# *Supplementary for:*
# "Data-free Defense of Black Box Models Against Adversarial Attacks"

Gaurav Kumar Nayak
University of Central Florida
Orlando, Florida, USA
gauravkumar.nayak@ucf.edu

Inder Khatri
New York University
New York, USA
ik2535@nyu.edu

Ruchit Rawal, Anirban Chakraborty
Indian Institute of Science
Bangalore, India
rawalruchit22@gmail.com,
anirban@iisc.ac.in

## 1. Ablation on coefficient selection strategy

In our method DBMA, we select only a limited number of coefficients(i.e., $k\%$ coefficients) to obtain a decent tradeoff between the adversarial and clean performance. As shown in Fig 1 (A) in the main paper, the least affected approximate coefficients have a high magnitude. Thus, we select the essential detail coefficients as the top-$k$ high magnitude. As they are likely to be less affected by the adversarial attack (Fig 1 (B)) (main paper), they can yield better performance. In this section, we perform an ablation to assess the effectiveness of selecting top-$k$ coefficients over the other possible choices. For comparative analysis, we perform the experiments using the bottom-$k$ and random-$k$ coefficients. Table 1 shows the adversarial and clean performance for the different coefficient selection strategies. The top-$k$ coefficient selection strategy, selecting the most important coefficient in terms of magnitude, improves both the clean and adversarial performance. On the other hand bottom-$k$ coefficient selection strategy, selecting the least significant coefficients shows the least performance as they primarily consist of contaminated high-frequency content. Although, randomly selecting $k\%$ coefficients showed improved performance than the bottom-$k$, it still performs poorly compared to top-$k$ coefficient selection strategy.

Table 1. Performance comparison when $k\%$ detail coefficients are selected in wavelet coefficient selection module (WCSM) using different methods. Selection of top-k coefficients yields better clean and adversarial accuracy than other strategies.

| Surrogate model (attacker) | coefficient selection strategy | Black Box Model : Alexnet Surrogate Model (defense): Resnet-18 | | | |
|---|---|---|---|---|---|
| | | clean | BIM | PGD | Auto Attack |
| Alexnet-half | bottom-$k$ | 31.25 | 8.59 | 7.32 | 12.57 |
| | random-$k$ | 42.58 | 13.13 | 11.74 | 19.77 |
| | top-$k$ (**ours**) | **77.92** | **26.66** | **24.55** | **34.02** |
| Alexnet | bottom-$k$ | 31.25 | 6.14 | 4.84 | 10.92 |
| | random-$k$ | 42.92 | 8.61 | 7.52 | 14.29 |
| | top-$k$ (**ours**) | **77.92** | **15.98** | **14.04** | **21.34** |

## 2. Performance of our method (DBMA) using different wavelets

The experiments in the main draft, have used Daubechies wavelets [4] in wavelet noise remover (WNR). Along with Daubechies, several other wavelets are available in the literature that varies in time, frequency, and rate of decay. This section analyzes the effect of different wavelets on the performance of proposed data-free black box defense (DBMA). We perform experiments with Coiflets [2] and Biorthogonal wavelets [3]. Table 2 summarizes the results obtained with Resnet18 as the defender's surrogate model and Alexnet-half as the attacker's surrogate model. We observe a similar performance trend over the adversarial and clean samples on using the different wavelet functions when compared to the Daubechies. This confirms DBMA yields consistent performance irrespective of the choice of the wavelet function used.

