A Defense Methods in Literature

There are significant gaps in literature for defense performance on various poisons. We surveyed multiple works on poison attacks [26,32,40,67,79,83] and poison defenses [53,62] for data on attack/defense evaluations. Results for CIFAR-10 are displayed in Table 7 and results for ImageNet-100 are displayed in Table 8. On CIFAR-10, there is a dearth of data available for the CUDA poison and for SSL as a defense. On ImageNet-100, the lack of data is more severe; CP8, CUDA, and OPS have not been evaluated on most defenses and SSL has not been evaluated across most poisons.

Table 7: CIFAR-10 poison defense in literature. Individual entries list where data for the corresponding poison and defense method can be found. Empty entries indicate no data found.

Poison	SL+AT	Cutout [21]	Mixup [87]	CutMix [85]	ISS [53]	SSL
AP8 [26]	[26, 27, 53, 62, 67]	[26, 53, 62]	[26, 53, 62]	[26, 53, 62]	[53, 62]	[32]
CP8 [32]	[32]	[32]	-	-	[62]	[32, 62]
CUDA [67]	[67]	-	-	-	-	-
LSP16 [83]	[53, 62]	[53, 62, 83]	[53, 62, 83]	[53, 62, 83]	[53, 62]	-
OPS [79]	[53, 62, 79]	[53, 62, 79]	[53, 62, 79]	[53, 62]	[53, 62]	-
RUE8 [27]	[27, 53, 62, 67]	[53, 62]	[53, 62]	[53, 62]	[53, 62]	-
UE8 [40]	[27, 53, 62, 67, 79]	[53, 62, 79, 83]	[53, 62, 79, 83]	[53, 62, 83]	[53, 62]	[32]

Table 8: ImageNet-100 poison defense in literature. Individual entries list where data for the corresponding poison and defense method can be found. Empty entries indicate no data found.

Poison	SL+AT	Cutout [21]	Mixup [87]	CutMix [85]	ISS [53]	SSL
AP8 [26]	[27, 53, 67]	[53]	[53]	[53]	[53]	-
CP8 [32]	-	-	-	-	-	[32]
CUDA [67]	[67]	-	-	-	-	-
LSP16 [83]	[53, 62]	[53, 62]	[53, 62]	[53, 62]	[53, 62]	-
OPS [79]	-	-	-	-	-	-
RUE8 [27]	[27, 53, 62, 67]	[53, 62]	[53, 62]	[53, 62]	[53, 62]	-
UE8 [40]	[27, 53, 62, 67]	[53, 62]	[53, 62]	[53, 62]	[53, 62]	-

B Experiment Setup

B.1 Datasets

We primarily test our methods on a subset of the ImageNet dataset [20], utilizing the first 100 classes for a total of 130,000 training images and 5,000 test images. We consider ImageNet images to be representative of real-world images. For benchmarking purposes, we additionally measure performance on the CIFAR-10 dataset [48], with a total of 50,000 training images and 10,000 test images.

We verify against seven popular poisons. When available, we generate poisons using the code presented by corresponding authors; otherwise we attempt to follow the poison formulation as closely as possible. Specifically, we utilize Adversarial Poisons[3] [26] with 8/255 perturbations, Contrastive Poisoning[4] [32] with 8/255 perturbations, Convolution-based Unlearnable Datasets[5] [67], Linearly Separable Poisons[6] [83] with 16/255 perturbations, One-Pixel Shortcut[7] [79], Robust Unlearnable Examples [27] with 8/255 perturbations, and Unlearnable Examples[8] [40] with 8/255 perturbations. Due to the lengthy optimization process required to generate some poisons on large images, AP8, CP8, and UE8 poisons for ImageNet-100 are generated as 20% datasets, containing 100 classes with a total of 26,000 training images. Though RUE8 generation also requires lengthy optimization, the full RUE8 ImageNet-100 dataset is provided by the RUE paper authors [27].

With the exception of CP8, all poisons studied were designed for SL. The CP8 poison follows the 'unlearnable examples' algorithm [40] but utilizes SSL architectures and losses [12,31,34]. Therefore, the CP8 poison is designed specifically for SSL. Our implementation of CP8 generates sample-wise poisons with SimCLR architecture and loss [12].

