A. Future Directions:

Given that $\hat{\nabla}_{\text{FREAK}}(x)$ allows us to locate the indices of top-k most sensitive frequencies, a question that arises is can we reconstruct those magnitude values similar to what was done in Februus [2]? Our results show that indeed this might be a feasible approach. This approach is summarized in Figure 1.

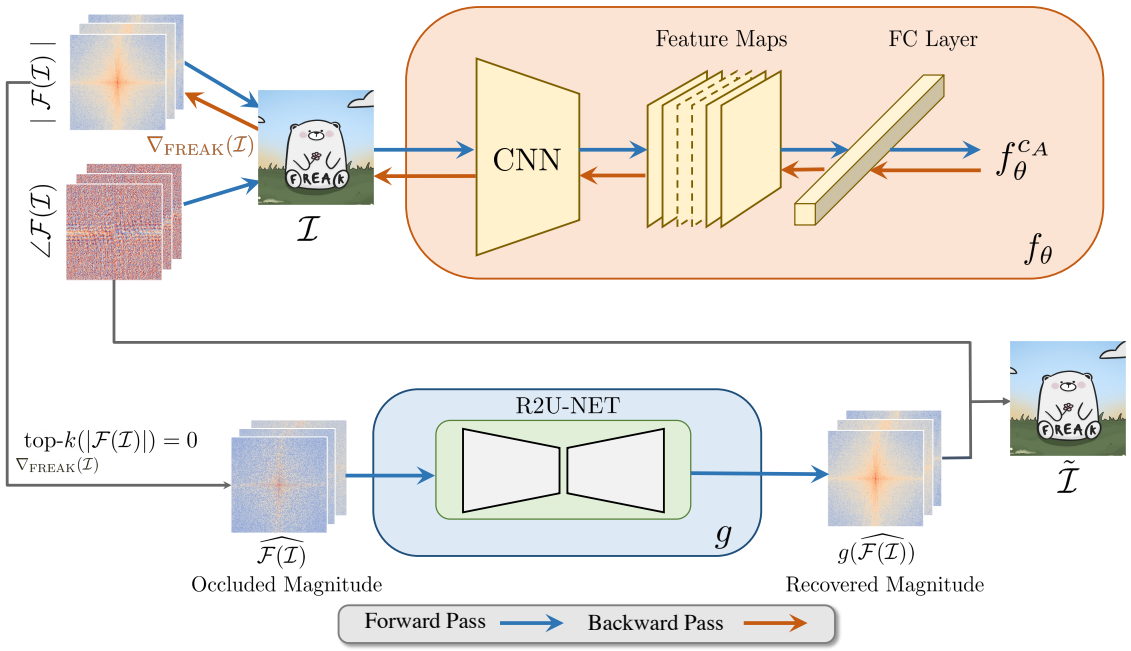

Figure 1: **FREAK for Image Purification.** We attempt to purify the images by first detecting the k most sensitive frequency components, masking them by zero (occluding them), and then reconstructing the magnitude components using an R2U-Net.

In simple terms, the idea is to locate the k most sensitive frequency components, set them to zero and attempt to reconstruct them using an R2U-Net [1]. The loss used for training this network is a simple MSE on the recovered images and an MSE on the log Fourier recovered magnitude. Mathematically, the loss can be written as

$$\mathcal{L} = \mathcal{L}_{\text{MSE}}(\tilde{\mathcal{I}}, \mathcal{I}) + \mathcal{L}_{\text{MSE}}(\log(g(|\widehat{\mathcal{F}(\mathcal{I})}|)), \log(|\mathcal{F}(\mathcal{I})|))$$

Method	**CDA**(%)	**ASR**(%)
No Defense	92.51	96.54
Gaussian (3x3)	65.12	92.85
Gaussian (5x5)	36.07	92.85
Weiner (3x3)	65.86	92.85
Weiner (5x5)	42.58	92.85
Highpass	31.34	14.28
Lowpass	33.16	71.42
Bandpass	22.98	0.00
JPEG Compression	83.80	92.85
Autoencoder	82.33	71.43
FREAK (ours)		
top-k = 0%	91.79	85.71
top-k = 25%	90.02	14.55
top-k = 50%	87.10	5.59

(a) FIBA [3]

Method	**CDA**(%)	**ASR**(%)
No Defense	92.84	100.00
Gaussian (3x3)	64.00	0.00
Gaussian (5x5)	34.80	7.14
Weiner (3x3)	67.14	7.14
Weiner (5x5)	46.93	0.00
Highpass	33.81	85.71
Lowpass	32.74	7.14
Bandpass	26.72	0.00
JPEG Compression	85.05	0.00
Autoencoder	82.05	0.00
FREAK (ours)		
top-k = 0	92.68	100.00
top-k = 25%	91.06	0.85
top-k = 50%	89.22	0.65

(b) FTrojan [5]

Method	**CDA**(%)	**ASR**(%)
No Defense	94.43	100.00
Gaussian (3x3)	63.96	0.00
Gaussian (5x5)	29.38	14.28
Weiner (3x3)	68.12	7.14
Weiner (5x5)	47.84	7.14
Highpass	36.96	64.28
Lowpass	26.41	7.14
Bandpass	27.14	0.00
JPEG Compression	86.91	0.00
Autoencoder	83.15	0.00
FREAK (ours)		
top-k = 0	94.38	92.89
top-k = 25%	93.54	7.03
top-k = 50%	92.29	5.59

(c) CYO [4]

Table 1: **Defending Against Frequency Backdoor Attacks (CIFAR10).** The results of applying various defenses against existing frequency backdoor attacks show that using FREAK approach allows for the best balance between CDA and ASR.

