A Proofs

A.1 Proof of Theorem 1

Lemma 1. *Given a random variable, X follows a uniform distribution with support $U(0, p_u)$. The p.d.f. of X^2 is $\frac{1}{2p_u\sqrt{x}}$ and $E(X^2) = \frac{p_u^2}{3}$*

Proof. Considering

$$\Pr(X^2 < x) = \Pr(X < \sqrt{x}) = \int_0^{\sqrt{x}} \frac{1}{p_u}\, dx = \frac{1}{p_u}\sqrt{x}, \tag{7}$$

which is the c.d.f. of X^2. To obtain the p.d.f. of X^2, we compute $\frac{d}{dx}\left(\frac{1}{p_u}\sqrt{x}\right) = \frac{1}{2p_u\sqrt{x}}$. Note that the support of X^2 is $[0, p_u^2]$ not $[0, p_u]$. To obtain

$$E(X^2) = \int_0^{p_u^2} \frac{1}{2p_u\sqrt{x}}\, x\, dx = \frac{1}{2p_u}\frac{2}{3}x^{3/2}\Big|_0^{p_u^2} = \frac{1}{3p_u}p_u^3 = \frac{p_u^2}{3} \tag{8}$$

q.e.d.

Theorem 1. *Given two different filters, R_i and R_c, generated by CUDA, if the condition $1 > (\kappa^2 - 2)p_u^2 + 2p_u$ is satisfied, $R_{c\pi} \star R_{i\pi}$ and $R_{c\pi} \star R_{c\pi}$ have the following properties.*

1. *The peak of $R_{c\pi} \star R_{c\pi}$ is located at the center.*
2. *The peak of $R_{c\pi} \star R_{i\pi}$ occurs at the position where two ones in $R_{c\pi}$ and $R_{i\pi}$ appear at the same position in $R_{c\pi} \star R_{i\pi}$.*
3. *The probability of peak of $R_{c\pi} \star R_{i\pi}$ at the center is $\frac{1}{\kappa^2}$.*
4. *The expected peak of $R_{c\pi} \star R_{c\pi}$ is higher than the expected peak of $R_{c\pi} \star R_{i\pi}$.*

Proof. Property 1: The center of $R_{c\pi} \star R_{c\pi}$ is at (κ, κ), where two filters completely overlap, and its minimum value is larger than 1. The maximum value at other positions is less than $(\kappa^2 - \kappa - 2)p_u^2 + 2p_u$. Using the assumption, we can deduce that $R_{c\pi} \star R_{c\pi}(\kappa, \kappa) > R_{c\pi} \star R_{c\pi}(i, j)$, for all $(i, j) \neq (\kappa, \kappa)$.

Property 2: When the one in $R_{c\pi}$ and the one in $R_{j\pi}$ appear at the same position in $R_{c\pi} \star R_{j\pi}$, the minimum value of $R_{c\pi} \star R_{j\pi}$ at the location is larger than one. The maximum value of other positions is less than $(\kappa^2 - 2)p_u^2 + 2p_u$. Based on the assumption, we achieve property 2.

Property 3: Based on property 2, the peak of $R_{c\pi} \star R_{j\pi}$ is at the position where the one in $R_{c\pi}$ and the one in $R_{j\pi}$ appear at the same position in $R_{c\pi} \star R_{j\pi}$. If the peak is at the center of $R_{c\pi} \star R_{j\pi}$, it means that the one in $R_{c\pi}$ and the one in $R_{j\pi}$ must be at the position. The probability that they are in the position is $\frac{1}{\kappa^2}$.

Property 4: Utilizing Lemma 1 and the preceding properties, we can derive the expected peak value of $R_{c\pi} \star R_{c\pi}$:

$$E(R_{c\pi} \star R_{c\pi}(\kappa, \kappa)) = 1 + (\kappa^2 - 1)\left(\frac{p_u^2}{3}\right). \tag{9}$$

When two ones are at the same position in $R_{c\pi}$ and $R_{i\pi}$, $R_{c\pi} \star R_{i\pi}(\kappa, \kappa)$ achieves its maximum value. Thus the expected maximum peak value of $R_{c\pi} \star R_{i\pi}$ is

$$E(R_{c\pi} \star R_{i\pi}(\kappa, \kappa)) = 1 + (\kappa^2 - 1)\left(\frac{p_u^2}{4}\right). \tag{10}$$