B.2 Implementation

For all experiments, we employ a ResNet18 encoder [35] with an encoding dimension of 512. When running experiments on CIFAR-10, we follow other CIFAR ResNet models [17] by modifying the first convolutional layer to use smaller kernel and stride, and omitting the first maxpool layer. We also implement a momentum encoder for BYOL [31] with a momentum updates rate of 0.999. All SSL methods utilize a 3-layer MLP projection network following the MoCo projector structure [34] with hidden and output dimensions of 2048. For SSL methods that utilize a prediction head (e.g., BYOL, SimSiam), we follow SimSiam [14] and use a 2-layer MLP with a hidden dimension of 512 and an output dimension of 2048. For experiments with a classifier, we attach a single linear layer to the encoder output.

[3] https://github.com/lhfowl/adversarial_poisons
[4] https://github.com/kaiwenzha/contrastive-poisoning
[5] https://github.com/vinusankars/Convolution-based-Unlearnability
[6] https://github.com/dayu11/Availability-Attacks-Create-Shortcuts
[7] https://github.com/cychomatica/One-Pixel-Shotcut
[8] https://github.com/HanxunH/Unlearnable-Examples

B.3 Loss Formulation

Experiments with BYOL and SimSiam utilize cosine similarity loss with normalized projections. Experiments with SimCLR utilize the InfoNCE loss [60] with normalized projections and a temperature of 0.2. All SSL losses are symmetrized. Classifiers (for linear probe and SL) utilize the cross-entropy loss. SSL+SL combined losses are calculated as in Eq. 3, with $\alpha = \beta = 0.5$.

B.4 Augmentations

The image augmentation procedure follows SimSiam [14]. SSL and SSL+SL pretrain augmentations involve random-resized-crop, color-jitter, grayscale, Gaussian blur, horizontal flip, and normalization. Linear probe augmentations involve random-resized-crop, horizontal flip, and normalization. SL augmentations for ImageNet-100 are identical to those for linear probe, though CIFAR-10 SL augmentations utilize random-crop instead. Test augmentations involve resize, center crop, and normalization. For ImageNet-100, we use a resize value of 256 and a crop size of 224. For CIFAR-10, we maintain the same resize/crop ratio as ImageNet, using a resize value of 32 and a crop size of 28.

Table 9: Training settings for various methods. For cells with two values, the left value is for CIFAR-10 and the right value is for ImageNet-100.

Setting	SSL+SL	SSL Pretrain	SSL LinProbe	SL
Augmentations	Pretrain	Pretrain	LinProbe	LinProbe
Epochs	400 | 200	400 | 200	100	100
Batch Size	512 | 128	512 | 128	512	256
LR x BatchSize/256	1.0 | 0.25	1.0 | 0.25	1.0 | 10.0	0.1
LR Warmup	10	10	0	0
LR Decay	Cosine	Cosine	Step (0.2x at 60,75,90)	Cosine
Optimizer	SGD	SGD	SGD	SGD
Momentum	0.9	0.9	0.9	0.9
Weight Decay	1e-4	1e-4	0	5e-4

B.5 Training Procedure

We present training settings for different training methods in Table 9. We conduct all experiments by first training on poisoned data and then evaluating on clean test data. For SL, SSL+SL, and VESPR methods, we simply evaluate the accuracy of the classifier head. For SSL methods, we evaluate the accuracy of a classifier head finetuned on poisoned data (i.e., linear probe with frozen encoder). Unless otherwise noted, all SSL, SSL+SL, and VESPR methods utilize the SimCLR [12] contrastive learning framework. Experiments with SSL largely follow the settings detailed in He et al. [32]. Experiments with SL methods follow the settings given by Liu et al. [53].