Table 2. Ablation over different choices of wavelets that are used in wavelet noise remover (WNR). The performance of DBMA remains consistent across different choices of wavelet functions.

| Surrogate model (Attacker) | Surrogate model (Defender) | Black Box Model : Alexnet Surrogate Model (defender): Resnet-18 | | | |
|---|---|---|---|---|---|
| | | clean | BIM | PGD | Auto Attack |
| Alexnet-half | Biorthogonal | 73.27 | 42.51 | 41.72 | 51.11 |
| | Daubechies | 73.77 | 42.71 | 42.71 | 50.63 |
| | Coiflets | 73.14 | 42.51 | 44.22 | 51.94 |

## 3. Defense against different data-free black box attacks

In all the previous experiments we assume that both defender and attacker use the same model stealing technique [1] for creating a surrogate model. To prove our data-free black box defense (DBMA) is robust to different model stealing strategies, we evaluate our method against two different approaches: a) Data-free model extraction (DFME) [7] and b) Data-free model stealing in hard label setting (DFMS-HL) [6] that are used by attacker to obtain

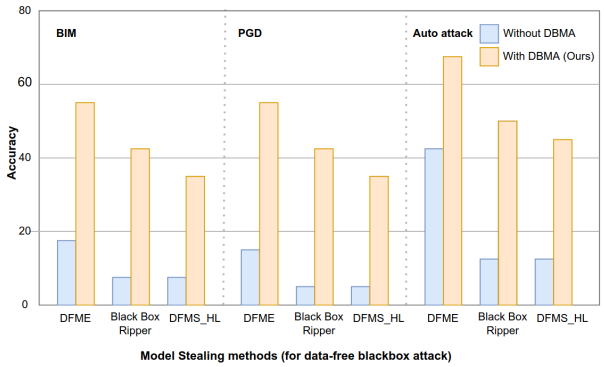


Figure 1. Performance of our approach DBMA for different model stealing methods used to get the attacker's surrogate model for data-free black box attacks. DBMA consistently improves performance against different attacks across all model stealing methods.

a surrogate model (Alexnet-Half) for crafting adversarial samples. As shown in Fig. 1, our method yields a consistent boost in adversarial accuracy on different data-free model stealing methods. In the case of DFME, we observe a massive improvement of $\approx$36%, $\approx$40% and $\approx$21% on BIM , PGD and Auto Attack respectively. On the other hand, in case of DFMS-HL the performance against the three attacks improves by $\approx$28%, $\approx$29%, and $\approx$31%. Overall across different model stealing methods, our method (DBMA) yields significant improvement in adversarial accuracy (i.e. $\approx 29-42\%$ in PGD, $\approx 28-38\%$ in BIM , and $\approx 25-31\%$ in state-of-the-art Auto Attack).

Further, we check our defense in a more tougher scenario, where attacker is aware of the black box model $B_m$'s architecture (i.e., Alexnet), and uses the same for attacker's surrogate model ($S_m^a$). The results for this setup are reported in Table 3. Compared to baseline, we observe an improvement of $\approx 29-32\%$, $\approx 38-40\%$ and $\approx 22-25\%$ in adversarial accuracy across attacks using the Black box ripper, DFME and DFMS-HL methods, respectively. This indicates that even if the attacker is aware of the black box model's architecture, our method DBMA can provide data-free black-box adversarial defense irrespective of the different model stealing methods.

Hence, our proposed method DBMA provides strong robustness even when the attacker uses a different model stealing strategy compared to the defender. Refer to next section for adversarial robustness results when the defender uses different model stealing methods to obtain the surrogate model.

## 4. DBMA using different model stealing methods

In all the experiments in the main draft, we used the Black Box Ripper (BBR) model stealing method to obtain surrogate model for defense. To demonstrate that our method DBMA can work across different model stealing techniques also, we obtain the defender's surrogate model using a different model stealing method (DFMS-HL). To get better insights, we analyze the performance of DBMA by varying the model stealing techniques for both attack and defense.

In Table 4 we observe the clean accuracy is not affected on using different model stealing techniques. DBMA obtains the best performance when the attacker uses BBR to attack while the defense is performed using model stealing DFMS-HL (third row). However, when the attacker uses the DFMS-HL method to create a surrogate model (second and fourth row), the adversarial accuracy decreases by $\approx 7-11\%$ across attacks compared to the performance with BBR method used by attacker. For attacks crafted using BBR, the defender's performance with DFMS-HL remains almost similar to BBR (first and third rows). These results suggest that the adversarial samples created using the DFMS-HL are stronger than the ones created using BBR method. This aligns with the Sec. 5.6 in the main draft, where we observed a similar trend indicating that the attacks crafted using DFMS-HL are powerful than BBR. Further, we observe that the choice of the defender's model stealing method does not significantly affect the adversarial and clean accuracy.

