. Higher mIOU and lower APSR indicate better results

Metric	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
				Alma				
mIOU (↑)	0.39	0.31	0.37	0.37	0.55	0.28	0.35	0.41
APSR (↓)	99.02	99.02	99.02	99.02	98.45	99.01	99.02	99.02
				Asma				
mIOU (↑)	76.06	72.84	75.32	72.84	70.42	69.84	74.09	76.74
APSR (↓)	6.01	7.23	6.14	7.58	5.98	8.18	6.75	5.98
				DAG				
mIOU (↑)	0.03	0.04	0.02	0.04	0.04	0.02	0.03	0.03
APSR (↓)	99.90	99.87	99.89	99.87	99.82	99.87	99.88	99.89
				DDN				
mIOU (↑)	10.81	9.76	6.91	10.74	6.62	2.95	8.57	11.12
APSR (↓)	79.62	75.93	82.58	78.71	75.20	87.30	83.48	80.41
				FGSM				
mIOU (↑)	35.16	31.90	30.88	35.18	36.25	27.70	32.37	34.99
APSR (↓)	33.29	33.63	36.12	33.63	27.35	36.99	36.10	33.69
				FMN				
mIOU (↑)	6.63	6.23	6.22	6.42	8.92	4.23	6.48	6.56
APSR (↓)	87.73	87.10	87.12	87.70	81.28	91.30	87.23	87.23
				PGD				
mIOU (↑)	14.13	12.12	12.12	13.25	12.23	10.49	12.31	13.51
APSR (↓)	76.16	75.49	75.49	76.60	73.38	78.37	80.82	77.62
Clean mIOU (↑)	76.90	76.69	77.01	76.19	75.62	74.20	76.54	77.89
Clean APSR (↓)	5.75	5.74	5.38	6.01	5.98	6.35	5.79	5.48
Adversarial mIOU (↑)	20.46↓73%	19.03↓75%	18.83↓76%	19.83↓74%	19.29↓74%	16.50 ↓78%	19.17 ↓75%	20.48↓74%
Adversarial APSR (↓)	68.82↑63%	68.32↑63%	69.48 ↑64%	69.02↑63%	65.92 ↑60%	71.57 ↑65%	70.47↑65%	69.12↑64%

Table 4. Performance metrics (mIOU and APSR) for various self-supervised and supervised models under different adversarial attacks, using frozen backbones. Clean and adversarial scores are reported, with percentage changes in adversarial performance noted. Higher mIOU and lower APSR indicate better results.

6.2.2. CityScapes

Metric	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
IOU (Higher Better)	65.48	62.05	65.87	63.88	58.82	62.57	66.48	65.64
APSR (Lower Better)	6.4	7.05	6.08	6.68	7.35	7.38	6.27	6.29
				Alma				
IOU	3.82	3.45	3.72	2.99	8.43	4.58	5.51	4.08
APSR	91.4	91.89	91.09	93.99	77.65	89.24	86.44	90.41
				Asma				
IOU	50.9	18.62	64.88	51.93	33.41	58.43	63.07	65.38
APSR	17.3	44.51	7.49	13.76	21.71	8.87	8.75	6.73
				DAG				
IOU	0.19	0.27	0.21	0.33	0.27	0.21	0.15	0.24
APSR	99.8	99.74	99.75	99.72	99.57	99.6	99.76	99.79
				DDN				
IOU	1.49	1.99	1.66	1.18	1.12	1.59	0.51	1.36
APSR	84.79	82.15	84.35	82.85	94.06	81	97.32	83.05
				FGSM				
IOU	30.31	26.22	28.75	30.92	30.64	28.19	28.68	20.33
APSR	31.97	39.57	30.72	26.8	24.97	31.38	33.62	31.99
				FMN				
IOU	8.03	9.19	6.04	8.2	16.77	7.81	8.48	8.2
APSR	77.75	72.11	81.38	77.35	58.44	74.76	77.34	75.44
				PGD				
IOU	10	10.71	7.76	12.23	9.94	11.59	11.43	9.62
APSR	71.92	66.84	75.79	72.8	70.34	68.75	71.46	70.16
Average IOU	14.96$\downarrow$77%	10.06$\downarrow$84%	16.15$\downarrow$75%	15.40$\downarrow$76%	14.37$\downarrow$76%	16.06$\downarrow$74%	16.83$\downarrow$75%	15.60$\downarrow$76%
Average APSR	67.85$\uparrow$61%	70.97$\uparrow$64%	67.22$\uparrow$61%	66.75$\uparrow$56%	63.82$\uparrow$57%	64.80$\uparrow$62%	67.81$\uparrow$61%	65.37$\uparrow$60%

Table 5. Updated performance metrics (mIOU and APSR) for various self-supervised and supervised models under different adversarial attacks, using frozen backbones. Clean and adversarial scores are reported, with percentage changes in adversarial performance noted. Higher mIOU and lower APSR indicate better results.

6.3. Detection

6.3.1. INRIA Person

Table 6. Adversarial Attack Results on Detection using unfrozen SSL and Supervised Models as backbones. The table presents performance metrics under clean and adversarial conditions for various attack types (Optim, BIM, MIM, SGD, PGD, Optim-Adam, Optim-Nesterov). The last two rows display clean mean Average Precision (mAP) and the average performance under adversarial attacks, with the percentage decrease in performance highlighted in red

	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
Optim	6.18	1.68	1.77	4.87	1.06	1.54	4.27	2.12
BIM	32.78	26.93	31.82	21.63	12.6	1.75	40.84	23.22
MIM	11.89	26.24	5.2	10.7	3.1	1.94	10.69	7.85
SGD	6.13	2.89	7.59	20.15	2.45	2.4	13.71	2.99
PGD	84.58	78.44	80.97	81.96	80.45	57.76	80.54	77.52
Optim-Adam	6.43	1.49	2.07	7.49	1.31	1.32	4.47	1.99
Optim-Nesterov	2.34	1.58	1.31	5.24	1.43	2.55	4.34	1.42
Clean mAP	89.14	88.98	89.74	89.74	88.16	86.45	88.60	89.45
Adv mAP.	21.48↓76%	19.89↓78%	18.68↓79%	21.72↓76%	14.63↓83%	9.89↓89%	22.69↓72%	16.73↓81%

Table 7. Adversarial Attack Results on Detection using frozen SSL and Supervised Models as backbones. The table presents performance metrics under clean and adversarial conditions for various attack types (Optim, BIM, MIM, SGD, PGD, Optim-Adam, Optim-Nesterov). The last two rows display clean mean Average Precision (mAP) and the average performance under adversarial attacks, with the percentage decrease in performance highlighted in red

	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
Optim	3.98	1.05	2	2.6	0.78	0.56	1.51	0.39
BIM	44.87	32.24	54.93	26.72	6.79	42.8	44.47	10.32
MIM	11.37	3.04	10.32	5.72	6.0	4.73	10.87	2.68
SGD	3.21	1.28	2.95	9.44	1.42	1.02	2.85	1.72
PGD	83.08	80.83	79.65	79.83	74.68	75.29	79.14	81.27
Optim-Adam	4.71	0.76	3.5	2.03	0.75	0.87	3.46	0.67
Optim-Nesterov	1.75	0.64	0.97	2.77	0.94	0.72	1.1	0.64
Clean mAP	88.39	87.44	87.63	87.36	87.67	87.67	86.55	88.43
Adv mAP	21.85↓75%	17.12↓80%	22.05↓75%	18.44↓79%	13.05↓85%	18.00↓79%	20.49↓76%	13.96↓84%

6.3.2. COCO

Table 8. Adversarial Attack Results on Detection using frozen SSL and Supervised Models as backbones. The table presents performance metrics under clean and adversarial conditions for various attack types (Optim, BIM, MIM, SGD, PGD, Optim-Adam, Optim-Nesterov). The last two rows display clean mean Average Precision (mAP) and the average performance under adversarial attacks, with the percentage decrease in performance highlighted in red

	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
Optim	18.6	16.47	11.05	15.43	11.71	9.12	13.8	13.97
BIM	18.6	18.58	21.02	18.23	13.06	16.74	20.03	18.47
MIM	21.07	18.37	16.38	16.47	14.05	13.01	17.55	17.67
SGD	17.62	18.38	19.19	21.7	12.96	9.22	17.41	16.69
PGD	21.05	20.5	21.68	22.22	14.2	19.61	21.44	20.96
Optim-Adam	19.15	16.36	11.12	15.83	12	9.06	13.92	14.09
Optim-Nesterov	15.76	16.62	13.37	15.31	11.67	9.08	13.91	13.44
Clean mAP	37.93	39.49	38.57	40.49	32.17	37.01	38.45	38.55
Adv Avg.	18.84↓50%	17.90↓55%	16.26↓58%	17.88↓56%	12.80↓60%	12.26↓67%	16.87↓56%	16.47↓57%

6.4. ResNet vs ViT

Table 9. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Imagenet-1k dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy

	MoCoV3-ViT-B	DINO-ViT-B	MoCo-ViT	DINO-ViT	DINO-ResNet	MoCoV3-ResNet
$FGSM_1$	46.69	58.36	34.63	51.42	24.68	42.67
$FGSM_2$	13.21	22.17	0.32	0.97	5.66	15.53
PGD_1	45.06	57.88	33.35	50.98	25.65	42.39
PGD_2	1.14	8.58	0.00	0.00	0.18	1.06
PGD_3	45.1	57.89	33.46	50.95	25.85	42.56
PGD_4	1.05	8.36	0.17	3.84	0.14	0.82
PGD_5	6.86	22.58	2.12	13.57	0.67	4.79
DIFGSM	45.13	57.89	51.91	59.81	41.12	54.09
CW	0	0	0	0	0.02	
Jitter	63.52	68.96	58.25	66.30	60.26	62.47
TIFGSM	66.17	68.58	61.84	65.23	56.98	61.47
PIFGSM	38.14	55.06	25.78	47.64	14.54	34.10
EADEN	0	0	0	0	0	0
OnePixel	74.93	76.67	71.28	75.47	72.83	72.64
Pixle	42.85	49.51	34.69	44.08	19.41	31.45
SPSA	71.18	74.35	66.20	72.47	68.11	69.93
Square	1.87	2.78	1.22	1.67	1.30	3.15
TAP	75.66	76.78	72.34	75.60	73.78	73.72
ASV	0.10	0.10	0.10	0.10	68.69	67.59
FFF (mean-std)	76.38	77.45	72.55	76.71	59.45	46.78
FFF (no-data)	76.34	77.45	72.66	76.71	40.73	57.39
FFF (one-sample)	76.34	77.45	72.55	76.71	16.15	57.18
FG-UAP	1.37	6.79	0.72	3.51	2.30	3.08
GD-UAP (no-data)	53.08	43.49	1.16	7.30	26.59	43.66
GD-UAP (mean-std)	56.97	68.29	28.80	64.17	49.85	53.59
GD-UAP (one-sample)	16.00	64.12	2.04	52.85	25.90	35.10
L4A-base	23.14	19.58	0.69	46.28	51.13	41.09
L4A-fuse	24.13	19.06	0.73	46.30	50.69	40.39
L4A-ugs	6.01	21.65	2.35	0.27	56.83	65.27
PD-UAP	74.72	76.81	70.61	74.91	45.06	56.25
SSP	29.27	60.91	2.52	56.53	38.55	48.56
STD	54.52	72.43	23.01	71.66	52.44	53.44
UAP (DeepFool)	2.19	8.50	1.15	8.56	15.18	24.25
UAPEPGD	73.86	76.26	69.34	74.51	70.60	71.57
Clean Accuracy	76.66	77.99	73.21	76.95	75.28	74.57
IAA Avg.	35.47 ↓54%	42.58 ↓45%	30.42↓58%	37.78↓51%	27.29↓64%	34.05 ↓54%
UAP Avg.	40.28 ↓47%	48.15 ↓38%	26.31↓64%	46.07↓40%	41.88↓44%	47.82 ↓36%
Adv Avg.	37.73 ↓51%	45.19 ↓42%	28.49↓61%	41.67↓46%	34.15↓55%	40.53 ↓46%