B.6 Adversarial Training

Adversarial training largely follows the procedure given in [53]. We use Projected Gradient Descent (PGD) [49,57] to generate L_∞-bounded perturbations to input augmentations. We employ 10 PGD steps of size $0.6/255$, a bound at $4/255$, random start, and zero restarts. For AT with SSL, we closely follow the Adversarial Contrastive Learning algorithm [42], specifically their A2S method where one augmentation is adversarially perturbed and the other is kept as is. We do not enforce separate batch norms for the adversarial and standard augmentations. Adversarial parameters for SSL are kept identical to those of AT for SL.

C Additional Results

C.1 Augmentation-Based Defense Comparison

In Table 10, we concatenate results from Table 2 and Table 4 to better compare the performance of augmentation-based defenses for SL. SL with SSL augmentations achieves highest minimum and average clean test accuracies across poisons, outperforming other commonly-used defenses.

Table 10: ImageNet-100 clean test set accuracy across augmentation-based defenses: Cutout, Mixup, CutMix, ISS, and SL with SSL augmentations.

Poison	SL	Cutout [21]	Mixup [87]	CutMix [85]	ISS [53]	SL (SSL Aug)
Clean	77.66	78.04	80.38	**81.08**	71.58	75.48
AP8 [26]	8.18	7.68	9.68	8.38	**24.22**	16.54
CP8 [32]	57.46	58.20	61.76	**62.28**	50.18	60.40
CUDA [67]	6.10	8.66	5.16	5.70	3.58	**17.30**
LSP16 [83]	6.62	5.06	5.50	7.84	33.70	**49.30**
OPS [79]	51.00	53.86	48.82	**65.12**	57.36	63.68
RUE8 [27]	14.18	15.60	33.08	16.10	**67.52**	49.08
UE8 [40]	6.32	6.42	12.38	5.80	**41.18**	34.46
Psn Min	6.10	5.06	5.16	5.70	3.58	**16.54**
Psn Avg	21.41	22.21	25.20	24.46	39.68	**41.54**

C.2 CIFAR-10 Architecture Ablation

Table 11 displays clean test accuracies for VESPR architecture ablations trained on various CIFAR-10 poisons. We find that the trends in poison defense for CIFAR-10 typically mirror those seen for ImageNet-100 in Table 4, though with smaller poison-clean discrepancies.

Table 11: CIFAR-10 clean test set accuracies for VESPR and multiple ablations to architecture and training procedure.

Poison	SL	SL+AT	SSL [12]	SSL+SL	VESPR
Clean	93.86	89.57	90.55	**94.45**	89.35
AP8 [26]	12.45	85.67	75.77	77.23	**86.78**
CP8 [32]	**93.59**	88.43	73.08	92.20	89.20
CUDA [67]	21.89	48.58	66.58	67.80	**80.31**
LSP16 [83]	21.47	85.67	88.87	**90.67**	88.96
OPS [79]	27.24	20.10	87.79	**91.88**	84.74
RUE8 [27]	18.91	34.45	86.70	**88.47**	87.82
UE8 [40]	32.53	89.31	88.45	**91.54**	89.39
Psn Min	18.91	20.10	66.58	67.80	**80.31**
Psn Avg	32.58	64.60	81.03	85.68	**86.74**

C.3 ImageNet-100 Classifier Ablations

Table 12 quantifies clean test classification performance of VESPR and multiple architecture ablations using a KNN classifier. This is similar to Table 4 but with a classification method that does **not** rely on a trained classifier network. To classify a single clean test image, rank the proximity (cosine similarity) of the clean image encoding to 10,240 randomly-sampled poisoned image encodings. The clean image class is decided its 20 nearest neighbors.

Intuitively, KNN accuracy is a distillation of the Psn In-Cls Sim and Psn-Cln Sim metrics from Tables 1 and 5. If poison image encodings are tightly clustered (high Psn In-Cls Sim) and well-aligned with their clean counterparts (high Psn-Cln Sim), then KNN accuracy should be high. As expected, the KNN accuracy results of Table 12 shadow those of linear classifier accuracy (Table 4), and VESPR achieves highest Psn Min and Psn Avg performance. This further demonstrates that VESPR's image encodings are robust to poison features across multiple poisoning methods and supports the analyses from Sec. 4.3.