Overall, for different model stealing methods, DBMA ensures good clean performance with decent adversarial performance.

## 5. Architecture details of Regenerator Network

The Regenerator network consists of U-net-based [5] generator with five downsampling and upsampling layers. Each $i^{th}$ downsampling layer has a skip connection to $(n-i)^{th}$ upsampling layer that concatenates channels of $i^{th}$ layer with those at layer $n-i$, $n$ represents number of upsampling and downsampling layers (i.e. $n=5$). The number of channels in the network's input and output are the same as the image channels in the training dataset (surrogate data $S_d$ in our case). Each Downsampling layer first filters input through Leaky relu with negative $slope = 0.2$, followed by convolution operation with $kernelsize = 4$, $stride = 2$, and $padding = 1$. The number of output channels for layers 1 to 5 are 64,128,256,512,512 respectively. Upsampling layers start with a Relu layer. Followed by transposed convolution. Each transposed convolution has $kernelsize = 4$ , $padding = 1$ and $stride = 2$. The number of output channels for layers 1 to 5 are 1024, 512, 256, 128, and 3 respectively.The output of convolution and deconvolution layers is normalized using instance normalization to avoid 'instance-specific mean and covariance shifts'. The output of the last upsampling layer is normalized using Tanh normalization to ensure output values lie in the range $[-1, 1]$.

Table 3. Performance of DBMA across several data-free black box attacks that are constructed using the surrogate model having same architecture as the black box model. We observe consistent significant improvements in adversarial performance using proposed DBMA across different adversarial attacks and model stealing techniques.

| Surrogate model (attacker) | Method | Model stealing (Attacker) | Black Box Model : Alexnet Surrogate Model (defense): Resnet-18 | | | |
|---|---|---|---|---|---|---|
| | | | clean | BIM | PGD | Auto Attack |
| Alexnet | Without DBMA | Black Box Ripper | 82.58 | 4.17 | 2.19 | 8.55 |
| | With DBMA (**Ours**) | Black Box Ripper | 73.77 | **33.31** (↑ 29.14) | **31.72** (↑ 29.53) | **40.56** (↑ 32.01) |
| | Without DBMA | DFME | 82.58 | 4.21 | 1.99 | 16.93 |
| | With DBMA (**Ours**) | DFME | 73.77 | **42.84** (↑ 38.63) | **42.49** (↑ 40.5) | **55.04** (↑ 38.11) |
| | Without DBMA | DFMS-HL | 82.58 | 3.30 | 1.82 | 7.84 |
| | With DBMA (**Ours**) | DFMS-HL | 73.77 | **26.0** (↑ 22.7) | **24.94** (↑ 23.12) | **32.57** (↑ 24.73) |

Table 4. Performance of DBMA across different model stealing methods used by defender and attacker. DBMA obtains respectable performance irrespective of the model stealing technique used by either the defender or attacker.

| Surrogate model (attacker) | Model stealing (defender) | Model stealing (attacker) | Black Box Model : Alexnet Surrogate Model (defense): Resnet-18 | | | |
|---|---|---|---|---|---|---|
| | | | clean | BIM | PGD | Auto Attack |
| Alexnet-half | BBR | BBR | 73.77 | 42.71 | 42.71 | 50.63 |
| | BBR | DFMS-HL | 73.77 | 35.4 | 34.58 | 43.05 |
| | DFMS-HL | BBR | 72.45 | **42.75** | **43.08** | **51.0** |
| | DFMS-HL | DFMS-HL | 72.45 | 32.85 | 31.86 | 40.6 |