6.5. DINOv2 and MAE

Table 10. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Imagenet-1k dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy

	DINOv2-S	DINOv2-B	MAE
$FGSM_1$	43.00	52.25	56.72
$FGSM_2$	6.05	14.16	41.61
PGD_1	38.84	48.53	45.13
PGD_2	0.26	0.75	2.73
PGD_3	38.93	48.60	45.33
PGD_4	0.25	0.69	1.92
PGD_5	2.83	4.17 8	3.39
DIFGSM	56.25	60.55	63.10
CW	0	0	0.06
Jitter	63.10	69.52	69.4
TIFGSM	68.83	73.35	72.17
PIFGSM	37.35	45.16	44.42
EADEN	0	0	0
OnePixel	79.48	82.84	82.39
Pixle	47.93	61.67	61.66
SPSA	76.05	80.22	78.56
Square	0.33	1.96	0.12
TAP	80.49	84.05	82.95
ASV	0.10	0.10	0.10
FFF (mean-std)	79.68	83.78	74.19
FFF (no-data)	79.68	83.80	78.05
FFF (one-sample)	79.68	83.78	80.00
FG-UAP	1.29	1.59	1.55
GD-UAP (no-data)	13.57	74.98	78.06
GD-UAP (mean-std)	0.14	35.22	78.89
GD-UAP (one-sample)	3.16	51.71	79.52
L4A-base	51.89	9.56	14.67
L4A-fuse	52.07	8.51	21.18
L4A-ugs	5.37	53.86	0.43
PD-UAP	18.74	83.33	17.66
SSP	0.13	21.93	28.77
STD	68.38	80.17	77.00
UAP (DeepFool)	12.24	20.24	28.99
UAPEPGD	79.51	83.45	82.25
Clean Accuracy	81.33	84.42	83.58
IAA Avg.	35.55 ↓56%	40.47 ↓52%	41.76↓50%
UAP Avg.	32.10 ↓61%	45.65 ↓46%	43.61↓48%
Adv Avg.	34.87 ↓57%	44.25 ↓48%	43.91↓47%

6.6. ImageNet Across Training Epochs

Table 11. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Imagenet-1k dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	MoCoV3-100	MoCoV3-300	MoCoV3-1000
$FGSM_1$	38.87	42.6	42.67
$FGSM_2$	7.94	8.38	15.53
PGD_1	37.89	41.99	42.39
PGD_2	0.49	0.09	1.06
PGD_3	38.06	42.14	42.56
PGD_4	1.75	1.22	0.82
PGD_5	5.4	5.49	4.79
DIFGSM	49.21	52.65	54.09
CW	0.02	0.02	0.02
Jitter	56.45	60.53	62.47
TIFGSM	57.39	61.86	61.47
PIFGSM	31.24	34.41	34.1
EADEN	0	0	0
OnePixel	66.79	70.76	72.64
Pixle	26.27	29.41	31.45
SPSA	64.05	68.02	69.93
Square	2.05	2.01	3.15
TAP	67.85	71.9	73.72
ASV	62.59	66.11	67.59
FFF (mean-std)	50.05	59.76	46.78
FFF (no-data)	32.96	36.39	57.39
FFF (one-sample)	40.81	58.32	57.18
FG-UAP	2.72	5.07	3.08
GD-UAP (mean-std)	49.56	53.45	43.66
GD-UAP (no-data)	37.00	46.68	53.59
GD-UAP (one-sample)	32.04	37.17	35.10
L4A-base	40.09	33.66	41.09
L4A-fuse	40.07	33.70	40.39
L4A-ugs	51.79	64.64	65.27
PD-UAP	40.96	55.29	56.25
SSP	42.24	26.80	48.56
STD	49.03	51.07	53.44
UAP (DeepFool)	24.71	26.99	24.25
UAPEPGD	65.51	69.75	71.57
Clean Accuracy	68.91	72.82	74.57
IAA Avg.	30.65 ↓56%	32.97↓55%	34.05↓54%
UAP Avg.	41.39 ↓40%	45.30↓38%	4782↓36%
Adv Avg.	35.70 ↓48%	38.77↓47%	40.53↓47%

Table 12. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Imagenet-1k dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy

	SwAV-100	SwAV-200	SwAV-400	SwAV-800
$FGSM_1$	18.08	19.99	21.9	24.71
$FGSM_2$	4.01	4.34	5.2	6.35
PGD_1	18.94	21.3	23.7	26.48
PGD_2	0.31	0.17	0.17	0.18
PGD_3	19.08	21.44	23.88	26.73
PGD_4	0.3	0.15	0.14	0.14
PGD_5	0.73	0.59	0.52	0.69
DIFGSM	39.31	42.01	42.31	45.65
CW	0.0	0.0	0.0	0.0
Jitter	56.67	59.15	60.43	61.16
TIFGSM	53.11	55.14	56.44	60.07
PIFGSM	10	10.87	11.76	14.12
EADEN	0	0	0	0
OnePixel	68.73	70.83	71.64	72.73
Pixle	13.21	16.03	18.08	16.95
SPSA	63.94	66.25	67.38	68.61
Square	0.35	0.36	0.5	1.99
TAP	71.79	73.56	74.37	75.05
ASV	64.00	63.90	67.06	67.66
FFF (mean-std)	55.81	58.52	62.60	61.14
FFF (no-data)	35.02	24.83	35.81	32.15
FFF (one-sample)	24.75	24.22	34.59	29.96
FG-UAP	1.86	3.56	2.29	2.17
GD-UAP (mean-std)	54.07	58.14	58.08	57.38
GD-UAP (no-data)	22.09	29.45	26.57	36.07
GD-UAP (one-sample)	21.33	35.80	37.18	36.37
L4A-base	18.63	17.66	33.43	10.90
L4A-fuse	18.58	17.908	34.73	10.94
L4A-ugs	34.02	37.03	50.20	49.91
PD-UAP	41.58	54.83	51.53	51.75
SSP	12.12	25.74	16.642	23.27
STD	45.74	53.71	45.88	56.08
UAP (DeepFool)	10.17	10.09	10.79	8.59
UAPEPGD	67.13	69.01	70.07	70.98
Clean Accuracy	72.02	73.82	74.57	75.27
IAA Avg.	24.36 ↓66%	25.68↓65%	26.58↓64%	27.87↓63%
UAP Avg.	32.93 ↓54%	36.52↓51%	39.84↓47%	37.83↓50%
Adv Avg.	28.40 ↓60%	30.78↓58%	32.82↓56%	32.55↓57%

6.7. Imagenet With Different MoCo Versions

Table 13. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Imagenet-1k dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy

	MoCoV1	MoCoV2	MoCoV3
$FGSM_1$	15.91	22.01	42.67
$FGSM_2$	6.25	5.17	15.53
PGD_1	17.89	24.00	42.39
PGD_2	0.09	0.54	1.06
PGD_3	17.96	24.14	42.56
PGD_4	0.06	0.52	0.82
PGD_5	0.21	1.33	4.79
DIFGSM	34.85	40.39	54.09
CW	0	0	0.02
Jitter	50.04	53.09	62.47
TIFGSM	48.70	49.50	61.47
PIFGSM	8.53	13.20	34.10
EADEN	0	0	0
OnePixel	56.67	64.63	72.64
Pixle	3.10	17.85	31.45
SPSA	50.62	60.57	69.93
Square	0.80	0.42	3.15
TAP	58.55	65.24	73.72
ASV	18.756	60.99	67.59
FFF (mean-std)	30.34	46.83	46.78
FFF (no-data)	19.47	33.01	57.39
FFF (one-sample)	5.40	42.08	57.18
FG-UAP	0.838	3.78	3.08
GD-UAP (mean-std)	5.88	51.28	53.39
GD-UAP (no-data)	18.34	44.34	43.66
GD-UAP (one-sample)	3.50	39.28	35.10
L4A-base	2.15	37.09	41.09
L4A-fuse	2.19	36.79	40.39
L4A-ugs	2.25	30.16	65.27
PD-UAP	33.96	50.06	56.25
SSP	4.59	42.66	48.56
STD	23.09	46.31	53.44
UAP (DeepFool)	4.34	30.64	24.25
UAPEPGD	47.67	63.44	71.57
Clean Accuracy	60.64	67.72	74.57
IAA Avg.	20.56 ↓66%	24.59↓64%	34.05↓54%
UAP Avg.	13.92 ↓77%	41.17↓39%	47.82↓36%
Adv Avg.	17.44 ↓71%	32.40↓52%	40.53↓46%

6.8. BYOL Ablations

Table 14. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Imagenet-1k dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	BYOL-NC	BYOL-CC	BYOL-128	BYOL-512	BYOL
$FGSM_1$	23.45	31.23	35.33	37.51	39.41
$FGSM_2$	10.6	11.36	13.62	13.92	13.47
PGD_1	25.57	30.86	35.32	37.69	39.63
PGD_2	0.68	0.79	1.3	1	0.65
PGD_3	25.83	30.89	35.52	37.69	39.82
PGD_4	0.59	0.66	1.1	0.83	0.5
PGD_5	1.22	1.89	3.51	3.69	3.44
DIFGSM	35.73	45.67	47.83	49.61	52.71
CW	0.02	0.01	0.1	0.02	0.02
Jitter	53	56.81	57.7	60.02	61.92
TIFGSM	40.89	58.33	56.77	60.11	62.27
PIFGSM	12.27	22.2	25.86	28.48	29.83
EADEN	0	0	0	0	0
OnePixel	61.18	66.27	67.31	68.89	72.5
Pixle	27.1	9.84	25	26.13	28.67
SPSA	55.76	63.8	64.24	66.94	69.59
Square	0.28	2.13	2.11	2.04	2.62
TAP	39.38	68.58	69.31	71.71	74.36
ASV	51.27	51.41	62.74	66.16	68.46
FFF (mean-std)	43.19	51.74	51.84	60.94	62.38
FFF (no-data)	42.24	30.11	39.31	43.11	31.94
FFF (one-sample)	16.55	23.80	36.39	37.32	54.41
FG-UAP	1.15	1.85	3.20	3.53	1.83
GD-UAP (no-data)	39.36	54.65	53.00	56.74	61.10
GD-UAP (mean-std)	32.64	38.54	41.85	42.88	37.96
GD-UAP (one-sample)	33.28	25.60	37.95	35.48	37.54
L4A-base	40.59	32.82	32.86	22.65	36.35
L4A-fuse	39.98	32.97	32.77	22.58	36.01
L4A-ugs	39.77	35.36	54.70	53.27	60.03
PD-UAP	39.90	41.47	50.45	50.02	54.06
SSP	21.82	35.16	26.31	22.75	41.05
STD	36.97	47.11	53.42	53.72	58.08
UAP (DeepFool)	18.58	10.96	19.48	22.53	20.25
UAPEPGD	56.11	65.48	66.16	68.83	71.21
Clean Accuracy	63.77	69.15	69.67	72.09	74.57
IAA Avg.	22.98 ↓60%	27.85↓60%	30.11↓57%	31.46↓56%	32.86↓56%
UAP Avg.	34.59 ↓46%	36.19↓48%	41.40↓41%	41.41↓43%	45.79↓39%
Adv Avg.	28.44↓55%	31.78↓54%	35.42↓49%	36.14↓50%	38.94↓48%

6.9. Transfer Learning (Linear)

Table 15. Combined results from transfer learning datasets showing Clean accuracy, UAP Avg., IAA Avg., and Adv Avg. with percentage drops relative to Clean accuracy.

		Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
Aircraft	Clean	56.88	56.34	60.25	58.75	46.77	44.89	54.01	56.43
	IAA	16.29↓71%	14.87↓73%	15.27↓75%	17.41↓70%	11.93↓74%	9.82↓78%	13.82↓74%	16.38↓71%
	UAP	20.22↓64%	19.64↓65%	19.25↓68%	25.57↓56%	16.30↓65%	15.54↓65%	16.01↓70%	19.82↓64%
	Adv	18.14↓68%	17.11↓70%	17.14↓72%	21.25↓64%	13.99↓70%	12.51↓72%	14.85↓73%	18.00↓68%
Caltech	Clean	90.54	90.99	90.31	92.89	89.10	90.25	90.36	90.57
	IAA	53.60↓41%	54.06↓41%	47.42↓47%	58.23↓37%	49.79↓44%	44.10↓51%	45.55↓50%	53.64↓41%
	UAP	66.48↓27%	69.06↓16%	65.18↓28%	74.95↓7%	66.50↓23%	65.08↓27%	59.24↓20%	65.95↓24%
	Adv	59.66↓34%	61.12↓33%	55.78↓38%	66.10↓29%	57.66↓35%	53.97↓40%	51.99↓42%	59.44↓34%
Cars	Clean	64.20	57.62	65.62	63.61	43.81	47.10	59.78	64.12
	IAA	19.90↓69%	15.84↓73%	17.54↓73%	20.12↓68%	11.14↓75%	9.56↓80%	14.95↓75%	19.66↓69%
	UAP	30.44↓53%	26.86↓54%	25.52↓61%	34.77↓45%	15.93↓64%	17.44↓63%	19.95↓67%	28.17↓56%
	Adv	24.86↓61%	21.02↓64%	22.13↓66%	27.01↓57%	13.90↓68%	13.27↓72%	17.93↓70%	24.58↓62%
CIFAR 10	Clean	92.78	93.05	93.85	94.67	90.98	91.40	93.90	92.79
	IAA	32.34↓65%	31.19↓66%	28.07↓70%	32.85↓65%	30.00↓67%	31.74↓65%	27.37↓71%	32.45↓65%
	UAP	23.40↓75%	22.13↓76%	25.83↓72%	25.63↓73%	27.29↓70%	18.81↓79%	21.49↓77%	22.63↓76%
	Adv	28.14↓70%	26.93↓71%	27.02↓71%	29.46↓69%	28.73↓68%	25.66↓72%	24.61↓74%	27.83↓70%
CIFAR 100	Clean	77.86	78.18	76.67	80.19	72.97	73.86	79.41	77.79
	IAA	23.34↓70%	22.65↓71%	20.45↓74%	22.77↓72%	18.36↓75%	21.72↓71%	19.59↓75%	24.05↓69%
	UAP	10.93↓86%	11.78↓85%	12.55↓84%	12.49↓84%	10.60↓85%	8.27↓89%	9.55↓88%	11.19↓86%
	Adv	17.50↓77%	17.54↓77%	16.68↓79%	17.94↓78%	14.71↓80%	15.39↓76%	14.87↓75%	18.00↓77%
DTD	Clean	79.97	76.76	77.02	75.43	73.19	72.13	77.45	77.61
	IAA	40.02↓50%	37.65↓51%	38.88↓50%	40.14↓50%	33.50↓54%	33.86↓53%	38.96↓50%	41.30↓47%
	UAP	63.57↓17%	61.54↓20%	61.52↓20%	62.51↓17%	59.00↓19%	53.20↓26%	59.85↓23%	64.98↓16%
	Adv	51.11↓34%	48.90↓36%	49.53↓36%	50.67↓33%	45.51↓38%	42.96↓40%	48.79↓37%	52.45↓32%
Flowers	Clean	94.92	93.36	95.23	94.07	90.57	90.59	93.84	94.92
	IAA	47.71↓50%	43.94↓53%	43.76↓54%	47.25↓50%	40.25↓56%	34.86↓62%	39.92↓58%	47.94↓50%
	UAP	57.51↓39%	59.58↓36%	58.50↓39%	66.71↓29%	55.73↓38%	46.55↓49%	46.21↓51%	57.29↓40%
	Adv	52.32↓45%	51.30↓45%	50.70↓47%	56.41↓40%	47.54↓48%	40.36↓55%	42.88↓54%	52.34↓45%
Food	Clean	76.09	73.07	78.42	73.83	67.24	69.05	76.51	75.81
	IAA	27.50↓64%	24.15↓67%	24.09↓69%	27.69↓62%	21.03↓69%	19.81↓71%	23.39↓69%	26.37↓65%
	UAP	40.06↓47%	37.88↓48%	39.60↓50%	42.42↓43%	30.05↓55%	25.37↓63%	30.51↓60%	38.54↓49%
	Adv	33.41↓56%	30.61↓58%	31.39↓60%	34.62↓53%	25.28↓62%	22.43↓68%	26.74↓65%	32.10↓58%
Pets	Clean	89.13	89.08	89.15	90.77	83.23	92.06	87.47	89.13
	IAA	45.87↓49%	44.48↓50%	39.48↓56%	50.74↓44%	37.75↓55%	41.79↓55%	36.73↓58%	45.95↓48%
	UAP	69.99↓21%	72.36↓19%	67.33↓24%	76.55↓16%	66.09↓21%	71.36↓22%	64.99↓26%	70.52↓21%
	Adv	57.22↓36%	57.60↓35%	52.58↓41%	62.88↓31%	51.09↓39%	55.71↓39%	50.03↓43%	57.52↓35%
All	**Clean**	80.26	78.71	80.72	80.47	73.09	74.59	79.19	79.90
	IAA	34.06↓58%	32.09↓59%	30.55↓62%	35.24↓56%	28.19↓62%	27.47↓63%	28.92↓63%	34.19↓57%
	UAP	41.89↓49%	41.67↓49%	41.32↓50%	45.98↓44%	38.39↓50%	35.27↓55%	36.20↓55%	41.72↓49%
	Adv	37.75↓54%	36.60↓55%	35.62↓57%	40.31↓51%	33.00↓57%	31.15↓60%	32.35↓60%	37.74↓54%

6.9.1. AirCraft

Table 16. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	21.31	25.20	28.18	39.65	22.87	20.31	23.38	21.18
FFF (mean-std)	17.59	25.94	23.81	28.10	17.91	13.76	20.00	17.40
FFF (no-data)	14.86	17.93	10.41	25.14	10.69	7.07	6.54	15.05
FFF (one-sample)	13.21	13.21	3.21	13.12	11.94	7.67	6.81	10.69
FG-UAP	2.36	1.52	2.14	1.99	1.55	1.27	3.32	2.05
GD-UAP (mean-std)	18.95	18.03	15.16	21.30	15.17	22.07	15.71	16.51
GD-UAP (no-data)	16.79	11.39	7.79	23.47	18.70	11.18	7.32	15.83
GD-UAP (one-sample)	3.41	1.90	3.24	5.61	1.56	4.48	6.11	3.48
L4A-base	31.12	24.98	32.35	37.42	27.29	15.81	15.24	31.54
L4A-fuse	31.47	25.71	32.00	38.29	27.28	15.84	16.27	31.11
L4A-ugs	41.55	36.54	42.27	40.98	25.42	26.53	32.97	41.93
PD-UAP	26.27	13.53	8.70	18.34	17.58	11.92	10.62	26.54
SSP	22.28	23.89	27.06	31.55	22.69	12.69	20.22	23.39
STD	20.66	26.74	26.68	33.18	12.77	25.20	23.94	19.70
UAP (DeepFool)	7.05	10.30	8.59	13.81	8.29	24.45	15.34	6.45
UAPEPGD	34.64	37.38	36.34	37.23	19.05	28.45	32.37	34.22
Clean Accuracy	56.88	56.34	60.25	58.75	46.77	44.89	54.01	56.43
IAA Avg.	16.29 ↓71%	14.87↓73%	15.27↓75%	17.41↓70%	11.93↓74%	9.82↓78%	13.82↓74%	16.38↓71%
UAP Avg.	20.22↓64%	19.64↓65%	19.25↓68%	25.57↓56%	16.30↓65%	15.54↓65%	16.01↓70%	19.82↓64%
Adv Avg.	18.14↓68%	17.11↓70%	17.14↓72%	21.25↓64%	13.99↓70%	12.51↓72%	14.85↓73%	18.00↓68%

6.9.2. Caltech 101

Table 17. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Caltech 101 dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	75.31	75.58	66.93	79.84	66.06	62.11	63.12	75.3
$FGSM_2$	53.82	52.44	37.84	59.58	47.67	27.38	36.13	53.82
PGD_1	74.27	75.19	65.57	79.35	64.94	58.96	61.96	74.34
PGD_2	9.61	10.47	2.24	17.17	11.14	1.64	2.05	9.34
PGD_3	74.43	75.39	65.7	79.81	65	59.24	62.28	74.68
PGD_4	7.62	9	1.81	14.79	10.22	1.19	1.69	7.53
PGD_5	17.17	18.64	5.48	25.45	13.11	4.35	3.91	16.86
DIFGSM	80.24	81.09	76.38	83.66	76.16	71.28	75.23	79.97
CW	0.68	0.94	0.3	0.79	0.49	0.22	0.31	0.68
Jitter	83.43	83.41	81.7	86.82	80.89	77.36	79.34	83.85
TIFGSM	85.73	86.72	83.63	88.69	82.73	79.58	81.98	85.98
PIFGSM	68.03	68.03	53.66	74.14	50.54	49	45.82	67.98
EADEN	0	0	0	0	0	0	0	0
OnePixel	89.85	90.57	89.43	92.25	87.67	88.7	89.52	89.88
Pixle	53.89	57.26	40.6	67.39	49.57	39.02	32.73	54.58
SPSA	88.89	88.82	87.45	91.08	86.51	85.73	87.2	89.04
Square	11.43	8.71	4.7	14.98	15.14	1.03	6.53	11.37
TAP	90.48	90.91	90.16	92.36	88.52	87.13	90.12	90.48
ASV	87.36	88.54	87.05	91.07	86.34	84.57	86.72	87.61
FFF (mean-std)	83.37	79.83	83.83	88.14	81.32	72.29	80.73	83.66
FFF (no-data)	73.92	51.35	58.54	77.23	75.82	55.89	55.47	71.03
FFF (one-sample)	42.23	51.33	65.30	79.74	44.95	65.27	65.96	37.53
FG-UAP	8.78	6.05	10.58	15.68	17.41	5.23	8.04	9.26
GD-UAP (mean-std)	83.06	81.74	65.43	78.92	80.73	78.18	72.84	81.71
GD-UAP (no-data)	58.14	62.99	51.64	73.74	73.92	58.89	62.15	59.71
GD-UAP (one-sample)	33.34	68.44	18.83	37.63	27.16	71.42	43.55	38.20
L4A-base	70.40	79.34	84.44	86.36	63.87	66.12	27.52	70.06
L4A-fuse	69.97	78.93	84.56	85.72	63.84	64.96	27.34	69.21
L4A-ugs	88.72	88.78	86.75	92.23	83.99	82.42	85.82	88.82
PD-UAP	88.58	75.06	73.89	86.96	84.04	71.07	75.22	88.18
SSP	73.96	82.74	78.66	85.34	80.62	57.39	64.53	71.47
STD	88.03	89.12	84.49	90.76	84.99	86.23	85.52	88.12
UAP (DeepFool)	24.67	30.67	21.12	37.96	29.01	33.89	18.49	21.36
UAPEPGD	89.13	90.02	87.70	91.76	86.07	87.37	87.94	89.29
Clean Accuracy	90.54	90.99	90.31	92.89	89.1	90.25	90.36	90.57
IAA Avg.	53.60 ↓41%	54.06↓41%	47.42↓47%	58.23↓37%	49.79↓44%	44.10↓51%	45.55↓50%	53.64↓41%
UAP Avg.	66.48 ↓26.58%	69.06↓15.8%	65.18↓27.9%	74.95↓7.4%	66.50↓23.2%	65.08↓26.8%	59.24↓20.3%	65.95↓24.2%
Adv Avg.	59.66 ↓34.1%	61.12↓32.8%	55.78↓38.2%	66.10↓28.8%	57.66↓35.3%	53.97↓40.2%	51.99↓42.5%	59.44↓34.4%