We also ablate the effect of training the VESPR classifier simultaneously with SSL. To this end, we freeze a VESPR-trained encoder and train a new classifier via linear probe on poisoned data with the same settings as SSL linear probe (9). Table 13 shows that the VESPR encoder with a linear-probe classifier

Table 12: ImageNet-100 KNN classifier accuracies for VESPR and multiple ablations. Bold numbers indicate the best defense method for each poison, while underlined numbers indicate second-best methods.

Poison	SL	SL (SSL Aug)	SL+AT	SSL [12]	SSL+AT [42]	SSL+SL	VESPR
Clean	70.31	<u>71.28</u>	68.79	56.48	41.88	**78.20**	69.14
AP8 [26]	9.80	10.86	**47.50**	35.16	24.84	28.48	<u>44.96</u>
CP8 [32]	51.33	**56.56**	52.07	5.20	28.98	<u>56.02</u>	50.47
CUDA [67]	5.12	17.23	<u>30.08</u>	23.98	21.56	20.70	**41.29**
LSP16 [83]	11.80	44.10	23.67	47.42	32.77	**52.27**	<u>49.96</u>
OPS [79]	50.12	<u>60.70</u>	45.43	51.17	38.01	**68.79**	55.66
RUE8 [27]	13.32	41.91	53.40	53.32	40.00	**68.16**	<u>65.90</u>
UE8 [40]	7.27	27.07	<u>46.64</u>	36.80	25.63	**55.74**	44.45
Psn Min	5.12	10.86	<u>23.67</u>	5.20	21.56	20.70	**41.29**
Psn Avg	21.25	36.92	42.68	36.15	30.26	<u>50.02</u>	**50.38**

(VESPR+LPC) performs worse than the original VESPR classifier. This is expected; though Tables 5 and 12 demonstrate that VESPR encodings are robust, training a new classifier with poison data allows poison features to determine classifier decision boundaries. Surprisingly, VESPR+LPC also underperforms SSL (which uses poison linear probe). We hypothesize that this is due to higher feature expressivity in SSL (see Psn E-Rank in Table 5) which may inhibit poison shortcut learning in the classifier. This ablation demonstrates that the VESPR classifier is imperative for poison robustness. Further research on VESPR's feature expressivity is required; possible solutions include uniformity regularization and utilizing other SSL methods (e.g., VICReg [5], DINO [7]) within VESPR.

Table 13: ImageNet-100 clean test set accuracies for SSL, VESPR, and VESPR encoder with linear probe classifier. Bold values highlight the top-performing methods.

Poison	SSL	VESPR	VESPR+LPC
Clean	70.96	74.84	69.74
AP8 [26]	41.80	**48.68**	44.82
CP8 [32]	14.80	**54.40**	50.88
CUDA [67]	26.12	**45.20**	20.50
LSP16 [83]	**62.06**	56.34	48.14
OPS [79]	62.22	**63.26**	55.98
RUE8 [27]	65.30	**70.36**	59.80
UE8 [40]	**50.02**	49.34	44.70
Psn Min	14.80	**45.20**	20.50
Psn Avg	46.05	**55.37**	46.40

C.4 SSL Method Ablation

In Table 14, we demonstrate the performance of multiple SSL methods. Methods that utilize momentum encoders (MoCo and BYOL) tend to outperform methods that do not (SimCLR and SimSiam). There seems to be no major difference in performance between contrastive methods (SimCLR and MoCo) and non-contrastive methods (SimSiam and BYOL), though contrastive methods tend to give better defense against AP8 and UE8. Relative to pretraining on clean data, all SSL methods perform poorly on multiple poisons, particularly CP8 and CUDA. Unsurprisingly, all SSL methods are weakest against CP8, the poison designed specifically to counter SSL.

Furthermore, we investigate the effect of using different SSL techniques within VESPR. We note that all VESPR variants give similar performance. This is encouraging, as it suggests that the VESPR framework is robust across different SSL techniques. Compared with SSL methods, all VESPR variants provide superior performance on AP8, CP8, and CUDA. Notably, VESPR with BYOL boosts clean test accuracy after CUDA training to 50.14%, which is the highest CUDA defense value yet seen. All VESPR variants are also resistant to the SSL-based poison, CP8. All VESPR variants boost Psn Min at least 20% higher than any SSL method. VESPR with SimCLR achieves the highest Psn Avg performance, with an 8% improvement over the best SSL method (BYOL) and a 0.5% improvement over the second-best VESPR variant (BYOL).