A.2 Proof of Theorem 2

Theorem 2. *Given a CUDA filter size being larger than 1 and x_{cij} being the vector form of the patch in X_{ci} involved in producing η_{cij}, i.e., $\eta_{cij} = \wp x_{cij}$, where $\wp$ is a $\kappa^2 \times (2\kappa - 1)^2$ matrix, if the column vectors of $\wp$, excluding the one corresponding to y_{cij}, span $\mathbb{R}^{\kappa^2}$, then, $\exists x_{cij} \in \mathbb{R}^{(2\kappa-1)^2}$ such that $y_{cij} \neq a_c^T \wp x_{cij}$.*

Proof. Let the patch in X_{ci} involved in computing the patch $(X_{ci} \star R_c)_j$ be $(X_{ci})_j$. Since the patch size of $(X_{ci} \star R_c)_j$ and the size of R_c are both $\kappa \times \kappa$, $((X_{ci})_j)$ has the size of $(2\kappa - 1) \times (2\kappa - 1)$. When $\kappa > 1$, $(X_{ci})_j$ is larger than $(X_{ci} \star R_c)_j$. Reshaping $(X_{ci})_j$ and $(X_{ci} \star R_c)_j$ as vectors, x_{cij} and η_{cij}, their relationship can be represented as a linear equation, i.e., $\eta_{cij} = \wp x_{cij}$, where $\wp$ is a $\kappa^2 \times (2\kappa - 1)^2$ matrix. The matrix $\wp$ is constructed by R_c. To achieve perfect reconstruction $y_{cij} = (\alpha_c^T \wp) x_{cij}$ for all $x_{cij} \in \mathbb{R}^{(2\kappa-1)^2}$, $\alpha_c^T \wp$ must be $[0 \cdots 0\, 1\, 0 \cdots 0]$, where the location of the one corresponds to y_{cij} in $(X_{ci})_j$. Since the column vectors of $\wp$, excluding the one corresponding to y_{cij}, span $\mathbb{R}^{\kappa^2}$, α_c cannot be orthogonal to all these column vectors. Thus, $\alpha_c^T \wp = [0 \cdots 0\, 1\, 0 \cdots 0]$ cannot be obtained.

A.3 Extending CUDA to CIFAR-10

Since CUDA uses $p_u = 0.3$ and $\kappa = 3$ for CIFAR-10, it fulfils the condition $1 > \frac{1}{4}(\kappa^2 - 2)p_u^2 + p_u$. According to Theorem 3, CUDA filters exhibit similar properties shown in Theorem 1 on CIFAR-10.

Theorem 3. *Given two different filters, R_i and R_c generated by CUDA, if the condition $1 > \frac{1}{4}(\kappa^2 - 2)p_u^2 + p_u$ is satisfied, $R_{c\pi} \star R_{i\pi}$ and $R_{c\pi} \star R_{c\pi}$ have the following properties.*

1. *The expected peak of $R_{c\pi} \star R_{c\pi}$ is located at the center.*
2. *The expected peak of $R_{c\pi} \star R_{i\pi}$ occurs at the position where two ones in $R_{c\pi}$ and $R_{i\pi}$ appear at the same position in $R_{c\pi} \star R_{i\pi}$.*
3. *The probability of the expected peak of $R_{c\pi} \star R_{i\pi}$ at the centre is $\frac{1}{k^2}$.*
4. *The expected peak of $R_{c\pi} \star R_{c\pi}$ is higher than the expected peak of $R_{c\pi} \star R_{i\pi}$.*

Proof. Property 1: The center of $R_{c\pi} \star R_{c\pi}$ is at (κ, κ), where two filters completely overlap, and its expected value is

$$E(R_{c\pi} \star R_{c\pi}(\kappa, \kappa)) = 1 + \frac{1}{3}(\kappa^2 - 1)p_u^2. \tag{11}$$

Let ω be the number of pixels overlapped when computing $E(R_{c\pi} \star R_{c\pi}(i, j))$, where $(i, j) \neq (\kappa, \kappa)$. We have the following cases:

Case 1: If both ones are not inside the overlapping region,

$$E(R_{c\pi} \star R_{c\pi}(i,j)) = \frac{1}{4}\omega p_u^2$$

Case 2: If both ones are inside the overlapping region but at different locations,

$$E(R_{c\pi} \star R_{c\pi}(i,j)) = \rho_u + \frac{1}{4}(\omega - 2)p_u^2$$

Case 3: If both ones are inside the overlapping region and at the same location,

$$E(R_{c\pi} \star R_{c\pi}(i,j)) = 1 + \frac{1}{4}(\omega - 1)p_u^2$$

Case 4: If only one is inside the overlapping region,

$$E(R_{c\pi} \star R_{c\pi}(i,j)) = \frac{1}{2}\rho_u + \frac{1}{4}(\omega - 1)p_u^2$$