6.9.3. Cars

Table 18. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Cars dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	14.55	8.27	6.34	16.32	3.18	2.1	3.48	14.48
$FGSM_2$	1.41	0.6	0.51	1.39	0.9	0.16	0.5	1.42
PGD_1	14.15	7.76	5.83	15.3	3.42	1.6	3.03	13.94
PGD_2	0.02	0	0	0	0.19	0.09	0	0.02
PGD_3	14.3	8.05	5.83	15.5	3.52	1.67	3.11	14.33
PGD_4	0.01	0	0	0	0.17	0.07	0	0.01
PGD_5	0	0.01	0	0	0.19	0	0.02	0
DIFGSM	33.68	24.49	28.83	32.76	17.55	14.05	22.04	30.92
CW	0	0	0	0	0	0	0	0
Jitter	44.21	36.41	45.44	42.21	26.4	23.85	40.67	43.86
TIFGSM	44.26	35.39	39.8	43.84	27.62	22.01	34.77	44
PIFGSM	6.32	3.3	1.54	8.54	0.75	0.6	0.65	6.39
EADEN	0	0	0	0	0	0	0	0
OnePixel	60.73	53.74	61.77	59.99	39.56	39.56	55.33	60.63
Pixle	6.63	5.1	5.02	8.17	4.14	1.92	2.3	6.33
SPSA	54.22	45.44	51.11	54.88	30.33	31.59	44.47	54.05
Square	0.06	0.01	0	0.04	0	0	0.01	0.05
TAP	63.71	56.62	63.75	63.26	42.74	32.92	58.79	63.51
ASV	43.05	35.78	42.81	51.73	28.26	27.47	33.37	41.80
FFF (mean-std)	35.13	34.85	34.66	38.75	22.62	19.05	26.25	34.98
FFF (no-data)	23.84	15.97	12.54	29.39	18.54	15.92	10.19	17.47
FFF (one-sample)	17.76	22.51	19.75	38.42	12.64	16.70	29.70	19.15
FG-UAP	2.69	0.82	1.78	1.04	2.18	0.92	1.27	2.90
GD-UAP (mean-std)	41.11	29.44	26.35	39.20	23.24	24.42	27.16	40.33
GD-UAP (no-data)	17.31	20.43	15.52	27.40	15.31	13.08	19.67	18.77
GD-UAP (one-sample)	5.92	17.22	4.99	12.35	3.51	5.88	16.04	6.47
L4A-base	33.94	28.37	34.16	37.86	9.80	8.62	6.39	33.17
L4A-fuse	33.29	28.72	33.13	38.13	8.72	8.68	6.72	33.03
L4A-ugs	56.10	38.76	47.90	55.12	29.59	21.00	34.34	55.49
PD-UAP	48.19	26.53	22.61	34.37	23.79	17.17	12.26	47.34
SSP	21.66	33.84	31.08	38.27	17.26	8.03	19.59	24.26
STD	37.64	31.02	34.57	39.36	16.42	24.72	28.23	36.97
UAP (DeepFool)	16.91	20.25	25.59	20.54	14.21	32.35	28.93	17.85
UAPEPGD	52.56	45.27	49.12	54.36	26.15	34.96	40.44	51.96
Clean Accuracy	64.20	57.62	65.62	63.61	43.81	47.10	59.78	64.12
IAA Avg.	19.90 ↓69%	15.84↓73%	17.54↓73%	20.12↓68%	11.14↓75%	9.56↓80%	14.95↓75%	19.66↓69%
UAP	30.44↓53%	26.86↓54%	25.52↓61%	34.77↓45%	15.93↓64%	17.44↓63%	19.95↓67%	28.17↓56%
Adv Avg.	24.86↓61%	21.02↓64%	22.13↓66%	27.01↓57%	13.90↓68%	13.27↓72%	17.93↓70%	24.58↓62%

6.9.4. CIFAR 10

Table 19. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the CIFAR 10 dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	32.95	31.04	27.57	33.04	37.86	42.84	19.38	33.04
$FGSM_2$	53.83	50.24	52.58	52.51	59.88	29.71	47.54	53.94
PGD_1	34.76	29.2	22.25	35.16	23.51	36.92	21.04	34.64
PGD_2	0.02	0	0	0	0.03	0	0	0.01
PGD_3	34.02	28.38	20.85	34.44	22.48	36.51	20.71	34.23
PGD_4	0.02	0.02	0	0	0.03	0	0	0
PGD_5	0	0	0	0	0.01	0	0	0
DIFGSM	56.24	52.78	42.53	55.48	52.39	55.9	39.2	54.64
CW	0	0	0	0	0.06	0	0	0
Jitter	66.67	62.37	59.8	66.97	55.15	70.7	58.5	67.63
TIFGSM	52.32	48.88	41.23	50.64	56.11	56.88	42.38	54.51
PIFGSM	0.39	0.22	0.04	0.28	0.45	5.18	0	0.41
EADEN	0	0	0	0	0	0	0	0
OnePixel	87.36	86.09	88.42	87.78	82.28	85.59	81.11	87.21
Pixle	5.55	2.15	4.44	3.02	1.82	2.22	1.93	5.41
SPSA	69.6	79.09	55.69	80.73	60.51	71.34	68.81	69.9
Square	0	0	0.05	0	0	0	0	0
TAP	88.51	91.06	89.82	91.4	87.56	77.66	92.14	88.59
ASV	50.49	71.15	66.54	67.20	51.96	36.61	70.25	50.81
FFF (mean-std)	22.27	23.90	30.37	30.11	30.74	13.79	26.08	22.83
FFF (no-data)	14.30	10.41	12.37	19.61	31.12	13.33	10.83	12.70
FFF (one-sample)	10.00	10.00	9.99	10.00	10.00	11.15	10.03	10.05
FG-UAP	10.16	6.68	11.38	9.99	8.25	10.01	10.00	10.06
GD-UAP (mean-std)	33.92	18.14	22.02	17.21	28.83	19.01	17.66	26.04
GD-UAP (no-data)	13.37	10.56	12.09	10.58	10.51	12.33	10.28	11.49
GD-UAP (one-sample)	10.00	11.94	10.00	10.00	10.00	9.79	10.22	10.00
L4A-base	10.12	10.62	26.26	12.13	10.08	13.28	10.97	10.16
L4A-fuse	10.12	10.51	28.99	12.59	10.01	13.29	11.16	10.09
L4A-ugs	12.89	16.29	50.54	41.20	50.43	14.48	10.19	12.90
PD-UAP	67.51	16.31	14.67	12.94	49.02	14.89	40.65	66.96
SSP	12.66	10.13	10.01	41.45	15.03	10.42	9.92	11.01
STD	23.12	41.06	27.91	30.18	52.46	27.27	20.45	24.00
UAP (DeepFool)	8.93	13.05	16.73	9.92	12.51	18.11	11.23	9.14
UAPEPGD	64.61	73.37	63.41	74.96	55.63	63.25	63.91	63.76
Clean Accuracy	92.78	93.05	93.85	94.67	90.98	91.4	93.9	92.79
IAA Avg.	32.34 ↓65%	31.19↓66%	28.07↓70%	32.85↓65%	30.00↓67%	31.74↓65%	27.37↓71%	32.45↓65%
UAP Avg.	23.40↓74.77	22.13↓76.21	25.83↓72.48	25.63↓72.93	27.29↓70.01	18.81↓79.42	21.49↓77.11	22.63↓75.62
Adv Avg.	28.14↓70%	26.93↓71%	27.02↓71%	29.46↓69%	28.73↓68%	25.66↓72%	24.61↓74%	27.83↓70%

6.9.5. CIFAR 100

Table 20. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the CIFAR 100 dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	20.52	19.01	16.03	19.29	19.49	24.51	11.07	22.34
$FGSM_2$	34.07	31.02	34.08	28.84	30.06	18.20	29.16	35.71
PGD_1	19.74	14.42	11.47	18.38	8.92	19.09	10.29	20.98
PGD_2	0.04	0	0	0.02	0.12	0	0.01	0.06
PGD_3)	19.33	14.18	11.09	17.69	8.24	18.85	9.92	20.67
PGD_4	0.06	0.01	0	0.01	0.08	0	0	0.02
PGD_5	0	0	0	0	0.18	0.01	0	0
DIFGSM	38.20	35.26	27.54	32.23	32.56	34.97	26.31	39.47
CW	0.01	0	0	0.06	0.02	0.02	0	0.04
Jitter	66.85	62.15	59.33	65.89	42.01	67.10	53.82	66.73
TIFGSM	34.84	36.35	27.80	30.79	35.15	36.82	29.15	37.30
PIFGSM	0.78	0.34	0.17	0.58	0.36	3.29	0.09	1.10
EADEN	0	0	0	0	0	0	0	0
OnePixel	67.73	66.25	69.87	67.64	58.73	64.76	61.41	68.19
Pixle	0.48	0.96	0.56	0.90	0.96	1.40	0.43	0.55
SPSA	47.25	54.96	38.62	53.70	28.82	48.30	44.86	49.46
Square	0.06	0.01	0.04	0	0	0	0.01	0.05
TAP	70.29	72.88	69.76	73.97	65.1	53.57	76.14	70.29
ASV	34.24	49.85	45.83	39.93	26.05	29.08	42.43	36.42
FFF (mean-std)	10.69	15.61	9.80	9.78	14.54	3.56	15.74	11.36
FFF (no-data)	2.50	1.02	3.68	3.11	2.47	3.00	1.60	1.61
FFF (one-sample)	1.03	1.02	1.02	0.90	1.30	1.21	1.11	1.00
FG-UAP	1.00	1.06	0.76	0.96	1.01	1.05	1.00	1.00
GD-UAP (mean-std)	9.68	13.00	6.89	5.14	11.58	8.36	7.28	9.36
GD-UAP (no-data)	1.05	2.15	1.11	1.40	4.18	2.51	1.11	1.06
GD-UAP (one-sample)	1.97	1.18	1.00	1.29	1.28	1.03	1.94	1.41
L4A-base	1.79	3.18	17.08	5.35	1.09	4.42	0.75	1.68
L4A-fuse	1.64	3.00	17.30	5.23	1.20	4.16	0.81	1.66
L4A-ugs	4.31	11.44	27.63	23.33	28.02	11.62	2.36	4.77
PD-UAP	47.61	6.95	6.56	10.54	22.54	2.65	20.47	48.33
SSP	1.13	1.69	4.49	29.29	6.20	1.75	1.08	1.16
STD	10.57	24.42	11.34	14.05	19.14	13.12	8.17	10.82
UAP (DeepFool)	1.72	2.44	3.88	4.03	2.57	2.65	4.80	1.69
UAPEPGD	43.92	50.47	42.35	45.49	26.42	42.15	42.14	45.77
Clean Accuracy	77.86	78.18	76.67	80.19	72.97	73.86	79.41	77.79
IAA Avg.	23.34 ↓70%	22.65↓71%	20.45↓74%	22.77↓72%	18.36↓75%	21.72↓71%	19.59↓75%	24.05↓69%
UAP Avg.	10.93 ↓86%	11.78 ↓85%	12.55 ↓84%	12.49 ↓84%	10.60 ↓85%	8.27 ↓89%	9.55 ↓88%	11.19 ↓86%
Adv Avg.	17.50 ↓77%	17.54 ↓77%	16.68 ↓79%	17.94 ↓78%	14.71 ↓80%	15.39 ↓76%	14.87 ↓75%	18.00 ↓77%

