Defending Against Patch-based Backdoor Attacks on Self-Supervised Learning (Supplementary Material)

Ajinkya Tejankar [*1] Maziar Sanjabi [2] Qifan Wang [2] Sinong Wang [2] Hamed Firooz [2]
Hamed Pirsiavash [1] Liang Tan [2]
[1] University of California, Davis [2] Meta AI

A. Appendix

A.1. Additional Ablations

We consider the effect of number of images scored by iterative search in Table A1 and Figure A1. Number of scored images can be changed by varying hyperparameters, l (number of clusters), s (number of samples scored per cluster) and r (percentage of least poisonous clusters removed in each iteration). The results show generally good performance when processing more than 6% of samples. Next, we consider the effect of varying the size of flip test set X^f in Table A2. We find that we can get reasonable results with $|X^f|$ as small as 316. Next, we evaluate intermediate checkpoints of one of our models to understand the relationship between overall performance of the model (measured with clean data Acc) and the attack effectiveness (measured with patched data FP) in Table A3. We can see that attack effectiveness improves as the overall model performance improves in the early epochs. In later epochs, the attack effectiveness fluctuates but is still high relative to earlier epochs. This indicates that it is possible to reduce the compute requirements for the model used during defense. We only need to train this model until the attack effectiveness becomes broadly comparable to a fully trained model. Next, we show that our defense is effective even when changing datasets in Table A4. Next, we consider the effect of varying top-k in Table A5, and find that any top-k $<= 20$ is a good choice. Next, we consider changing the trigger size w during the attack in Table A6. We find that our defense works despite changing w from 50 to 75 and 100. Next, we consider attacking images from a downstream dataset instead of pre-training dataset. We use Food101 [6] classification dataset for this purpose. It consists 101 fine-grained food categories and 750 images per category. For each category, we randomly select 700 images for training and 50 images for validation. We backdoor BYOL, ResNet-18 models by picking random category from Food101, poisoning all of its 700 training images, and adding them to the pre-training dataset (ImageNet-100). The resulting pre-trained model is evaluated on the task of Food101 classification. We use the same linear evaluation setup as other ResNet-18 models in our experiments. We use the training set of Food101 (700×101) for training the linear layer while the evaluation set (50×101) is used for evaluation. The results are presented in Table A7 We find that the attack is successful, but the models can also be successfully defended with PatchSearch.

A.2. Implementation Details

Below, we describe implementation details for various settings. Note that running PatchSearch is relatively inexpensive compared to the pre-training stage. For instance, with ImageNet-100 (126K images) and ViT-B on 4x3090 GPUs, PatchSearch takes 1.5 hrs, which is small as compared to the training time, about 16 hrs, for MoCo-v3.

Poison classifier training in PatchSearch. The training has following parameters: SGD (lr=0.01, batch size=32, max iterations=2000, weight decay=$1e^{-4}$, and cosine lr scheduler). The architecture of the classifier is ResNet-18 but each layer has only a single `BasicBlock` instead of the default two [1].

ResNet-18 model training. As mentioned in the main paper, MoCo-v2 and BYOL training for ResNet-18 models is exactly the same as code [2] from [41]. The models are trained on 4 NVIDIA A100 GPUs.

MoCo-v3 model training. We use the code [3] from MoCo-v3 [12] for training ViT-B models. We use the default hyperparameters from the code except SGD (batch size=1024, epochs=200). The models are trained on 8 NVIDIA A100 GPUs.

MAE model training. We use the code [4] from MAE [24] to train ViT-B models. All hyperparameters are unchanged except following: SGD (batch size=32 and accum

*Corresponding author <atejankar@ucdavis.edu>. Work done while interning at Meta AI.

[1] https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py
[2] https://github.com/UMBCvision/SSL-Backdoor
[3] https://github.com/facebookresearch/moco-v3
[4] https://github.com/facebookresearch/mae