Table 14: ImageNet-100 clean test set accuracies for various SSL and VESPR methods. Underlined and bolded values highlight the top-performing SSL and VESPR methods, respectively.

Poison	SimCLR [12]	MoCo [15]	SimSiam [14]	BYOL [31]	VESPR SimCLR	VESPR MoCo	VESPR SimSiam	VESPR BYOL
Clean	70.96	75.00	70.68	75.34	74.84	74.84	74.78	**75.02**
AP8 [26]	41.80	43.44	37.88	33.84	48.68	47.74	**48.82**	48.50
CP8 [32]	14.80	15.80	12.82	23.64	54.40	54.20	54.26	**54.80**
CUDA [67]	26.12	29.84	27.62	31.44	45.20	44.88	45.94	**50.14**
LSP16 [83]	62.06	67.10	61.98	67.04	56.34	**56.94**	51.20	50.22
OPS [79]	62.22	68.04	64.54	69.00	**63.26**	63.08	59.50	61.70
RUE8 [27]	65.30	70.38	66.84	70.40	70.36	**70.56**	70.04	70.34
UE8 [40]	50.02	51.04	43.92	37.64	**49.34**	48.54	48.12	48.58
Psn Min	14.80	15.80	12.82	23.64	45.20	44.88	45.94	**48.50**
Psn Avg	46.05	49.38	45.09	47.57	**55.37**	55.13	53.98	54.90

C.5 VESPR Loss Weight Ablation

Fig. 4 ablates the value of α from Eq. 3, effectively varying the relative importance of contrastive loss in VESPR. The set of tested α values is $\{0.05,\ 0.25,$

0.50, 1.00, 5.00}. To ensure that the classifier head is adequately trained, β is held constant at 0.5 throughout the ablations. The dataset is ImageNet-100. Poison robustness is largely consistent for all values of α, though it often decreases slightly for excessively large values of α. Poison defense is usually highest for moderate α ($\{0.25, 0.50, 1.00\}$).

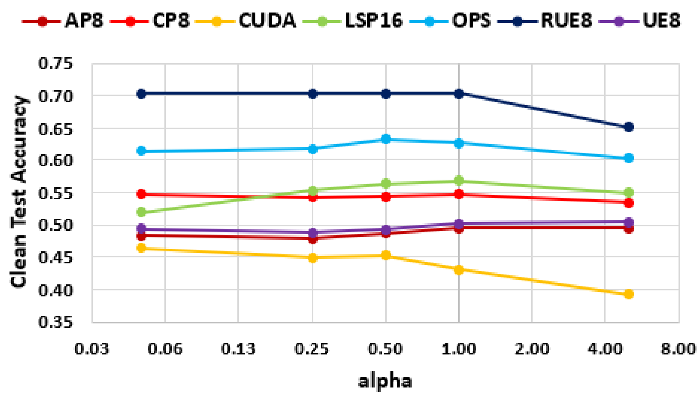

Fig. 4: Ablation of the α weight from Eq. 3. β is held constant at 0.5. The dataset is ImageNet-100.

C.6 t-SNE Graphs

Figure 5 displays additional t-SNE plots across multiple training methods for both AP8 and CUDA. We note that, across training methods, class clusters for AP8 tend to be more diffuse than those of CUDA, suggesting that CUDA perturbations enforce strong class-based signals. In particular, CUDA enforces strong class separation for SL, SSL, and SSL+SL, with clean representations forming their own out-of-distribution cluster. Alignment between clean representations and poison representations is generally higher for AP8-poisoned encoders. For both poisons, adding AT (e.g., SL+AT or VESPR) greatly improves the alignment between poison and clean representations. VESPR consistently enforces high poison-clean alignment as well as reducing class cluster separation.