6.9.6. DTD

Table 21. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the DTD dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	50.43	46.76	48.88	51.65	38.4	42.02	48.99	52.71
$FGSM_2$	23.24	21.28	23.94	24.63	17.87	17.66	25.80	26.54
PGD_1	50.05	46.01	47.93	51.17	39.31	40.05	48.99	51.65
PGD_2	6.91	4.57	3.46	6.38	2.13	3.19	3.35	6.91
PGD_3	50.11	46.54	48.14	51.17	39.04	40.27	48.62	51.65
PGD_4	6.54	3.94	2.82	5.96	1.7	2.93	3.03	6.60
PGD_5	14.89	12.23	11.81	16.76	3.99	10.37	10.53	16.22
DIFGSM	59.84	52.87	60.05	59.79	52.02	54.47	60.27	64.20
CW	0.32	0.32	0.74	0.69	0.43	0.64	0.90	0.90
Jitter	67.39	65.90	66.91	66.17	62.02	60.48	68.51	68.30
TIFGSM	67.77	65.32	67.93	66.06	62.07	62.34	67.39	68.88
PIFGSM	42.77	38.83	40.16	45.53	26.76	35.43	38.40	43.94
EADEN	0	0	0	0	0	0	0	0
OnePixel	75.32	75.43	76.17	74.41	71.12	70.69	75.96	76.28
Pixle	49.89	46.28	46.97	49.57	40.48	37.62	41.38	50.90
SPSA	72.87	71.81	73.51	72.39	67.98	66.91	73.78	74.15
Square	8.09	5.96	6.7	7.77	5.74	1.49	8.46	8.67
TAP	74.10	73.78	73.72	72.50	72.07	62.98	76.97	75.05
ASV	71.01	70.59	72.18	68.83	67.39	62.98	70.69	72.07
FFF (mean-std)	62.93	59.15	65.27	61.70	61.65	50.85	65.90	64.20
FFF (no-data)	58.62	54.15	55.32	56.33	62.61	43.99	49.73	60.59
FFF (one-sample)	55.32	53.88	49.36	57.23	40.69	37.02	38.62	57.02
FG-UAP	21.70	20.90	19.26	26.86	23.14	19.95	19.15	23.83
GD-UAP (mean-std)	62.13	58.46	59.41	57.39	61.97	54.89	61.38	63.94
GD-UAP (no-data)	53.62	55.27	52.18	57.18	54.04	44.89	53.62	55.80
GD-UAP (one-sample)	50.69	48.56	42.82	46.01	32.45	35.05	47.23	55.43
L4A-base	72.61	72.82	73.88	73.30	67.07	63.51	67.61	73.99
L4A-fuse	72.82	73.24	73.94	73.51	66.17	62.93	68.19	73.56
L4A-ugs	74.31	74.26	74.10	73.24	69.04	68.40	74.63	74.84
PD-UAP	72.82	60.74	59.84	62.82	65.74	49.15	60.16	73.14
SSP	72.50	71.86	73.09	72.98	66.17	59.89	71.38	73.83
STD	73.40	71.91	73.19	71.28	69.31	68.56	72.98	73.72
UAP (DeepFool)	69.26	65.69	67.02	68.67	68.19	61.01	62.29	69.89
UAPEPGD	73.46	73.19	73.40	72.82	68.40	68.14	74.10	73.78
Clean Accuracy	79.97	76.76	77.02	75.43	73.19	72.13	77.45	77.61
IAA Avg.	40.02 ↓50%	37.65↓51%	38.88↓50%	40.14↓50%	33.50↓54%	33.86↓53%	38.96↓50%	41.30↓47%
UAP Avg.	63.57↓17%	61.54↓20%	61.52↓20%	62.51↓17%	59.00↓19%	53.20↓26%	59.85↓23%	64.98↓16%
Adv Avg.	51.11↓34%	48.90↓36%	49.53↓36%	50.67↓33%	45.51↓38%	42.96↓40%	48.79↓37%	52.45↓32%

6.9.7. Flowers

Table 22. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Flowers dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	66.36	57.69	57.37	64.52	48.50	41.85	46.97	66.36
$FGSM_2$	25.96	17.49	19.44	24.96	19.00	7.68	13.33	25.96
PGD_1	66.03	55.99	55.60	63.31	50.45	36.97	46.65	65.81
PGD_2	1.51	0.37	0.17	1.10	0.15	0.00	0.06	1.65
PGD_3	66.19	56.37	55.95	63.50	51.00	37.31	46.72	66.44
PGD_4	1.21	0.38	0.13	0.90	0.13	0.00	0.02	1.29
PGD_5	8.03	4.90	2.81	7.17	0.92	0.72	0.89	8.05
DI2FGSM	74.42	72.08	69.73	75.75	62.56	56.94	67.56	78.12
CW	0.00	0.00	0.05	0.00	0.00	0.02	0.00	0.00
Jitter	84.93	80.12	81.87	82.53	79.85	73.62	79.24	84.33
TIFGSM	86.85	84.35	87.48	86.17	81.29	75.36	84.39	87.88
PIFGSM	53.81	43.06	39.04	51.65	29.16	27.46	28.63	53.85
EADEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OnePixel	94.47	92.77	94.79	93.10	89.27	88.38	92.94	94.49
Pixle	35.32	38.34	31.21	45.08	32.07	20.88	24.05	35.09
SPSA	93.03	90.21	92.91	91.84	85.56	84.31	90.60	92.84
Square	6.70	4.17	4.40	5.60	4.90	0.06	3.32	6.70
TAP	94.01	92.77	94.76	93.31	89.76	75.93	93.14	94.01
ASV	81.20	81.11	88.69	86.44	77.76	68.66	80.83	81.60
FFF (mean-std)	69.41	71.24	72.18	73.24	63.71	69.07	69.08	69.32
FFF (no-data)	52.74	49.93	43.70	63.13	58.11	49.88	32.65	52.55
FFF (one-sample)	23.79	50.12	37.17	77.48	31.42	16.90	20.18	22.73
FG-UAP	7.47	9.08	9.72	10.72	22.59	5.65	5.19	7.28
GD-UAP (mean-std)	78.75	67.02	64.74	76.07	69.99	70.11	65.18	78.08
GD-UAP (no-data)	48.38	51.64	39.25	57.48	57.36	42.94	35.00	47.88
GD-UAP (one-sample)	22.13	45.88	13.95	21.55	19.79	13.95	16.12	25.00
L4A-base	56.73	69.04	82.30	76.55	47.71	40.11	31.88	56.64
L4A-fuse	57.26	68.86	82.66	76.63	48.40	38.92	31.94	56.94
L4A-ugs	89.36	86.32	91.14	90.80	81.58	63.86	77.50	89.35
PD-UAP	85.11	59.04	48.22	68.65	75.03	53.21	52.31	84.59
SSP	53.40	47.61	69.40	87.01	70.49	22.10	42.08	54.17
STD	72.29	72.70	72.37	75.94	61.46	63.29	68.90	72.30
UAP (DeepFool)	31.95	34.80	28.73	34.62	25.15	44.85	22.05	28.10
UAPEPGD	90.23	88.85	91.75	91.00	81.15	81.37	88.44	90.18
Clean Accuracy	94.92	93.36	95.23	94.07	90.57	90.59	93.84	94.92
IAA Avg.	47.71 ↓50%	43.94↓53%	43.76↓54%	47.25↓50%	40.25↓56%	34.86↓62%	39.92↓58%	47.94↓50%
UAP Avg.	57.51↓39%	59.58↓36%	58.50↓39%	66.71↓29%	55.73↓38%	46.55↓49%	46.21↓51%	57.29↓40%
Adv Avg.	52.32↓45%	51.30↓45%	50.70↓47%	56.41↓40%	47.54↓48%	40.36↓55%	42.88↓54%	52.34↓45%

6.9.8. Food

Table 23. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Food dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	26.40	19.34	14.13	28.69	12.10	13.18	12.95	23.48
$FGSM_2$	3.24	1.50	1.39	4.02	1.41	1.29	0.95	2.52
PGD_1	26.60	19.03	13.87	28.54	13.69	11.30	13.15	23.91
PGD_2	0.04	0.01	0.01	0.05	0.00	0.02	0.00	0.04
PGD_3	26.72	19.21	14.13	28.76	13.92	11.42	13.48	24.12
PGD_4	0.04	0.01	0.00	0.04	0.00	0.01	0.00	0.03
PGD_5	0.59	0.19	0.10	0.82	0.04	0.13	0.01	0.47
DI2FGSM	44.15	37.23	37.35	44.94	33.02	32.45	37.32	40.14
CW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jitter	60.70	56.00	61.34	58.14	55.13	53.14	61.79	59.79
TIFGSM	57.43	51.93	53.38	56.41	48.65	45.76	54.04	56.51
PIFGSM	17.53	11.71	6.67	19.93	5.17	6.80	5.46	14.85
EADEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OnePixel	73.54	69.95	76.00	71.41	63.59	64.63	73.65	73.22
Pixle	14.94	12.97	9.65	17.11	8.34	5.84	4.93	13.66
SPSA	68.75	64.12	69.31	66.70	57.49	56.88	67.11	68.04
Square	0.19	0.05	0.09	0.19	0.16	0.02	0.07	0.16
TAP	74.21	71.43	76.25	72.68	65.96	53.74	76.18	73.75
ASV	62.46	59.82	66.04	61.50	52.91	47.60	62.86	62.02
FFF (mean-std)	51.98	49.45	55.35	49.68	44.52	41.29	53.77	49.99
FFF (no-data)	45.78	29.52	39.07	47.63	44.67	25.60	24.66	43.53
FFF (one-sample)	20.82	34.82	29.72	36.80	4.63	20.67	10.41	19.50
FG-UAP	3.98	2.50	2.22	4.28	3.90	1.84	1.67	3.88
GD-UAP (mean-std)	59.88	49.04	52.61	51.72	47.22	44.65	51.52	58.62
GD-UAP (no-data)	42.31	38.30	30.52	46.00	37.85	22.85	32.35	40.77
GD-UAP (one-sample)	13.89	29.22	8.69	14.46	5.33	7.26	15.75	12.92
L4A-base	26.08	35.83	48.51	43.13	5.36	15.57	7.30	25.89
L4A-fuse	25.76	35.50	47.07	43.72	5.11	15.43	7.70	25.86
L4A-ugs	58.75	46.16	53.52	61.22	41.61	18.25	39.10	57.71
PD-UAP	63.56	46.55	33.67	50.59	50.14	29.66	37.59	62.90
SSP	34.22	26.85	35.80	41.64	26.52	6.07	23.58	28.30
STD	45.39	37.86	42.17	40.49	34.37	34.74	43.51	44.38
UAP (DeepFool)	16.28	18.95	16.98	17.35	19.64	15.37	7.83	10.85
UAPEPGD	69.89	65.73	71.68	68.50	57.03	59.03	68.59	69.47
Clean Accuracy	76.09	73.07	78.42	73.83	67.24	69.05	76.51	75.81
IAA Avg.	27.50 ↓64%	24.15↓67%	24.09↓69%	27.69↓62%	21.03↓69%	19.81↓71%	23.39↓69%	26.37↓65%
UAP Avg.	40.06↓47%	37.88↓48%	39.60↓50%	42.42↓43%	30.05↓55%	25.37↓63%	30.51↓60%	38.54↓49%
Adv Avg.	33.41↓56%	30.61↓58%	31.39↓60%	34.62↓53%	25.28↓62%	22.43↓68%	26.74↓65%	32.10↓58%