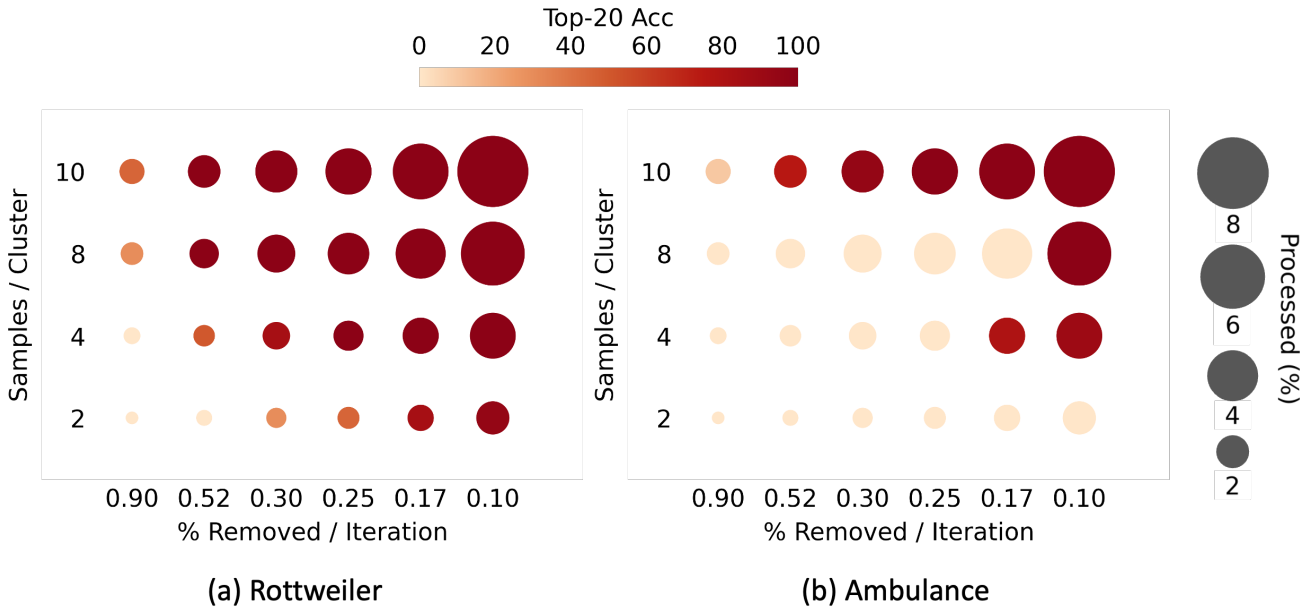

Figure A1. **Effect of number of images scored.** We vary different hyperparameters that control the number of images scored by PatchSearch and observe the effect on the accuracy of finding poisons in top-20 ranked images. The color of the circles denotes the accuracy while their size denotes the number of scored images. We find that a large value for samples per cluster (s) has better performance compared to small s values with comparable number of processed images. Finally, processing more than 6% of images generally results in a good model performance. See Table A1 for detailed results. Seeting: MoCo-v3, ViT-B, and poison rate 0.5%.

Num. Clusters	Samples Per Cluster	Pruned per iteration (%) 0.90	0.52	0.30	0.25	0.17	0.10
100	2	0.0 / 0.18	0.0 / 0.31	0.0 / 0.54	0.0 / 0.70	0.0 / 0.97	0.0 / 1.60
100	4	0.0 / 0.35	0.0 / 0.62	0.0 / 1.07	0.0 / 1.30	80.0 / 1.94	90.0 / 3.21
100	8	0.0 / 0.70	0.0 / 1.24	0.0 / 2.14	0.0 / 2.60	0.0 / 3.87	100.0 / 6.42
100	10	10.0 / 0.88	75.0 / 1.55	95.0 / 2.68	100.0 / 3.30	100.0 / 4.84	100.0 / 8.02
1000	2	20.0 / 1.80	45.0 / 3.00	75.0 / 5.30	80.0 / 6.40	100.0 / 9.40	100.0 / 16.00
1000	4	65.0 / 3.50	100.0 / 6.10	100.0 / 10.60	95.0 / 12.70	100.0 / 18.60	100.0 / 30.90
1000	8	100.0 / 7.00	100.0 / 12.20	100.0 / 21.00	100.0 / 25.00	100.0 / 35.40	100.0 / 52.30
1000	10	100.0 / 8.80	100.0 / 15.20	100.0 / 26.10	100.0 / 30.80	100.0 / 42.40	100.0 / 59.30

Table A1. **Effect of varying scored count.** We explore the effect of processing different amount of images by varying number of clusters, s (samples per cluster) and r (percentage of clusters pruned per iteration). Each table entry has the format *accuracy of finding poisons (%) in top-20 / percentage of training set scored*. We find that even with large r and a small number of clusters, increasing s can improve the model performance. Finally, we observe that processing more than 6% of the training set results in good model performance most of the times. Setting: MoCo-v3, ViT-B, poison rate 0.5%, target category Ambulance.

iter=4). Here, *accum iter* refers to the number iterations used for averaging the gradients before updating parameters. The models are trained on 8 NVIDIA A100 GPUs. Hence, the effective batch size is $32 \times 8 \times 4 = 1024$.

ResNet-18 linear evaluation. We use the linear layer training procedure proposed in code [5] from ComPress [4] for evaluating ResNet-18 models. A single linear layer is trained on top of a frozen backbone. The output of the backbone is processed according to following steps before passing it to the linear layer. (1) A mini-batch of features is normalized to have unit l_2 norm. (2) The mini-batch is also normalized to have zero mean and unit variance. Note that the mean and variance used for normalization in the second step comes from the l_2 normalized features of the

[5] https://github.com/UMBCvision/CompRess/blob/master/eval_linear.py

	Target Category	Flip test set size				
		10	32	100	316	1000
top-20 Acc (%)	Ambulance	5.0	20.0	60.0	80.0	80.0
	Rottweiler	55.0	100.0	100.0	100.0	100.0

Table A2. **Effect of flip test set size**. We explore the effect of changing the flip test set size $|X^f|$. The larger this set the more diverse samples will be used to obtain the poison score for an image. We find that PatchSearch is not greatly sensitive to this hyperparameter and even a small value like 316 could be effective. Setting: ViT-B, MoCo-v3, and poison rate 0.5%.