## 6. Importance of Regenerator Network

To evaluate the effectiveness of proposed Regenerator network $R_n$ in our approach DBMA, we analyze the performance of DBMA with and without $R_n$ for different values of $k$ (i.e. $k$=1, 2, 4, 8, 16 (ours), 50). In Table 5, we observe that across different $k$, appending $R_n$ to the WNR improves the adversarial accuracy against various attacks crafted using Alexnet-half and Alexnet by $\approx 13-18\%$. Further, we observe a similar trend for clean accuracy, which also improves on adding $R_n$ to WNR. However, the improvement margin for clean accuracy gradually drops on increasing the value of coefficient percent $k$. For higher $k$'s (e.g., 16 , 50), there is a small drop in clean performance using $R_n$ compared to the performance with only WNR but leads to significant increase in adversarial accuracy. This implies regenerator network enhances the output image of WNR to increase the adversarial accuracy. In this process, for smaller values of $k$, it increases clean accuracy too, but for a large value of $k$, the decrease in clean accuracy is compensated by the increase in adversarial accuracy to achieve the best trade-off. Combining WNR with the regenerator network at our $\hat{k}$ (i.e., $k = 16$) produces the best adversarial accuracy.

From Table 6, we observe that DBMA with only WNR improves adversarial accuracy with a small drop in clean accuracy compared to baseline. Similarly, with only regenerator network $R_n$, adversarial accuracy increases compared to the baseline. Also, $R_n$ performs better than WNR. However, using WNR and $R_n$ together in DBMA gives the best adversarial accuracy. Hence this demonstrates the importance of both the defense components (WNR and $R_n$) in our method DBMA.

## 7. Ablation on choice of surrogate architecture

To better analyze the performance of DBMA against different combinations of defender surrogate ($S_m^d$) and attacker surrogate models ($S_m^a$), we perform experiments with different choices of surrogate models (i.e., Alexnet-half, Alexnet, and Resnet18). For all the experiments, Alexnet is used as the black box model. For Resnet18 as $S_m^d$, the wavelet coefficient selection module (WCSM) yields optimal $k$ ($\hat{k}$) as 16. Similarly, for other choices of $S_m^d$ (i.e., Alexnet and Alexnet-half), we obtain $\hat{k}$ as 15. This shows that the value of $\hat{k}$ is not much sensitive to the choice of architecture of the defender's surrogate model $S_m^d$. The results are reported in Table 7.

We obtain the best performance for Resnet-18 as $S_m^d$ and Alexnet-half as $S_m^a$ ($1^{st}$ row), whereas the lowest for Alexnet-half as $S_m^d$ and Resnet18 as $S_m^a$ ($6^{th}$ row). Further, on carefully observing the results, we deduce some key insights. Clean accuracy remains similar across different choices of surrogate models, but adversarial accuracy depends on the surrogate model of defender ($S_m^d$) and attacker ($S_m^a$).

We observe that the bigger the network size, the more accurate the surrogate models. With accurate surrogate models, the gradients with respect to the input are better estimated. Thus better black-box attacks and defenses can be obtained using the bigger architectures for surrogate models. For better defense, $S_m^d$ should have a relatively higher capacity than $S_m^a$. This can be confirmed by rows 1, 4,

Table 5. Performance of DBMA with and without regenerator network across different values of $k$. For low values, regenerator network improves both clean and adversarial accuracy. For $k=16$, small decrease in clean accuracy, but adversarial accuracy increases significantly.