We test that approach against a various set of defenses such as filtering approaches, some of which were proposed in [4, 5], namely, Gaussian, Weiner, Highpass, Lowpass, and Bandpass filtering and compression approaches such as: JPEG and Autoencoder compression. As shown in Table 1, this approach proves to be a solid approach to defend against backdoor attacks. More precisely, using this frequency reconstruction approach during test time allows us to maintain a high clean data accuracy while dropping the attack success rate to a very low level. This is observed for all three studied frequency-backdoor attacks.

However, our experiments on ImageNet show that this approach might not be scalable on large scale images where we observed a large drop in performance in terms of CDA and ASR trade-off. This calls for further research to develop a different loss function and architecture for applying frequency reconstruction.

B. Additional Results:

B.1. Results on ResNet34

In this subsection we present evaluations of FREAK against the CYO, FTrojan and FIBA using ResNet34 model instead of ResNet18. FREAK still proves to be a useful defense for detecting poisoned samples.

	TPR (%)	**FPR (%)**
CYO	99.61	1.95
FTrojan	99.80	4.29
FIBA	91.20	7.31

Table 2: **FREAK Against Frequency Backdoor Attacks.** FREAK proves to be capable of achieving a high TPR while maintaining a low FPR against frequency backdoor-attacks.

B.2. Hyperparameter Sensitivity of FREAK

Now we study the sensitivity of FREAK to the different hyperparameters. Unless the hyperparameter is being ablated, the value is fixed to that presented in the manuscript *i.e.* $\alpha = 1$, $|\mathcal{D}_h| = 32$, $|\mathcal{D}_c| = 128$, $\beta = 12$, and $k = 5000$.

Top-k Value. As shown in table 3, increasing the value of k allows for a lower FPR at the cost of a drop in TPR.

	k	**TPR (%)**	**FPR (%)**
CYO	2500	99.68	2.97
	7500	99.02	1.95
FTrojan	2500	100.00	3.90
	7500	100.00	2.73
FIBA	2500	87.89	5.85
	7500	85.46	5.63

Table 3: **Effect of k in top-k Operation of FREAK**

Size of Pooling Filter (β) Table 4, shows the effect of changing the filter size β.

	β	TPR (%)	FPR (%)
CYO	9	99.22	1.57
	16	99.22	1.95
FTrojan	9	100.00	2.34
	16	100.00	4.29
FIBA	9	91.79	8.98
	16	88.08	4.98

Table 4: **Effect of Pooling Size β on FREAK**

Size of Held Out Set Table 5, shows the effect of increasing the size of the held-out set. Our results show little to no change in the performance of FREAK with increased size of held-out set.

	$\mid \mathcal{D}_h \mid$	TPR (%)	FPR (%)
CYO	64	99.22	1.95
	256	99.22	1.95
FTrojan	64	100.00	2.73
	256	100.00	2.73
FIBA	64	89.68	5.15
	256	90.47	5.31

Table 5: **Effect of Size of Held-Out Set on FREAK**

Size of Clean Set Table 6, shows the effect of changing the size of the clean set. Our results show little to no change in the performance of FREAK with increased size of held-out set.

	$\mid \mathcal{D}_c \mid$	TPR (%)	FPR (%)
CYO	64	99.22	2.23
	256	99.22	1.56
FTrojan	64	100.00	2.45
	256	100.00	2.73
FIBA	64	86.48	4.02
	256	90.53	5.47

Table 6: **Effect of Size of Clean Set on FREAK**

Trade-Off Parameter α Table 7, shows the effect of changing the confidence parameter α. As expected, as α increases we are less

	α	**TPR (%)**	**FPR (%)**
CYO	2	99.02	0.78
	4	96.09	0.39
FTrojan	2	100.00	1.17
	4	100.00	1.17
FIBA	2	81.49	2.96
	4	65.00	1.25

Table 7: **Effect of Changing α on FREAK**

References

[1] Md. Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *ArXiv*, abs/1802.06955, 2018.

[2] Bao Gia Doan, Ehsan Abbasnejad, and Damith Chinthana Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. *Annual Computer Security Applications Conference*, 2020.

[3] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. *ArXiv*, abs/2112.01148, 2021.

[4] Hasan Abed Al Kader Hammoud and Bernard Ghanem. Check your other door! creating backdoor attacks in the frequency domain. 2021.

[5] Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. Backdoor attack through frequency domain. *ArXiv*, abs/2111.10991, 2021.