Checkpoint Epoch	Clean Data		Patched Data		Clean Data		Patched Data	
	Acc	FP	Acc	FP	Acc	FP	Acc	FP
	Rottweiler				*Ambulance*			
20	27.2	58.8	24.9	72.8	26.9	47.8	24.7	67.6
40	43.6	47.4	36.4	391.8	43.3	20.6	39.6	66.2
60	53.1	36.6	17.1	3145.0	53.0	14.0	46.5	284.6
80	58.3	32.8	35.5	1787.4	58.0	13.4	51.2	262.4
100	60.9	28.8	28.0	2689.6	61.4	12.6	48.0	1041.6
120	62.9	27.6	35.5	2118.4	63.1	12.4	51.6	804.2
140	64.2	30.0	40.7	1675.0	64.6	9.2	51.4	908.6
160	65.8	26.4	43.5	1420.4	66.1	9.8	54.8	704.6
180	66.3	25.8	40.9	1895.8	66.9	7.8	54.4	800.6
200	66.4	26.0	39.9	1926.8	66.8	7.6	53.6	895.6

Table A3. **Relationship between overall performance and attack effectiveness**. We evaluate intermediate checkpoints to understand the relationship between overall model performance (measured with *Clean Data Acc*) and the attack effectiveness (measured with *Patched Data FP*). We find that attack effectiveness has a strong correlation with overall model performance in early epochs ($<= 60$). While this correlation is not strict for later epochs, the attack effectiveness maintains its overall magnitude. This observation suggests that a model used during defense need not be trained to convergence. We only need to train it until the attack effectiveness is comparable to the fully trained model in overall magnitude. Setting: BYOL, ResNet-18, and poison rate 0.5%.

	Clean Data			Patched Data		
Model	Acc	FP	ASR	Acc	FP	ASR
Clean	79.2	205	2.1	54.3	741	7.4
Backdoored	79.0	206	2.1	26.5	6187	61.9
PatchSearch	78.7	224	2.2	57.0	816	8.2

Table A4. **CIFAR-10**. We show results for the attack and defense on CIFAR-10. We find that the defense is successfully able to mitigate the attack. In fact, we find that the poison classifier is not needed since the iterative search itself is sufficient. Simply removing the top 10% of the clusters removes 100% of poisons. To account for smaller, 32x32, images, we also reduce the size of the trigger to 8x8. Setting: ResNet-18, BYOL, poison rate 0.5% and target category Airplane.

	Acc (%) in top-k				
Model	5	10	20	50	100
ResNet-18, BYOL, 0.5%	100.0	100.0	99.5	93.8	84.0
ResNet-18, MoCo-v2, 0.5%	52.0	55.0	52.5	44.6	30.9
ViT-B, MoCo-v3, 0.5%	100.0	100.0	96.5	87.6	73.5
ViT-B, MoCo-v3, 1.0%	98.0	98.0	97.5	84.0	77.3

Table A5. **Effect of top-k in PatchSearch.** Note that PatchSearch is not able to find the patches for ResNet-18, MoCo-v2 for any k. Hence, we use an easy to backdoor model like ViT-B to defend it.

entire training dataset. All hyperparameters values are set to default values from the original code.

ViT-B linear evaluation. We use the linear layer training procedure from MoCo-v3 [12] code [6]. We set all hyperparameters to their default values from the code except SGD (epochs=30, batch size=256).

ViT-B fine-tuning evaluation. We use the code [7] from MAE [24] for fine-tuning ViT-B models. Strong augmentations like mixup, random erase, and cutmix are turned off

[6] https://github.com/facebookresearch/moco-v3/blob/main/main_lincls.py

[7] https://github.com/facebookresearch/mae/blob/main/engine_finetune.py

	Clean Data			Patched Data		
Model	Acc	FP	ASR	Acc	FP	ASR
$w = 75$						
Clean	65.7	28.8	0.6	55.0	19.0	0.4
Backdoored	66.6	32.4	0.7	36.4	1588.0	32.0
PatchSearch	66.5	29.2	0.6	55.8	25.2	0.5
$w = 100$						
Clean	65.7	28.8	0.6	45.6	15.2	0.3
Backdoored	65.8	30.2	0.6	20.4	2481.6	50.6
PatchSearch	65.6	32.8	0.7	46.7	34.6	0.7

Table A6. **Changing trigger size w in the attack.**. We explore the effect of changing the trigger size w during the attack. We find that our defense can successfully defend against different w by searching for most effective w during the defense (outlined in Figure 4). Note that since we are dealing with large w we paste the trigger anywhere in the image without any boundary. This is different from the default setting of 25% margin on all sides. Setting: ResNet-18, BYOL, poison rate 0.5% and target category Rottweiler.