| Surrogate model (attacker) | Coefficients ($k\%$) | Method | Black Box Model : Alexnet Surrogate Model (defense): Resnet-18 | | | |
|---|---|---|---|---|---|---|
| | | | clean | BIM | PGD | Auto Attack |
| Alexnet-half | 1 | WNR | 42.75 | 14.89 | 13.79 | 21.41 |
| | | WNR + $R_n$ | 56.08 | 30.58 (↑ 15.69) | 30.14 (↑ 16.35) | 37.04 (↑ 15.63) |
| | 2 | WNR | 50.17 | 17.34 | 16.38 | 25.36 |
| | | WNR + $R_n$ | 60.35 | 34.94 (↑ 17.60) | 34.61 (↑ 18.23) | 41.61 (↑ 16.25) |
| | 4 | WNR | 59.14 | 21.99 | 20.77 | 29.43 |
| | | WNR + $R_n$ | 65.82 | 39.65 (↑ 17.66) | 39.62 (↑ 18.85) | 46.94 (↑ 17.51) |
| | 8 | WNR | 69.89 | 24.9 | 23.54 | 33.04 |
| | | WNR + $R_n$ | 70.37 | 41.89 (↑ 16.99) | 42.21 (↑ 18.67) | 49.88 (↑ 16.84) |
| | 16 | WNR | 77.92 | 26.66 | 24.55 | 34.02 |
| | | WNR + $R_n$ | 73.77 | **42.71** (↑ 16.05) | **42.71** (↑ 18.16) | **50.63** (↑ 16.61) |
| | 50 | WNR | 82.58 | 11.36 | 8.23 | 17.34 |
| | | WNR + $R_n$ | 75.19 | 33.12 (↑ 21.76) | 31.60 (↑ 23.37) | 40.11 (↑ 22.77) |
| Alexnet | 1 | WNR | 42.75 | 10.20 | 8.72 | 15.80 |
| | | WNR + $R_n$ | 56.08 | 24.02 (↑ 14.82) | 23.13 (↑ 14.41) | 30.55 (↑ 14.75) |
| | 2 | WNR | 50.17 | 15.37 | 14.14 | 21.92 |
| | | WNR + $R_n$ | 60.34 | 32.89 (↑ 17.52) | 32.24 (↑ 18.10) | 40.50 (↑ 18.58) |
| | 4 | WNR | 59.14 | 17.54 | 16.03 | 25.08 |
| | | WNR + $R_n$ | 65.82 | 31.00 (↑ 13.46) | 30.85 (↑ 14.82) | 38.26 (↑ 13.18) |
| | 8 | WNR | 69.89 | 19.65 | 18.30 | 27.73 |
| | | WNR + $R_n$ | 70.37 | 34.13 (↑ 14.48) | 33.61 (↑ 15.31) | 41.08 (↑ 13.35) |
| | 16 | WNR | 77.92 | 15.98 | 14.04 | 21.34 |
| | | WNR + $R_n$ | 73.77 | **33.31** (↑ 17.33) | **31.72** (↑ 17.68) | **40.56** (↑ 19.22) |
| | 50 | WNR | 82.58 | 5.58 | 3.33 | 10.44 |
| | | WNR + $R_n$ | 75.19 | 19.65 (↑ 14.07) | 18.38(↑ 15.04) | 25.41 (↑ 14.97) |

Table 6. Ablation on defense components of proposed DBMA and comparison with baseline. Both the components individually provide better defense than baseline. DBMA yields best performance when WNR and $R_n$ are used together.

| Surrogate model (attacker) | Defense Components | Black Box Model : Alexnet Surrogate Model (defense): Resnet-18 | | | |
|---|---|---|---|---|---|
| | | clean | BIM | PGD | Auto Attack |
| Alexnet-half | Baseline | 82.58 | 7.02 | 4.53 | 11.65 |
| | WNR | 77.92 | 26.66 (↑ 19.69) | 24.55 (↑ 20.02) | 34.02 (↑ 22.37) |
| | $R_n$ | 77.03 | 29.40 (↑ 22.38) | 28.32 (↑ 23.79) | 37.16 (↑ 25.51) |
| | WNR + $R_n$ | 73.77 | **42.71** (↑ 35.69) | **42.71** (↑ 38.18) | **50.63** (↑ 38.98) |
| Alexnet | Baseline | 82.58 | 4.17 | 2.19 | 8.55 |
| | WNR | 77.92 | 15.98 (↑ 11.81) | 14.04 (↑ 11.85) | 21.34 (↑ 12.79) |
| | $R_n$ | 77.03 | 16.52 (↑ 12.35) | 15.09 (↑ 12.9) | 22.40 (↑ 13.85) |
| | WNR + $R_n$ | 73.77 | **33.31** (↑ 29.14) | **31.72** (↑ 29.53) | **40.56** (↑ 32.01 ) |