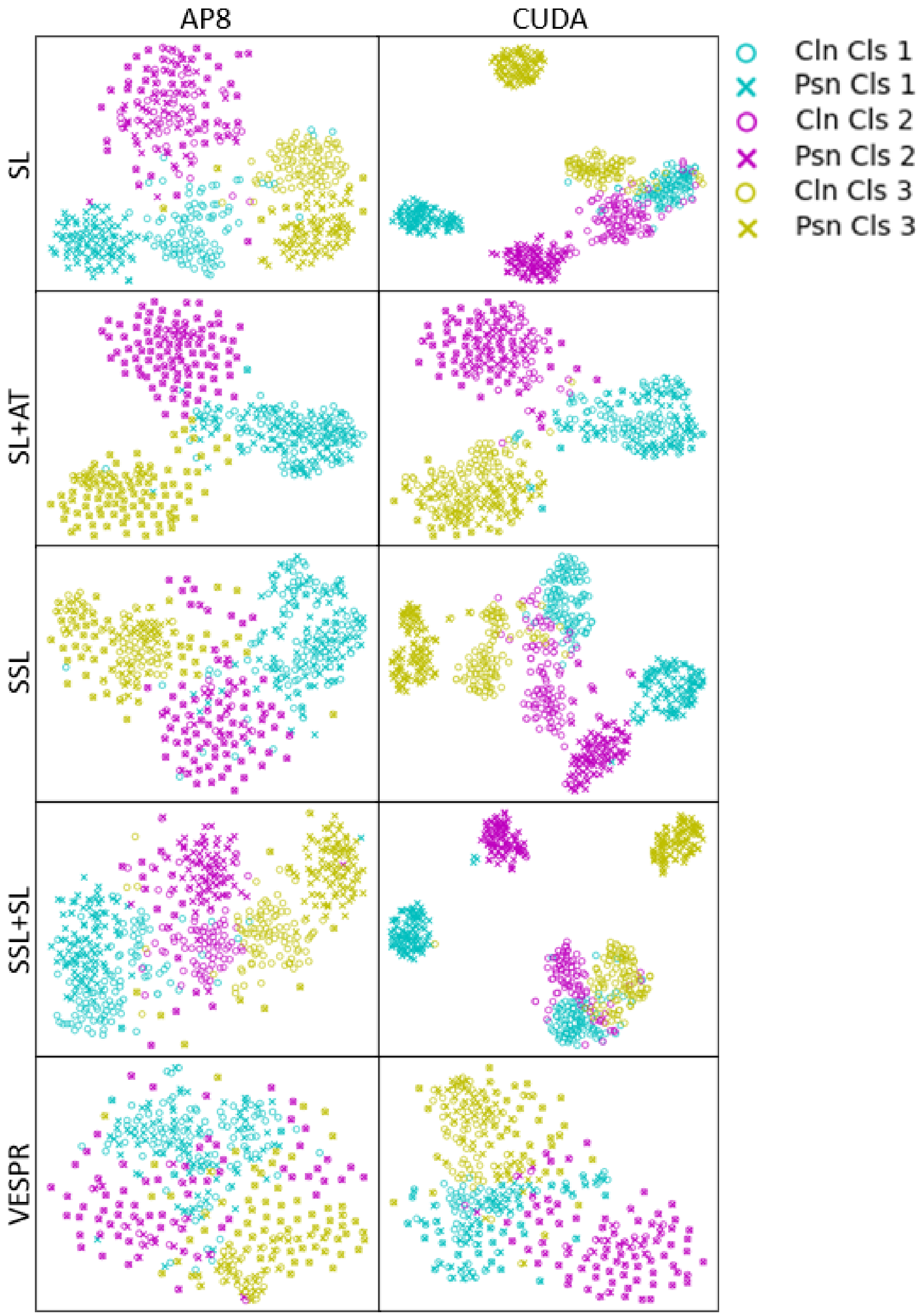

Fig. 5: 2-dimensional t-SNE plots [38] of clean and poison image representations for SL, SL+AT, SSL, SSL+SL, and VESPR models trained on AP8 and CUDA data. Color schemes denote different image classes.