Target Category	Model	Clean Data		Patched Data	
		Acc	FP	Acc	FP
Chicken Curry (10)	Clean	47.6	27	39.1	24
	Backdoored	48.2	19	9.2	3438
	PatchSearch	46.6	29	40.3	21
Steak (11)	Clean	47.6	17	39.5	24
	Backdoored	48.0	26	13.5	2596
	PatchSearch	47.4	22	40.7	57
Panna Cotta (12)	Clean	47.6	41	39.8	23
	Backdoored	47.5	42	19.1	2158
	PatchSearch	47.4	37	40.9	16
Deviled Eggs (13)	Clean	47.6	50	39.3	41
	Backdoored	47.9	23	7.9	4317
	PatchSearch	48.0	52	41.0	35
Mean	Clean	47.6	33.8	39.4	28.0
	Backdoored	47.9	27.5	12.4	3127.3
	PatchSearch	47.4	35.0	40.7	32.3

Table A7. **Attack on downstream task: Food101 classification**. Instead of attacking one of the categories in the pre-training dataset, we consider an attack on the downstream task of Food101 classification [6]. We poison 700 images of a category from Food101 and add them to the pre-training dataset (ImageNet-100). We evaluate the resulting models on Food101 classification and find that the targeted attack is successful. Next, we defend the backdoored models with PatchSearch and find that it is able to defend against the attack successfully. Averaged across the four categories, PatchSearch has 99.7% recall and 54.8% precision for filtering out poisons. Setting: BYOL & ResNet-18.

during fine-tuning since we find that their presence hurts the overall performance of the model (Table A8). For MoCo-v3

	Clean Data		Patched Data	
Model	Acc	FP	Acc	FP
with strong aug	63.1	21.0	53.5	63.1
without strong aug	65.7	18.7	53.8	97.6

Table A8. **Effect of strong augmentations during fine-tuning**. We find that strong augmentations like cutmix, mixup, and random erase can degrade the overall model performance. Setting: ViT-B, MAE and average of 4 target categories (Rottweiler, Tabby Cat, Ambulance, and Laptop).

models, fine-tuning runs for 30 epochs while for MAE models it runs for 90 epochs. Finally, MAE uses global pooling of tokens following the default option in the original code while MoCo-v3 models use `[CLS]` token. Rest of the hyperparameters are unchanged from their default values. For fine-tuning results on ImageNetx-1k reported in Table 6, all settings except epochs=50 are the same as their default values. Note that strong augmentations is kept on as in the default settings in order to be comparable to the numbers reported in MAE [24].

A.3. Per Category Results

We list per category results for Table 2. The results are in Tables A9, A10, A11, and A12.