and 7, where the defense becomes more effective on increasing the $S_m^d$'s capacity against various attacks using Alexnet-half as $S_m^a$. A similar trend is observed against the attacks crafted using Alexnet and Resnet18. For the other way around, i.e., when $S_m^a$ has relatively higher capacity than $S_m^d$, more powerful attacks can be crafted. This is evident from rows 1, 2, and 3, where stronger adversarial samples are obtained on increasing the capacity of $S_m^a$ for a given $S_m^d$. For instance, for Resnet18 as $S_m^d$, stronger attacks (lower adversarial accuracy) are obtained by using Resnet18 as $S_m^a$, followed by Alexnet and Alexnet-half. A similar pattern is observed on other choices of $S_m^d$.

## 8. Our defense (DBMA) on larger black box model

Throughout all our experiments, we used Alexnet as the black-box model $B_m$ . To check the consistency of our approach DBMA across different architecture, especially for bigger and high-capacity networks, we perform experiments using Resnet34 as black-box model and report corresponding results in Table 8. With Resnet18 and Alexnet as the defender's and attacker's surrogate model ($S_m^d$ and $S_m^a$)

Table 7. Investigating the effect of surrogate model's architecture (for both defender and attacker) on the performance of our proposed approach (DBMA). Given defender's surrogate model, the attack is stronger if larger surrogate model is used by the attacker.

| Surrogate model (defender) | Surrogate model (attacker) | Black Box Model : Alexnet | | | |
|---|---|---|---|---|---|
| | | clean | BIM | PGD | Auto Attack |
| Resnet-18 | Alexnet-half | 73.77 | **42.71** | **42.71** | **50.63** |
| Resnet-18 | Alexnet | 73.77 | 33.31 | 31.72 | 40.56 |
| Resnet-18 | Resnet-18 | 73.77 | 22.48 | 21.93 | 29.48 |
| Alexnet-half | Alexnet-half | **74.94** | 38.98 | 39.3 | 47.83 |
| Alexnet-half | Alexnet | 74.94 | 29.04 | 27.58 | 35.37 |
| Alexnet-half | Resnet-18 | 74.94 | 20.72 | 19.11 | 26.79 |
| Alexnet | Alexnet-half | 74.67 | 40.08 | 39.6 | 48.5 |
| Alexnet | Alexnet | 74.67 | 29.49 | 28.63 | 36.93 |
| Alexnet | Resnet-18 | 74.67 | 21.51 | 20.24 | 28.44 |

Table 8. Performance of our approach DBMA in defending the larger black-box network (i.e., Resnet34). Across various combinations, DBMA shows a consistent improvement against the different attacks with a small drop in the clean accuracy.

| Method | Surrogate Model (Defender) | Surrogate Model (Attacker) | Black Box Model : Resnet34 | | | |
|---|---|---|---|---|---|---|
| | | | Clean | BIM | PGD | Auto-Attack |
| Baseline | - | Resnet18 | 95.66 | 3.11 | 1.23 | 11.82 |
| DBMA (**Ours**) | Alexnet | Resnet18 | 87.06 | **20.76** | **17.33** | **25.11** |
| Baseline | - | Alexnet | 95.66 | 21.36 | 12.83 | 27.85 |
| DBMA (**Ours**) | Resnet18 | Alexnet | 88.40 | **48.16** | **44.80** | **56.65** |

Table 9. Performance of our approach DBMA with fourier transform and wavelet transform based noise removal technique (FNR and WNR, respectively). WNR defense outperforms the FNR across different attacks. Also, WNR-based DBMA (WNR + $R_n$) yields more significant gains in performance on CIFAR10.