6.9.9. Pets

Table 24. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the Pets dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

Method	Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	63.58	61.00	48.74	71.38	44.60	55.10	41.59	63.58
$FGSM_2$	25.08	21.62	11.81	34.65	17.20	14.17	8.74	25.08
PGD_1	64.38	60.82	48.07	71.07	46.76	52.20	43.00	64.30
PGD_2	0.82	0.41	0.08	2.96	0.16	0.00	0.03	0.79
PGD_3	64.52	61.21	48.10	71.29	47.25	52.21	43.42	64.52
PGD_4	0.63	0.27	0.03	2.39	0.11	0.00	0.03	0.57
PGD_5	6.54	5.69	0.89	14.03	0.98	1.38	0.43	6.51
DI2FGSM	73.92	71.18	63.92	78.63	61.06	68.25	59.70	74.18
CW	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
Jitter	80.75	79.82	75.82	84.06	74.50	78.41	75.60	80.83
TIFGSM	81.43	80.30	78.13	84.89	75.31	80.60	76.11	82.31
PIFGSM	54.24	51.35	34.23	64.67	31.70	41.02	26.11	54.24
EADEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OnePixel	88.28	87.90	87.65	89.82	81.51	90.60	85.85	88.31
Pixle	42.04	41.46	38.47	59.12	31.61	43.01	29.16	42.31
SPSA	87.27	86.87	85.81	88.79	79.93	88.54	83.95	87.46
Square	3.28	1.71	0.49	4.79	3.88	0.05	0.46	3.30
TAP	88.97	89.04	88.41	90.66	83.01	86.83	87.09	88.97
ASV	84.04	86.07	86.52	89.11	79.46	86.89	83.59	84.07
FFF (mean-std)	79.82	82.97	82.04	85.08	73.86	83.03	81.53	79.69
FFF (no-data)	65.60	47.88	70.12	80.33	53.54	71.01	53.59	62.52
FFF (one-sample)	49.62	85.48	61.54	83.37	63.28	73.35	61.44	56.39
FG-UAP	12.32	6.27	8.53	14.09	12.36	8.46	5.18	12.34
GD-UAP (mean-std)	84.78	81.57	75.38	84.37	75.67	84.82	80.65	85.93
GD-UAP (no-data)	67.48	63.61	55.35	79.27	69.36	65.81	60.80	71.45
GD-UAP (one-sample)	31.59	67.83	23.04	39.50	19.63	34.28	37.79	26.46
L4A-base	84.68	84.02	82.28	88.44	74.45	79.91	61.65	84.55
L4A-fuse	84.54	83.83	81.98	88.16	73.96	79.23	61.69	83.96
L4A-ugs	88.28	88.31	87.47	90.22	80.84	87.68	86.17	88.39
PD-UAP	84.38	75.44	65.26	78.86	77.81	75.73	69.66	84.25
SSP	86.49	85.07	84.23	89.73	78.70	76.96	78.96	86.29
STD	85.41	85.83	84.30	88.53	78.35	88.91	85.46	85.54
UAP (DeepFool)	42.47	45.23	41.35	56.05	67.08	55.57	44.72	48.42
UAPEPGD	88.36	88.44	87.83	89.74	79.14	90.05	87.04	88.11
Clean Accuracy	89.13	89.08	89.15	90.77	83.23	92.06	87.47	89.13
IAA Avg.	45.87 ↓49%	44.48↓50%	39.48↓56%	50.74↓44%	37.75↓55%	41.79↓55%	36.73↓58%	45.95↓48.4%
UAP Avg.	69.99↓21%	72.36↓19%	67.33↓24%	76.55↓16%	66.09↓21%	71.36↓22%	64.99↓26%	70.52↓21%
Adv Avg.	57.22 ↓36%	57.60↓35%	52.58↓41%	62.88↓31%	51.09↓39%	55.71↓39%	50.03↓43%	57.52↓35%

6.10. Transfer Learning (Finetune)

Table 25. Combined results from transfer learning datasets showing Clean accuracy, UAP Avg. , IAA Avg. , and Adv Avg. with percentage drops relative to Clean accuracy.

		Barlow	BYOL	DINO	MocoV3	SimCLR	Supervised	SwAV	VICReg
Aircraft	Clean	86.71	83.14	80.38	82.74	79.35	85.08	86.17	86.95
	IAA	39.62 ↓54%	39.21↓53%	32.36↓60%	40.18↓51%	30.68↓61%	45.4↓47%	42.47↓51%	39.21↓55%
	UAP	36.47↓58%	30.12↓64%	30.45↓62%	38.95↓53%	33.90↓57%	42.25↓50%	26.72↓69%	35.36↓59%
	Adv	38.14↓56%	34.93↓58%	31.46↓61%	39.60↓52%	32.20↓59%	43.92↓48%	35.06↓59%	37.40↓57%
Caltech	Clean	91.44	92.03	89.77	92.92	90.39	91.81	90.37	91.32
	IAA	53.63 ↓41%	57.03↓38%	49.71↓45%	59.32↓36%	53.86↓40%	54.18↓41%	48.01↓47%	53.78↓41%
	UAP	62.03↓32%	63.09↓31%	55.59↓38%	64.37↓31%	67.38↓25%	58.60↓36%	37.55↓58%	62.47↓32%
	Adv	57.58↓37%	59.88↓35%	52.48↓42%	61.70↓34%	60.23↓33%	56.26↓39%	43.09↓52%	57.87↓37%
Cars	Clean	90.54	89.91	89.80	90.11	89.62	88.99	89.73	90.41
	IAA	48.69 ↓46%	54.91↓39%	52.72↓52%	55.39↓39%	47.63↓47%	55.47↓38%	52.04↓42%	48.70↓48%
	UAP	45.75 ↓49%	60.09↓33%	49.85↓44%	65.77↓27%	34.10↓62%	55.54↓38%	50.81↓43%	44.46↓51%
	Adv	47.31 ↓48%	57.35↓36%	52.74↓41%	60.28↓33%	42.02↓53%	55.50↓38%	52.88↓41%	47.90↓47%
CIFAR 10	Clean	97.13	96.89	96.90	96.86	97.22	96.16	96.75	97.07
	IAA	48.74 ↓50%	46.26↓52%	54.18↓44%	49.66↓49%	49.03↓50%	43.37↓55%	55.75↓43%	48.31↓50%
	UAP	55.07 ↓43%	31.96↓67%	58.16↓40%	44.72↓54%	55.76↓43%	36.20↓62%	53.76↓44%	50.00↓48%
	Adv	51.72 ↓47%	39.53↓59%	56.05↓42%	47.34↓51%	52.20↓46%	40.00↓58%	54.81↓43%	49.10↓49%
CIFAR 100	Clean	84.60	83.88	84.69	84.49	84.44	82.63	84.37	84.27
	IAA	35.85 ↓57%	44.11↓47%	39.68↓53%	44.95↓47%	35.54↓58%	42.33↓49%	40.22↓52%	35.59↓58%
	UAP	36.41 ↓57%	36.79↓56%	42.02↓50%	31.57↓63%	28.24↓67%	24.94↓70%	41.03↓51%	31.40↓63%
	Adv	36.12 ↓57%	40.67↓52%	40.78↓52%	38.65↓54%	32.10↓62%	34.14↓59%	40.60↓52%	33.62↓60%
DTD	Clean	76.12	76.28	78.09	75.43	75.90	73.30	74.47	77.29
	IAA	41.42 ↓46%	41.89↓45%	40.50↓48%	42.91↓43%	42.42↓44%	38.99↓47%	39.65↓47%	41.90↓46%
	UAP	62.74 ↓18%	62.74↓18%	61.26↓22%	63.58↓16%	57.58↓24%	57.78↓21%	62.41↓16%	62.54↓19%
	Adv	51.46 ↓32%	51.70↓32%	50.27↓36%	52.64↓30%	49.56↓35%	47.83↓35%	50.36↓32%	51.62↓33%
Flowers	Clean	97.41	96.74	97.18	96.70	95.18	96.73	96.68	96.86
	IAA	60.24 ↓38%	61.32↓37%	54.30↓44%	60.98↓37%	51.34↓46%	60.79↓37%	55.66↓42%	59.97↓38%
	UAP	60.69 ↓38%	67.75↓30%	54.18↓44%	68.26↓29%	60.09↓37%	56.33↓42%	54.51↓44%	63.09↓35%
	Adv	60.45 ↓38%	64.34↓33%	54.25↓44%	64.41↓33%	55.46↓42%	58.69↓39%	55.12↓43%	61.44↓37%
Food	Clean	83.93	85.63	87.65	85.82	82.30	84.35	87.16	83.71
	IAA	27.80 ↓67%	36.30↓58%	34.12↓61.06%	36.09↓58%	28.89↓65%	34.74↓59%	34.26↓59%	31.71↓62%
	UAP	35.93 ↓57%	43.98↓49%	46.95↓46%	44.88↓48%	32.33↓61%	29.07↓66%	39.09↓55%	41.18↓51%
	Adv	35.94 ↓57%	39.91↓53%	40.16↓54%	40.23↓53%	30.51↓63%	32.07↓62%	36.53↓58%	36.17↓57%
Pets	Clean	90.89	91.34	90.24	92.19	88.51	93.94	90.50	90.92
	IAA	43.21 ↓52%	44.69↓51%	40.01↓56%	48.33↓48%	42.03↓53%	46.53↓50%	39.53↓56%	43.76↓52%
	UAP	67.27 ↓26%	71.96↓21%	63.83↓29%	73.20↓21%	67.90↓23%	75.66↓19%	64.06↓29%	67.63↓26%
	Adv	54.53 ↓40%	57.52↓37%	51.23↓43%	60.03↓35%	54.20↓39%	60.24↓36%	51.07↓44%	54.99↓40%
All	**Clean**	88.75	88.43	88.30	88.58	86.99	88.11	88.47	88.76
	IAA	44.36↓50%	47.30↓47%	44.18↓50%	48.65↓45%	42.38↓51%	46.87↓47%	45.29↓49%	44.77↓50%
	UAP	51.37↓42%	52.05↓41%	51.37↓42%	55.03↓38%	48.59↓44%	48.49↓45%	47.77↓46%	50.90↓43%
	Adv	48.14↓46%	49.54↓44%	47.71↓46%	51.65↓42%	45.39↓48%	47.63↓46%	46.61↓47%	47.79↓46%