Target				Clean Data		Patched Data	
Category	Model	PatchSearch	i-CutMix	Acc	FP	Acc	FP
Rottweiler (10)	clean	✗	✗	65.7	29.4	60.6	24.4
	clean	✗	✓	65.8	33.0	64.0	28.6
	backdoored	✗	✗	66.4	26.0	39.9	1926.8
	defended	✗	✓	66.8	26.4	60.4	272.0
	defended	✓	✗	66.9	31.8	61.1	43.8
	defended	✓	✓	65.6	32.6	63.4	32.2
Tabby Cat (11)	clean	✗	✗	65.7	5.0	60.8	9.4
	clean	✗	✓	65.8	4.4	63.8	3.4
	backdoored	✗	✗	65.9	1.6	31.9	2338.0
	defended	✗	✓	67.1	6.4	62.5	299.0
	defended	✓	✗	66.1	4.2	61.9	7.0
	defended	✓	✓	67.0	3.2	64.6	1.6
Ambulance (12)	clean	✗	✗	65.7	8.6	60.5	10.2
	clean	✗	✓	65.8	9.2	63.9	10.4
	backdoored	✗	✗	66.8	7.6	53.6	895.6
	defended	✗	✓	66.5	7.8	62.4	116.0
	defended	✓	✗	66.4	9.2	61.2	9.2
	defended	✓	✓	67.3	7.2	65.1	7.0
Pickup Truck (13)	clean	✗	✗	65.7	14.2	60.6	13.2
	clean	✗	✓	65.8	14.2	63.7	15.6
	backdoored	✗	✗	66.4	13.6	48.0	1430.6
	defended	✗	✓	68.3	16.8	64.1	131.0
	defended	✓	✗	65.7	14.8	60.8	13.0
	defended	✓	✓	67.1	19.0	64.6	18.8
Laptop (14)	clean	✗	✗	65.7	30.8	60.4	33.0
	clean	✗	✓	65.8	27.4	64.1	15.8
	backdoored	✗	✗	65.6	29.6	16.1	355.8
	defended	✗	✓	66.2	25.2	60.5	356.2
	defended	✓	✗	65.8	30.4	60.9	37.6
	defended	✓	✓	66.3	29.0	63.9	17.0
Goose (15)	clean	✗	✗	65.7	10.4	60.3	18.0
	clean	✗	✓	65.8	8.6	63.6	9.0
	backdoored	✗	✗	66.4	10.0	26.5	3391.0
	defended	✗	✓	67.4	11.2	62.4	288.2
	defended	✓	✗	66.4	12.2	61.6	23.6
	defended	✓	✓	67.0	25.6	65.1	24.8
Pirate Ship (16)	clean	✗	✗	65.7	4.2	60.4	4.6
	clean	✗	✓	65.8	5.4	63.5	6.0
	backdoored	✗	✗	66.9	4.6	38.3	2438.6
	defended	✗	✓	67.2	4.4	61.4	328.2
	defended	✓	✗	66.4	4.0	61.5	9.8
	defended	✓	✓	66.9	3.6	64.4	2.4
Gas Mask (17)	clean	✗	✗	65.7	19.8	60.7	28.6
	clean	✗	✓	65.8	22.0	63.8	36.4
	backdoored	✗	✗	66.6	21.2	34.8	2722.4
	defended	✗	✓	66.8	18.6	57.2	931.6
	defended	✓	✗	66.8	30.8	62.1	52.8
	defended	✓	✓	67.2	29.0	64.9	55.8
Vacuum Cleaner (18)	clean	✗	✗	65.7	60.0	60.6	74.8
	clean	✗	✓	65.8	58.6	63.4	38.2
	backdoored	✗	✗	66.6	49.2	20.4	3210.0
	defended	✗	✓	67.0	51.0	59.6	328.4
	defended	✓	✗	66.8	58.4	61.8	85.8
	defended	✓	✓	67.0	59.0	64.8	39.2
American Lobseter (19)	clean	✗	✗	65.7	24.0	61.0	31.4
	clean	✗	✓	65.8	21.8	63.5	32.6
	backdoored	✗	✗	67.0	22.2	43.5	2086.8
	defended	✗	✓	66.5	17.6	59.5	602.2
	defended	✓	✗	66.3	32.8	61.9	51.0
	defended	✓	✓	66.6	15.8	64.2	26.6
Mean and STD	clean	✗	✗	65.7 ± 0.0	20.6 ± 16.8	60.6 ± 0.2	24.8 ± 20.2
	clean	✗	✓	65.8 ± 0.0	20.5 ± 16.5	63.7 ± 0.2	19.6 ± 13.1
	backdoored	✗	✗	66.5 ± 0.4	18.6 ± 14.3	35.3 ± 11.9	2079.6 ± 967.5
	defended	✗	✓	67.0 ± 0.6	18.5 ± 13.7	61.0 ± 2.0	365.3 ± 239.4
	defended	✓	✗	66.4 ± 0.4	22.9 ± 17.1	61.5 ± 0.5	33.4 ± 25.6
	defended	✓	✓	66.8 ± 0.5	22.4 ± 16.8	64.5 ± 0.5	22.5 ± 17.1

Table A9. **Per category results for BYOL, ResNet-18, and poison rate 0.5%**. Detailed results for Table 2.