| Surrogate model (attacker) | Method | Black Box Model : Alexnet<br>Surrogate Model (defense): Resnet-18 | | | |
|---|---|---|---|---|---|
| | | clean | BIM | PGD | Auto Attack |
| Alexnet-half | Baseline | 82.58 | 7.02 | 4.53 | 11.65 |
| | FNR | 79.14 | 13.30 (↑ 6.28) | 10.86 (↑ 6.33) | 20.47 (↑ 8.82) |
| | FNR+ $R_n$ | 74.13 | 28.38 (↑ 21.36) | 27.05 (↑ 22.52) | 35.47 (↑ 23.82) |
| | WNR | 77.92 | 26.66 (↑ 19.64) | 24.55 (↑ 20.02) | 34.02 (↑ 22.37) |
| | WNR + $R_n$ (**Ours**) | **73.77** | **42.71** (↑ 35.69) | **42.71** (↑ 38.18) | **50.63** (↑ 38.98) |
| Alexnet | Baseline | 82.58 | 4.17 | 2.19 | 8.55 |
| | FNR | 79.14 | 5.87 (↑ 1.70) | 4.03 (↑ 1.84) | 10.97 (↑ 2.42) |
| | FNR+ $R_n$ | 74.13 | 19.24 (↑ 15.07) | 17.88 (↑ 15.69) | 24.55 (↑ 16.00) |
| | WNR | 77.92 | 15.98 (↑ 11.81) | 14.04 (↑ 11.85) | 21.34 (↑ 12.79) |
| | WNR + $R_n$ (**Ours**) | **73.77** | **33.31** (↑ 29.14) | **31.72** (↑ 29.53) | **40.56** (↑ 32.01) |

respectively, we observe the improvement of $\approx 27-32\%$ in adversarial accuracy across attacks compared to baseline (rows $3^{rd}$ and $4^{th}$). With Alexnet as the defender's surrogate and Resnet18 as the attacker's surrogate model, we get an improvement of $\approx 17-23\%$ across attacks (rows $1^{st}$ and $2^{nd}$). As observed in previous experiments, compared to baseline, clean accuracy drops by $\approx 7-8\%$. Overall, across different black box models, our proposed defense DBMA has obtained decent performance. Hence we conclude, DBMA is even effective on bigger black box architectures.

## 9. Comparison with Fourier Transform based Noise removal

Apart from the wavelet transformations, some recent works utilised the fourier transformations to remove the adversarial noise from the adversarial images, and further found it to be effective in denoising [8]. In this section, we do an ablation on our choice of Wavelet-based Noise Remover (WNR) over the other possible choice of fourier-based Noise Remover (FNR).

As observed in recent works [8], adversarial attack affects the high-frequency components more than low-

frequency components. Therefore, In FNR we apply a low pass filter on an image with threshold radius $\hat{r}$. Similar to WCSM, we compute $LCR_C$ , $LCR_A$, $LCR$ and $ROC$ for different values of $r$. The optimal value of $r$ (i.e., $\hat{r}$) is selected at which $ROC$ starts saturating ($\hat{r} = 11$). In Table 9, we observe, compared to baseline, Fourier-based DBMA gives an improvement of $\approx 21 - 34\%$ in adversarial accuracy across attacks for Alexnet half (rows 1 and 3).Compared to Fourier-based DBMA, the results using our wavelet-based DBMA are significantly better in terms of adversarial accuracy (rows 1 and 5) with similar clean performance. Similarly, for Alexnet, Wavelet-based DBMA gives $\approx 16 - 24\%$ better adversarial accuracy compared to Fourier based DBMA across all attacks (rows 8 and 10). Hence, our wavelet-based DBMA is more robust across different adversarial attacks than Fourier-based DBMA.