6.10.1. AirCraft

Table 26. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	53.03	55.42	49.24	59.86	51.83	64.20	62.10	52.77
FFF (mean-std)	48.07	41.76	42.26	39.16	43.73	66.04	48.39	54.66
FFF (no-data)	6.41	17.70	8.80	35.53	53.12	7.79	1.09	2.77
FFF (one-sample)	3.71	3.05	1.51	18.25	4.41	1.21	2.30	2.01
FG-UAP	1.03	1.34	1.36	1.27	1.00	1.47	1.03	1.12
GD-UAP (mean-std)	46.54	19.30	38.41	38.56	17.72	56.61	35.26	47.09
GD-UAP (no-data)	7.43	9.67	7.29	7.59	28.10	9.88	1.18	1.33
GD-UAP (one-sample)	2.30	2.52	1.18	2.92	3.94	1.69	1.06	1.03
L4A-base	63.04	52.01	58.56	56.52	42.81	69.48	10.52	56.20
L4A-fuse	64.09	49.22	58.50	55.87	40.42	68.70	10.54	57.69
L4A-ugs	74.26	64.69	60.52	73.28	55.91	73.75	67.25	74.52
PD-UAP	5.03	11.22	7.77	18.77	37.70	13.39	1.66	1.54
SSP	60.82	28.61	38.99	69.89	42.61	75.01	20.45	55.00
STD	63.15	54.63	48.31	65.25	49.60	75.97	76.34	66.58
UAP (DeepFool)	13.69	4.30	6.35	8.80	12.43	14.44	12.04	20.83
UAPEPGD	70.90	66.42	58.16	71.67	57.00	76.42	76.22	70.58
Clean Accuracy	86.71	83.14	80.38	82.74	79.35	85.08	86.17	86.95
IAA Avg.	39.62 ↓54%	39.21↓53%	32.36↓60%	40.18↓51%	30.68↓61%	45.4↓47%	42.47↓51%	39.21↓55%
UAP Avg.	36.47↓58%	30.12↓64%	30.45↓62%	38.95↓53%	33.90↓57%	42.25↓50%	26.72↓69%	35.36↓59%
Adv Avg.	38.14↓56%	34.93↓58%	31.46↓61%	39.60↓52%	32.20↓59%	43.92↓48%	35.06↓59%	37.40↓57%

6.10.2. Caltech 101

Table 27. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	87.68	84.73	81.98	90.28	81.40	88.77	63.56	87.29
FFF (mean-std)	86.13	80.89	71.22	89.66	74.99	78.03	47.85	83.06
FFF (no-data)	77.25	74.04	42.23	48.19	86.39	41.23	64.82	70.62
FFF (one-sample)	44.59	70.90	24.11	46.35	29.45	32.53	19.60	42.75
FG-UAP	9.42	9.03	2.68	11.08	9.40	6.14	1.76	9.35
GD-UAP (mean-std)	84.10	78.40	59.99	86.99	71.77	83.20	54.39	84.00
GD-UAP (no-data)	61.24	66.08	51.16	63.14	76.93	61.03	36.66	65.29
GD-UAP (one-sample)	53.10	67.34	44.78	51.65	56.50	62.41	24.78	62.74
L4A-base	40.18	38.60	63.97	59.38	79.84	46.59	6.71	40.71
L4A-fuse	39.33	39.98	65.70	58.64	79.27	45.31	6.28	39.46
L4A-ugs	84.46	78.05	69.87	85.20	82.89	82.04	38.86	84.11
PD-UAP	89.44	71.47	62.44	65.06	76.64	58.79	56.72	88.92
SSP	38.38	48.11	64.43	50.72	81.07	39.81	5.19	43.45
STD	90.32	91.14	87.86	91.82	88.33	90.47	87.83	90.52
UAP (DeepFool)	17.40	22.13	14.11	40.02	19.29	31.40	9.62	17.39
UAPEPGD	89.50	88.60	82.84	91.80	83.98	89.76	76.12	89.86
Clean Accuracy	91.44	92.03	89.77	92.92	90.39	91.81	90.37	91.32
IAA Avg.	53.63 ↓41%	57.03↓38%	49.71↓45%	59.32↓36%	53.86↓40%	54.18↓41%	48.01↓47%	53.78↓41%
UAP Avg.	62.03↓32%	63.09↓31%	55.59↓38%	64.37↓31%	67.38↓25%	58.60↓36%	37.55↓58%	62.47↓32%
Adv Avg.	57.58↓37%	59.88↓35%	52.48↓42%	61.70↓34%	60.23↓33%	56.26↓39%	43.09↓52%	57.87↓37%

6.10.3. Cars

Table 28. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	72.91	82.15	78.42	82.19	66.07	80.09	78.46	76.31
FFF (mean-std)	70.69	75.79	69.49	72.04	46.51	72.89	69.07	65.95
FFF (no-data)	31.91	37.45	46.05	60.81	35.42	20.62	49.68	47.43
FFF (one-sample)	47.34	39.57	44.55	53.91	2.18	55.48	43.60	65.58
FG-UAP	0.82	1.87	1.79	3.93	0.60	2.11	1.36	0.63
GD-UAP (mean-std)	60.14	66.63	69.87	71.37	38.22	56.11	70.99	53.70
GD-UAP (no-data)	32.32	40.18	19.64	68.37	33.95	40.78	31.25	18.19
GD-UAP (one-sample)	8.66	27.60	16.48	34.76	7.62	21.95	16.20	7.20
L4A-base	43.38	87.19	63.86	76.66	41.94	71.56	63.09	56.05
L4A-fuse	40.58	86.36	65.68	77.37	42.44	75.14	63.01	54.18
L4A-ugs	68.64	85.40	82.13	86.22	35.49	79.16	84.80	75.12
PD-UAP	55.27	64.43	36.02	77.24	57.46	46.35	57.79	31.04
SSP	28.33	56.45	52.54	68.45	13.66	60.51	39.88	26.18
STD	75.91	79.33	82.30	80.89	68.19	76.01	80.03	78.03
UAP (DeepFool)	11.27	42.93	28.39	49.53	5.38	43.58	24.75	13.23
UAPEPGD	83.88	88.09	87.02	88.56	76.12	86.32	87.23	83.14
Clean Accuracy	90.54	89.91	89.80	90.11	89.62	88.99	89.73	90.41
IAA Avg.	48.69 ↓46%	54.91↓39%	52.72↓52%	55.39↓39%	47.63↓47%	55.47↓38%	52.04↓42%	48.70↓48%
UAP Avg.	45.75 ↓49%	60.09↓33%	49.85↓44%	65.77↓27%	34.10↓62%	55.54↓38%	50.81↓43%	44.46↓51%
Adv Avg.	47.31 ↓48%	57.35↓36%	52.74↓41%	60.28↓33%	42.02↓53%	55.50↓38%	52.88↓41%	47.90↓47%

6.10.4. CIFAR 10

Table 29. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	93.41	71.89	94.31	86.68	88.88	76.26	93.55	91.31
FFF (mean-std)	85.96	63.01	44.25	72.36	86.67	55.1	29.72	77.9
FFF (no-data)	68.68	12.87	52.85	52.54	71.1	25.55	35.1	36.08
FFF (one-sample)	16.72	13.51	47.16	12.91	12.22	10.25	23.85	14.84
FG-UAP	10.03	10.27	11.52	11.15	11.46	10.76	14.94	10.63
GD-UAP (mean-std)	71.66	20.62	34.9	61.37	85.5	33.53	24.17	72.08
GD-UAP (no-data)	71.78	10.31	39.89	33.17	15.38	17.34	38.98	54.72
GD-UAP (one-sample)	12.43	10.12	26.46	10.35	10.26	10.05	22.34	12.13
L4A-base	33.41	14.15	71.00	25.06	58.13	14.48	73.48	29.14
L4A-fuse	32.70	13.32	71.38	27.32	57.63	14.83	74.09	29.74
L4A-ugs	80.68	52.12	91.30	78.32	88.76	61.03	79.24	68.42
PD-UAP	86.15	18.99	87.32	27.60	92.59	27.82	82.89	81.42
SSP	17.16	10.66	43.62	23.58	16.46	21.83	61.30	20.30
STD	88.90	86.48	92.52	81.73	89.29	88.23	88.49	90.13
UAP (DeepFool)	15.98	12.61	26.41	17.99	14.46	20.88	22.31	17.42
UAPEPGD	95.39	90.42	95.70	93.44	93.36	91.23	95.63	93.73
Clean Accuracy	97.13	96.89	96.90	96.86	97.22	96.16	96.75	97.07
IAA Avg.	48.74 ↓50%	46.26↓52%	54.18↓44%	49.66↓49%	49.03↓50%	43.37↓55%	55.75↓43%	48.31↓50%
UAP Avg.	55.07 ↓43%	31.96↓67%	58.16↓40%	44.72↓54%	55.76↓43%	36.20↓62%	53.76↓44%	50.00↓48%
Adv Avg.	51.72 ↓47%	39.53↓59%	56.05↓42%	47.34↓51%	52.20↓46%	40.00↓58%	54.81↓43%	49.10↓49%

6.10.5. CIFAR 100

Table 30. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	76.76	64.38	74.69	54.69	49.74	61.44	72.13	75.5
FFF (mean-std)	61.98	35.03	38.93	19.71	18.57	5.46	48.25	36.28
FFF (no-data)	27.47	34.56	37.48	21.76	34.83	12.85	28.24	21.22
FFF (one-sampl)	1.83	14.42	27.3	15.76	1.53	5.2	22.05	5.92
FG-UAP	1.65	3.13	6.22	2.38	1.49	3.99	4.3	1.02
GD-UAP (mean-std)	48.83	22.66	17.57	14.95	43.36	7.59	51.07	36.65
GD-UAP (no-sample)	52.7	31.68	48.65	36.84	4.43	18.53	42.65	25.12
GD-UAP (one-sample)	4.45	11.91	9.57	11.5	4.57	7.88	18.26	2.17
L4A-base	8.49	24.86	41.34	16.66	17.31	19.07	27.93	8.92
L4A-fuse	8.92	24.55	40.87	17.12	18.88	20.3	26.87	9.18
L4A-ugs	57.08	61.96	73.13	46.77	49.34	45.16	70.88	59.27
PD-UAP	66.13	64.98	65.27	60.42	53.51	17.84	53.14	66.23
SSP	24.29	32.31	29.39	31.57	7.48	13.24	36.67	8.23
STD	57.19	66.71	64.13	57.24	65.29	68.65	55.1	61.38
UAP (DeepFool)	5.2	14.25	16.71	16.07	7.15	11.73	18.48	6.07
UAPEPGD	79.63	81.17	81.13	81.71	74.3	80.04	80.4	79.19
Clean Accuracy	84.60	83.88	84.69	84.49	84.44	82.63	84.37	84.27
IAA Avg.	35.85 ↓57%	44.11↓47%	39.68↓53%	44.95↓47%	35.54↓58%	42.33↓49%	40.22↓52%	35.59↓58%
UAP Avg.	36.41 ↓57%	36.79↓56%	42.02↓50%	31.57↓63%	28.24↓67%	24.94↓70%	41.03↓51%	31.40↓63%
Adv Avg.	36.12 ↓57%	40.67↓52%	40.78↓52%	38.65↓54%	32.10↓62%	34.14↓59%	40.60↓52%	33.62↓60%