Target				Clean Data		Patched Data	
Category	Model	PatchSearch	i-CutMix	Acc	FP	Acc	FP
Rottweiler (10)	clean	✗	✗	49.7	36.8	46.3	28.4
	clean	✗	✓	55.9	32.0	54.0	29.4
	backdoored	✗	✗	50.2	39.4	33.4	1094.4
	defended	✗	✓	55.1	38.6	52.4	117.2
	defended	✓	✗	50.0	40.2	46.3	26.6
	defended	✓	✓	55.6	39.0	54.4	35.2
Tabby Cat (11)	clean	✗	✗	49.7	7.2	46.4	8.2
	clean	✗	✓	55.9	8.0	54.4	6.6
	backdoored	✗	✗	50.0	5.8	33.5	1901.6
	defended	✗	✓	55.3	5.6	52.4	181.8
	defended	✓	✗	49.9	11.6	46.7	13.8
	defended	✓	✓	55.7	8.8	54.2	9.0
Ambulance (12)	clean	✗	✗	49.7	18.4	46.4	15.0
	clean	✗	✓	55.9	13.6	54.4	19.2
	backdoored	✗	✗	50.3	14.0	46.2	103.2
	defended	✗	✓	55.4	10.6	53.8	27.6
	defended	✓	✗	49.0	14.4	45.4	15.0
	defended	✓	✓	55.7	16.4	54.1	15.8
Pickup Truck (13)	clean	✗	✗	49.7	16.6	46.6	15.4
	clean	✗	✓	55.9	16.6	54.2	18.6
	backdoored	✗	✗	50.6	13.0	46.4	115.0
	defended	✗	✓	55.8	15.2	53.8	75.2
	defended	✓	✗	49.1	18.2	45.8	17.8
	defended	✓	✓	55.5	16.6	53.9	18.6
Laptop (14)	clean	✗	✗	49.7	37.0	46.0	35.8
	clean	✗	✓	55.9	33.0	54.3	24.2
	backdoored	✗	✗	49.8	33.8	41.8	466.2
	defended	✗	✓	55.4	24.0	53.7	92.6
	defended	✓	✗	49.7	41.6	45.8	47.6
	defended	✓	✓	54.8	46.4	53.5	29.4
Goose (15)	clean	✗	✗	49.7	37.2	46.6	39.4
	clean	✗	✓	55.9	38.2	54.2	39.2
	backdoored	✗	✗	49.6	33.8	45.2	194.0
	defended	✗	✓	55.5	31.0	53.8	46.8
	defended	✓	✗	49.5	39.6	46.3	46.0
	defended	✓	✓	55.1	41.0	53.4	37.6
Pirate Ship (16)	clean	✗	✗	49.7	8.2	46.1	9.0
	clean	✗	✓	55.9	5.0	54.4	6.4
	backdoored	✗	✗	49.7	6.2	42.1	573.4
	defended	✗	✓	55.7	4.6	53.6	68.8
	defended	✓	✗	49.7	12.8	46.8	17.4
	defended	✓	✓	55.0	5.6	53.3	8.2
Gas Mask (17)	clean	✗	✗	49.7	43.8	46.5	57.0
	clean	✗	✓	55.9	39.6	54.2	56.8
	backdoored	✗	✗	49.7	43.6	42.3	561.2
	defended	✗	✓	55.2	37.2	53.2	92.6
	defended	✓	✗	50.0	47.4	46.6	62.4
	defended	✓	✓	55.6	48.6	53.8	61.8
Vacuum Cleaner (18)	clean	✗	✗	49.7	51.4	46.5	63.2
	clean	✗	✓	55.9	53.6	54.3	40.0
	backdoored	✗	✗	49.8	48.8	41.7	424.4
	defended	✗	✓	55.4	54.2	53.7	55.6
	defended	✓	✗	49.9	58.0	46.4	64.6
	defended	✓	✓	54.8	51.2	53.4	37.0
American Lobseter (19)	clean	✗	✗	49.7	34.0	46.5	35.4
	clean	✗	✓	55.9	23.4	54.1	30.4
	backdoored	✗	✗	49.8	34.0	36.1	1599.2
	defended	✗	✓	55.2	23.2	49.7	482.2
	defended	✓	✗	50.0	47.2	46.7	61.2
	defended	✓	✓	55.6	32.2	54.3	39.2
Mean and STD	clean	✗	✗	49.7 ± 0.0	29.1 ± 15.3	46.4 ± 0.2	30.7 ± 19.2
	clean	✗	✓	55.9 ± 0.0	26.3 ± 15.6	54.3 ± 0.1	27.1 ± 15.6
	backdoored	✗	✗	50.0 ± 0.3	27.2 ± 16.0	40.9 ± 4.9	703.3 ± 626.1
	defended	✗	✓	55.4 ± 0.2	24.4 ± 16.1	53.0 ± 1.3	124.0 ± 132.9
	defended	✓	✗	49.7 ± 0.4	33.1 ± 17.1	46.3 ± 0.5	37.2 ± 21.3
	defended	✓	✓	55.3 ± 0.4	30.6 ± 17.2	53.8 ± 0.4	29.2 ± 16.6

Table A10. **Per category results for MoCo-v2, ResNet-18, and poison rate 0.5%**.Detailed results for Table 2.