Among the four cases, the third case is the largest one, since $\rho_u < 1$. Clearly $E(R_{c\pi} \star R_{c\pi}(\kappa,\kappa))$ is larger than $E(R_{c\pi} \star R_{c\pi}(i,j))$ in the third case. Thus, the expected peak of $E(R_{c\pi} \star R_{c\pi}(\kappa,\kappa))$ is at the center.

Property 2: When two ones overlap, the minimum value of $E(R_{c\pi} \star R_{i\pi}(i,j))$ is larger than 1. The expected maximum value of $E(R_{c\pi} \star R_{i\pi}(i,j))$ when two ones do not overlap is

$$E(R_{c\pi} \star R_{i\pi}(i,j)) = \rho_u + \frac{1}{4}(\kappa^2 - 2)p_u^2.$$

Using the assumption, the expected peak appears at the position when two ones overlap.

Property 3: Based on property 2, the expected peak appears at the position when two ones are at the same position. Thus, the probability of the expected peak appearing at the center is $\frac{1}{\kappa^2}$

Property 4: The proof is the same as the proof of property 4 in Theorem 1.

B Experiment Setup

B.1 Training details

The training details are provided in Table 7. However, there are two exceptions: for VGG-19 and ViT models, the learning rate in SL is set to 0.01, and the training epoch for ViT is adjusted to 200. For the SL training, we utilize the standard linear probing augmentation of SSL as it has demonstrated an enhancement in clean test accuracy. This implies that the linear probing augmentation acts as an effective defense against certain DAAs. Specifically, in ImageNet training, the data augmentation comprises RandomResizedCrop with an expected output size of 224 along with RandomHorizontalFlip. For CIFAR-10 training, RandomCrop with an output size of 32 and padding set to 4 is employed, alongside RandomHorizontalFlip. For SSL pretraining and linear probing, we follow the standard augmentations used in [4–6, 16].

Table 7: Training settings for SL and SSL. For cells with two values, the left value is for CIFAR-10 and the right value is for ImageNet-100. 'Pretrain' is the standard augmentation utilized in SSL pretraining, while 'LinProbe' is the standard augmentation employed in SSL linear probing.

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

C Additional Results

C.1 Images with Double Blur and Random Sharpness

The examples presented in Figure 7 depict outputs generated through Double Blur, Random Sharpness, and IRP. The images produced by Double Blur exhibit the highest level of blurriness. The degree of blurriness in the images generated by Random Sharpness varies, as the sharpness filters are randomly generated class-wise. Conversely, the images produced by IRP are the clearest, owing to the optimization process customized for the blurriness filter in IRP.

Fig. 7: Example images generated by Double Blur (the first row), Random Sharpness (the second row), and IRP (the third row) on ImageNet-100.

C.2 Additional Defenses

We further evaluate IRP on three recently invented poison defenses. AVATAR [10] adds Gaussian noise to poisoned images and 'purifies' them with a pretrained diffusion model. UEraser [26] applies multiple augmentations to poisoned images and selects the one that maximizes loss. COIN [36] applies random pixel interpolations to reduce the impact of filter-based poisons.

Figure 8 shows that AVATAR and COIN are useful defenses against both CUDA and IRP poisons. Under COIN purification, CUDA-poisoned CIFAR-10 permits a clean test accuracy of 71.90%. AVATAR raises the max accuracy of IRP-poisoned CIFAR-10 to 54.78%, though this value is still severely low compared to training on clean data. UEraser is less effective as a defense than standard methods like SSL and AT. Across all defenses, IRP is a more potent DAA than CUDA.

Table 8: Top-1 clean test accuracies (%) of ResNet-18 trained with data poisoned by various DAAs on CIFAR-10. The "Max" column represents the highest clean test accuracy, which corresponds to the worst performance of each DAA under different training scenarios. Bold indicates the best performance.

DAA	SL	SSL	AT	AVATAR	UEraser	COIN	Max
Clean	93.86	90.55	89.57	-	-	-	93.86
CUDA	21.89	66.58	48.58	55.93	40.56	71.90	71.90
IRP	**10.39**	**43.24**	**32.21**	**54.78**	**26.55**	**53.32**	**54.78**