6.10.6. DTD

Table 31. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	68.51	67.77	69.20	65.11	61.76	65.74	69.15	68.56
FFF (mean-std)	67.87	64.95	69.52	66.54	56.97	52.34	67.50	68.72
FFF (no-data)	57.18	57.13	46.91	59.52	57.39	46.22	58.56	51.97
FFF (one-sample)	40.16	54.63	51.86	57.29	49.95	47.29	60.16	38.56
FG-UAP	19.10	21.54	13.30	21.28	24.84	18.94	13.46	17.23
GD-UAP (mean-std)	64.47	61.91	69.68	65.27	52.82	57.18	63.67	66.33
GD-UAP (no-data)	57.13	59.47	46.54	58.62	57.23	50.48	56.86	56.22
GD-UAP (one-sample)	57.66	52.29	42.55	48.94	45.64	47.02	47.71	55.64
L4A-base	72.61	73.35	72.13	73.51	64.26	68.14	71.01	72.66
L4A-fuse	72.55	72.71	72.29	73.40	64.20	68.30	71.38	72.93
L4A-ugs	71.60	70.69	72.39	73.30	66.65	71.12	72.93	72.87
PD-UAP	72.13	69.79	65.74	70.90	60.11	60.64	63.35	73.24
SSP	72.45	70.21	73.03	71.70	59.73	66.17	69.63	72.18
STD	71.44	71.49	73.30	72.39	69.73	71.54	72.93	72.18
UAP (DeepFool)	68.03	66.49	69.47	67.71	63.09	62.77	67.93	68.56
UAPEPGD	70.96	69.41	72.23	71.76	66.91	70.59	72.39	72.77
Clean Accuracy	76.12	76.28	78.09	75.43	75.90	73.30	74.47	77.29
IAA Avg.	41.42 ↓46%	41.89↓45%	40.50↓48%	42.91↓43%	42.42↓44%	38.99↓47%	39.65↓47%	41.90↓46%
UAP Avg.	62.74 ↓18%	62.74↓18%	61.26↓22%	63.58↓16%	57.58↓24%	57.78↓21%	62.41↓16%	62.54↓19%
Adv Avg.	51.46 ↓32%	51.70↓32%	50.27↓36%	52.64↓30%	49.56↓35%	47.83↓35%	50.36↓32%	51.62↓33%

6.10.7. Flowers

Table 32. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	69.45	77.27	69.56	82.40	75.33	77.53	63.57	73.27
FFF (mean-std)	67.79	90.99	54.47	71.41	61.67	77.85	56.39	57.47
FFF (no-data)	57.49	74.01	53.72	61.22	58.68	39.71	52.00	72.22
FFF (one-sample)	57.62	68.44	36.91	62.99	62.40	25.02	66.46	51.71
FG-UAP	14.02	17.87	2.56	13.45	6.25	10.20	4.10	11.32
GD-UAP (mean-std)	63.39	70.67	58.55	75.25	61.16	70.12	56.21	63.13
GD-UAP (no-data)	66.62	95.02	46.14	62.45	60.30	47.44	55.74	74.80
GD-UAP (one-sample)	30.40	14.74	26.38	41.49	59.61	11.85	22.44	40.30
L4A-base	67.70	81.02	81.03	75.93	56.36	70.48	64.16	72.19
L4A-fuse	68.34	83.02	81.59	76.85	56.86	71.46	67.54	72.42
L4A-ugs	84.09	84.89	63.96	88.48	75.71	85.65	81.73	84.32
PD-UAP	81.61	87.36	47.50	77.88	70.19	71.09	52.30	78.08
SSP	55.77	43.54	62.99	77.63	63.71	43.64	64.75	65.42
STD	69.43	74.60	66.85	71.59	67.20	69.42	67.86	67.67
UAP (DeepFool)	30.68	30.28	31.87	61.57	40.14	38.88	13.47	37.46
UAPEPGD	86.74	90.21	82.86	91.61	85.85	90.96	83.43	87.68
Clean Accuracy	97.41	96.74	97.18	96.70	95.18	96.73	96.68	96.86
IAA Avg.	60.24 ↓38%	61.32↓37%	54.30↓44%	60.98↓37%	51.34↓46%	60.79↓37%	55.66↓42%	59.97↓38%
UAP Avg.	60.69 ↓38%	67.75↓30%	54.18↓44%	68.26↓29%	60.09↓37%	56.33↓42%	54.51↓44%	63.09↓35%
Adv Avg.	60.45 ↓38%	64.34↓33%	54.25↓44%	64.41↓33%	55.46↓42%	58.69↓39%	55.12↓43%	61.44↓37%

6.10.8. Food

Table 33. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	66.85	72.56	76.44	67.11	63.10	68.71	73.28	68.88
FFF (mean-std)	56.35	64.24	72.43	62.39	46.44	21.16	70.21	63.32
FFF (no-data)	27.32	42.11	34.44	43.99	30.61	24.10	35.49	36.72
FFF (one-sample)	16.11	20.92	49.03	22.00	11.71	6.63	23.65	18.35
FG-UAP	3.60	3.51	5.25	3.18	2.35	5.27	5.69	3.48
GD-UAP (mean-std)	51.73	47.56	64.19	69.35	51.11	34.44	68.26	66.41
GD-UAP (no-data)	20.73	49.44	30.21	45.98	25.59	15.42	25.43	39.66
GD-UAP (one-sample)	14.78	18.09	24.88	41.16	22.17	7.45	7.39	31.94
L4A-base	34.63	40.48	42.81	32.70	26.26	34.83	18.31	31.63
L4A-fuse	33.93	40.76	42.48	33.41	24.73	34.40	17.74	24.70
L4A-ugs	59.73	67.54	72.25	60.77	10.41	35.42	58.80	34.90
PD-UAP	28.06	67.88	57.45	62.27	65.10	34.61	36.81	75.30
SSP	18.81	25.50	25.47	32.13	8.99	6.28	32.83	21.57
STD	57.58	55.80	60.02	50.13	42.12	52.72	61.46	55.82
UAP (DeepFool)	7.20	7.26	12.26	10.01	13.10	5.28	8.71	8.13
UAPEPGD	77.46	80.01	81.65	81.49	73.45	78.37	81.32	78.15
Clean Accuracy	83.93	85.63	87.65	85.82	82.30	84.35	87.16	83.71
IAA Avg.	27.80 ↓67%	36.30↓58%	34.12↓61.06%	36.09↓58%	28.89↓65%	34.74↓59%	34.26↓59%	31.71↓62%
UAP Avg.	35.93 ↓57%	43.98↓49%	46.95↓46%	44.88↓48%	32.33↓61%	29.07↓66%	39.09↓55%	41.18↓51%
Adv Avg.	35.94 ↓57%	39.91↓53%	40.16↓54%	40.23↓53%	30.51↓63%	32.07↓62%	36.53↓58%	36.17↓57%

6.10.9. Pets

Table 34. This table presents the results of various instance and universal adversarial perturbation (UAP) attacks on the AirCraft dataset, with all UAP attack names in *italics*. Different configurations of FGSM and PGD are denoted, such as $FGSM_1$ and PGD_1. Average results for universal adversarial perturbations (UAP Avg.), instance adversarial attacks (IAA Avg.), and overall adversarial performance (Adv Avg.) are reported at the bottom, including percentage drops relative to clean accuracy.

	Barlow	BYOL	DINO	MoCoV3	SimCLR	Supervised	SwAV	VICReg
$FGSM_1$	8.92	5.94	4.84	11.41	2.7	2.58	3.64	8.86
$FGSM_2$	1.52	0.69	0.45	1.95	0.78	0.81	2.57	1.8
PGD_1	10.03	5.72	4.54	10.96	3.44	1.61	4	10.18
PGD_2	0.06	0	0	0.12	0.24	0.18	0.64	0.06
PGD_3	10.27	6.02	4.63	11.09	3.27	1.61	3.83	10.06
PGD_4	0.06	0	0	0.12	0.18	0.12	0.61	0.06
PGD_5	0.12	0.03	0	0.24	0.18	0.24	0.79	0.12
DIFGSM	24.56	24.16	20.83	28.01	19.39	19.43	16.74	27.41
CW	0	0	0	0	0	0	0	0
Jitter	45.87	44.28	48.39	45.42	37.43	31.98	43.75	44.73
TIFGSM	32.78	31.08	29.68	35.76	28.31	18.99	29.83	33.04
PIFGSM	3.62	2.1	1.62	4.46	0.9	0.6	1.71	3.44
EADEN	0	0	0	0	0	0	0	0
OnePixel	51.75	49.39	54.93	53.41	41.4	36.01	47.55	51.54
Pixle	3.67	1.9	2.17	6.16	2.8	1.48	2.26	3.8
SPSA	44.36	42.91	44.2	46.6	30.76	28.51	38.42	44.31
Square	0.03	0	0	0.03	0.03	0	0	0.03
TAP	55.53	53.4	58.55	57.72	42.93	32.54	52.48	55.35
ASV	86.65	87.60	86.78	85.55	78.08	87.31	86.05	86.12
FFF (mean-std)	84.69	84.25	79.83	85.91	73.65	85.76	68.35	83.45
FFF (no-data)	48.95	75.83	47.83	74.82	66.53	70.05	73.33	64.11
FFF (one-sample)	26.50	79.70	45.52	68.01	50.15	74.97	54.46	29.36
FG-UAP	8.10	6.85	7.12	13.08	25.21	6.23	5.79	5.42
GD-UAP (mean-std)	82.46	85.03	76.63	85.23	66.49	83.99	81.65	84.72
GD-UAP (no-data)	65.76	68.77	53.73	66.01	72.91	75.24	76.37	62.12
GD-UAP (one-sample)	34.98	28.91	16.56	50.63	52.54	55.38	25.65	31.27
L4A-base	83.76	81.41	83.84	82.30	72.93	88.88	58.86	83.19
L4A-fuse	83.03	81.08	83.04	82.39	72.59	88.36	57.81	82.03
L4A-ugs	88.82	88.76	86.91	90.43	77.88	90.76	87.71	88.77
PD-UAP	86.63	81.05	63.30	79.72	74.99	76.75	74.91	86.63
SSP	79.54	81.60	83.18	82.57	72.05	84.12	70.15	76.15
STD	86.94	86.94	81.23	89.15	79.74	90.81	84.33	86.63
UAP (DeepFool)	40.81	44.95	38.77	44.69	69.07	60.11	31.96	43.56
UAPEPGD	88.68	88.69	87.07	90.78	81.50	91.88	87.52	88.64
Clean Accuracy	90.89	91.34	90.24	92.19	88.51	93.94	90.50	90.92
IAA Avg.	43.21 ↓52%	44.69↓51%	40.01↓56%	48.33↓48%	42.03↓53%	46.53↓50%	39.53↓56%	43.76↓52%
UAP Avg.	67.27 ↓26%	71.96↓21%	63.83↓29%	73.20↓21%	67.90↓23%	75.66↓19%	64.06↓29%	67.63↓26%
Adv Avg.	54.53 ↓40%	57.52↓37%	51.23↓43%	60.03↓35%	54.20↓39%	60.24↓36%	51.07↓44%	54.99↓40%