Target				Clean Data		Patched Data	
Category	Model	PatchSearch	i-CutMix	Acc	FP	Acc	FP
Rottweiler (10)	clean	✗	✗	70.5	25.4	65.1	16.6
	clean	✗	✓	75.6	31.0	74.5	26.8
	backdoored	✗	✗	70.5	24.8	42.9	1412.4
	defended	✗	✓	75.6	28.0	70.5	268.4
	defended	✓	✗	70.4	28.2	64.8	23.8
	defended	✓	✓	75.7	32.6	74.9	28.6
Tabby Cat (11)	clean	✗	✗	70.5	3.2	64.4	4.0
	clean	✗	✓	75.6	5.0	74.5	3.4
	backdoored	✗	✗	70.7	4.6	42.0	2354.6
	defended	✗	✓	75.4	5.2	73.3	133.4
	defended	✓	✗	70.6	6.0	65.2	7.6
	defended	✓	✓	74.7	4.8	73.8	4.4
Ambulance (12)	clean	✗	✗	70.5	9.8	64.4	13.4
	clean	✗	✓	75.6	8.2	74.4	8.0
	backdoored	✗	✗	70.5	10.2	56.9	748.8
	defended	✗	✓	75.3	8.6	74.4	49.6
	defended	✓	✗	70.1	9.6	63.8	16.6
	defended	✓	✓	75.2	6.8	74.1	8.6
Pickup Truck (13)	clean	✗	✗	70.5	13.8	65.3	10.4
	clean	✗	✓	75.6	11.0	74.3	13.0
	backdoored	✗	✗	70.7	11.8	57.9	805.0
	defended	✗	✓	75.7	12.4	74.3	54.2
	defended	✓	✗	69.9	14.0	65.9	12.8
	defended	✓	✓	75.1	13.0	74.3	13.6
Laptop (14)	clean	✗	✗	70.5	29.6	64.9	39.4
	clean	✗	✓	75.6	34.4	74.5	26.6
	backdoored	✗	✗	70.5	34.6	42.3	2172.4
	defended	✗	✓	75.8	31.4	71.7	365.8
	defended	✓	✗	69.8	43.0	62.9	93.2
	defended	✓	✓	75.3	40.8	74.3	31.0
Goose (15)	clean	✗	✗	70.5	6.4	64.8	9.6
	clean	✗	✓	75.6	6.6	74.4	6.8
	backdoored	✗	✗	70.7	6.8	48.6	1693.2
	defended	✗	✓	76.1	5.2	74.0	66.6
	defended	✓	✗	69.6	10.0	64.8	17.4
	defended	✓	✓	75.1	7.2	74.0	5.0
Pirate Ship (16)	clean	✗	✗	70.5	2.0	64.2	1.4
	clean	✗	✓	75.6	2.6	74.1	1.6
	backdoored	✗	✗	70.7	2.2	56.2	915.6
	defended	✗	✓	75.8	2.0	73.5	93.4
	defended	✓	✗	70.4	2.0	62.1	1.6
	defended	✓	✓	75.7	1.6	74.6	1.2
Gas Mask (17)	clean	✗	✗	70.5	29.0	64.5	55.6
	clean	✗	✓	75.6	12.4	74.3	23.0
	backdoored	✗	✗	70.3	20.0	39.2	2558.0
	defended	✗	✓	75.4	14.2	71.0	381.6
	defended	✓	✗	70.4	42.6	64.6	71.2
	defended	✓	✓	74.9	28.0	73.8	44.2
Vacuum Cleaner (18)	clean	✗	✗	70.5	52.0	64.4	113.8
	clean	✗	✓	75.6	36.2	74.5	25.4
	backdoored	✗	✗	70.6	49.8	43.8	1847.4
	defended	✗	✓	75.2	35.4	66.3	723.4
	defended	✓	✗	70.4	60.8	65.2	128.0
	defended	✓	✓	75.0	49.4	73.8	36.2
American Lobseter (19)	clean	✗	✗	70.5	13.6	64.5	8.0
	clean	✗	✓	75.6	8.2	74.3	11.4
	backdoored	✗	✗	70.7	9.4	39.0	2581.8
	defended	✗	✓	75.9	6.4	73.3	185.4
	defended	✓	✗	70.8	14.4	65.7	25.8
	defended	✓	✓	75.1	12.4	74.2	16.8
Mean and STD	clean	✗	✗	70.5 ± 0.0	18.5 ± 15.6	64.6 ± 0.4	27.2 ± 34.8
	clean	✗	✓	75.6 ± 0.0	15.6 ± 13.0	74.4 ± 0.1	14.6 ± 10.0
	backdoored	✗	✗	70.6 ± 0.1	17.4 ± 15.1	46.9 ± 7.5	1708.9 ± 714.2
	defended	✗	✓	75.6 ± 0.3	14.9 ± 12.2	72.2 ± 2.5	232.2 ± 212.5
	defended	✓	✗	70.2 ± 0.4	23.1 ± 19.6	64.5 ± 1.2	39.8 ± 42.6
	defended	✓	✓	75.2 ± 0.3	19.7 ± 16.8	74.2 ± 0.4	19.0 ± 15.0

Table A11. **Per category results for MoCo-v3, ViT-B, and poison rate 0.5%**. Detailed results for Table 2.

Target				Clean Data		Patched Data	
Category	Model	PatchSearch	i-CutMix	Acc	FP	Acc	FP
Rottweiler (10)	clean	✗	✗	70.5	25.4	65.1	16.6
	clean	✗	✓	75.6	31.0	74.5	26.8
	backdoored	✗	✗	70.8	21.8	30.6	3127.4
	defended	✗	✓	75.6	21.4	69.0	510.8
	defended	✓	✗	70.0	32.6	64.8	26.8
	defended	✓	✓	75.3	37.4	74.3	32.0
Tabby Cat (11)	clean	✗	✗	70.5	3.2	64.4	4.0
	clean	✗	✓	75.6	5.0	74.5	3.4
	backdoored	✗	✗	70.9	3.4	23.8	3669.8
	defended	✗	✓	75.7	5.0	71.0	429.6
	defended	✓	✗	70.4	24.4	62.9	86.8
	defended	✓	✓	74.8	29.6	74.0	18.0
Ambulance (12)	clean	✗	✗	70.5	9.8	64.4	13.4
	clean	✗	✓	75.6	8.2	74.4	8.0
	backdoored	✗	✗	70.7	7.4	49.0	1511.2
	defended	✗	✓	75.4	7.4	73.0	180.8
	defended	✓	✗	69.8	25.8	64.2	36.0
	defended	✓	✓	75.0	21.0	73.9	21.0
Pickup Truck (13)	clean	✗	✗	70.5	13.8	65.3	10.4
	clean	✗	✓	75.6	11.0	74.3	13.0
	backdoored	✗	✗	70.7	12.4	54.6	1007.0
	defended	✗	✓	75.2	9.6	73.0	117.8
	defended	✓	✗	69.7	18.4	64.9	19.2
	defended	✓	✓	75.2	18.6	73.9	21.4
Laptop (14)	clean	✗	✗	70.5	29.6	64.9	39.4
	clean	✗	✓	75.6	34.4	74.5	26.6
	backdoored	✗	✗	70.8	27.0	39.4	2566.4
	defended	✗	✓	75.4	25.8	66.9	724.8
	defended	✓	✗	69.9	61.2	60.3	112.8
	defended	✓	✓	75.6	54.2	74.7	55.0
Goose (15)	clean	✗	✗	70.5	6.4	64.8	9.6
	clean	✗	✓	75.6	6.6	74.4	6.8
	backdoored	✗	✗	70.5	5.8	38.6	2437.8
	defended	✗	✓	75.7	5.2	71.2	314.2
	defended	✓	✗	70.1	32.8	63.4	59.0
	defended	✓	✓	74.7	33.4	73.7	25.0
Pirate Ship (16)	clean	✗	✗	70.5	2.0	64.2	1.4
	clean	✗	✓	75.6	2.6	74.1	1.6
	backdoored	✗	✗	70.3	2.6	44.9	2093.8
	defended	✗	✓	75.8	2.4	73.5	90.8
	defended	✓	✗	70.3	20.2	62.4	36.0
	defended	✓	✓	75.0	18.2	74.0	24.2
Gas Mask (17)	clean	✗	✗	70.5	29.0	64.5	55.6
	clean	✗	✓	75.6	12.4	74.3	23.0
	backdoored	✗	✗	70.4	21.2	31.6	3112.0
	defended	✗	✓	76.1	8.8	67.7	728.6
	defended	✓	✗	70.2	79.8	64.7	138.8
	defended	✓	✓	74.6	66.2	73.5	102.8
Vacuum Cleaner (18)	clean	✗	✗	70.5	52.0	64.4	113.8
	clean	✗	✓	75.6	36.2	74.5	25.4
	backdoored	✗	✗	71.2	38.4	40.6	2203.4
	defended	✗	✓	75.3	30.6	64.9	903.0
	defended	✓	✗	70.4	85.8	64.8	182.8
	defended	✓	✓	74.6	69.2	73.8	63.2
American Lobseter (19)	clean	✗	✗	70.5	13.6	64.5	8.0
	clean	✗	✓	75.6	8.2	74.3	11.4
	backdoored	✗	✗	70.3	8.0	24.1	3630.2
	defended	✗	✓	75.1	9.0	71.6	341.6
	defended	✓	✗	70.0	53.4	64.3	61.6
	defended	✓	✓	75.1	42.2	74.3	51.0
Mean and STD	clean	✗	✗	70.5 ± 0.0	18.5 ± 15.6	64.6 ± 0.4	27.2 ± 34.8
	clean	✗	✓	75.6 ± 0.0	15.6 ± 13.0	74.4 ± 0.1	14.6 ± 10.0
	backdoored	✗	✗	70.7 ± 0.3	14.8 ± 11.8	37.7 ± 10.2	2535.9 ± 873.6
	defended	✗	✓	75.5 ± 0.3	12.5 ± 9.8	70.2 ± 2.9	434.2 ± 279.3
	defended	✓	✗	70.1 ± 0.2	43.4 ± 25.0	63.7 ± 1.5	76.0 ± 53.9
	defended	✓	✓	75.0 ± 0.3	39.0 ± 18.8	74.0 ± 0.3	41.4 ± 27.0

Table A12. **Per category results for MoCo-v3, ViT-B, and poison rate 1.0%**. Detailed results for Table 2.