## 10. Visualization

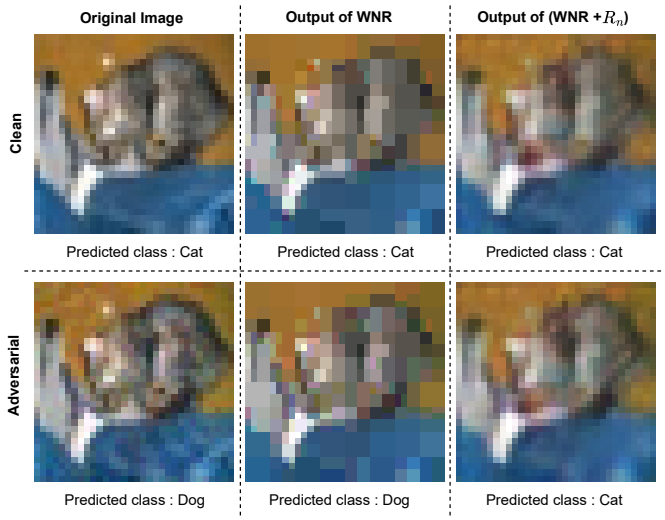


Figure 2. Visualization of images: The top row indicates input as clean image and bottom row corresponds to adversarial image. The predictions obtained by the black-box network on inputs: (a) Original clean image (b) Output of wavelet noise remover on clean image (c) Output of WNR with regenerator network ($R_n$) on clean image (d) Original adversarial image (e) Output of wavelet noise remover (WNR) on adversarial image (f) Output of WNR with regenerator network ($R_n$) on adversarial Image. Here, the ground truth class is Cat. Our method (DBMA) produces correct output using regenerated image as input.

## 11. Algorithm

**Algorithm 1** Algorithm for our proposed method (DBMA)

**Require:** Black box model $B_m$, max coefficients $k^{max}$
**Ensure:** $\hat{B}_m$
**Step 1:** ***Model Stealing***
Surrogate model $S_m$, Synthetic data $S_d \leftarrow$ Model Stealing on $B_m$
**Step 2:** ***Wavelet Coefficient Selection Module***
Obtain adversarial samples ($S_{da}$) corresponding to $S_d$ using adversarial attack on $S_m$
**for** $k = 1 : k^{max}$: **do**
  $\bar{S}^k_{da} \leftarrow$ WNR($S_{da}$,$k$)
  $N_{flips} = 0$
  **for** $i = 1 : (|\bar{S}^k_{da}|)$ **do**
    $\bar{x}^i_{sa} \leftarrow \bar{S}^k_{da}[i]$ $\{i^{th}$ element of $\bar{S}^k_{da}\}$
    $x^i_s \leftarrow S_d[i]$ $\{i^{th}$ element of $S_d\}$
    **if** $label(B_m(\bar{x}^i_{sa})) \neq label(B_m(x^i_s))$ **then**
      $N_{flips} = N_{flips} + 1$
    **end if**
  **end for**
  $LFR^k = N_{flips}/|S_d|$
**end for**
$\hat{k} = \underset{k}{\text{argmin}} \;\; LFR^k$
**Step 3:** ***Training Regenerator network*** $(R_n)$
$\bar{S}^{\hat{k}}_d \leftarrow$ WNR($S_d$,$\hat{k}$)
$\bar{S}^{\hat{k}}_{da} \leftarrow$ WNR($S_{da}$,$\hat{k}$)
Initialize $R^\theta_n$
**for** $epoch < MaxEpoch$ **do**
  **for** $i = 1 : (|S_d|)$ **do**
    $x^i_s \leftarrow S_d[i]$ $\{i^{th}$ element of $S_d\}$
    $\bar{x}^i_s \leftarrow \bar{S}^{\hat{k}}_d[i]$ $\{i^{th}$ element of $\bar{S}^{\hat{k}}_d\}$
    $\bar{x}^i_{sa} \leftarrow \bar{S}^{\hat{k}}_{da}[i]$ $\{i^{th}$ element of $\bar{S}^{\hat{k}}_{da}\}$
    $L_{cs} = CS(S_m(R_n(\bar{x}^i_s)), S_m(x^i_s))$ $\{CS$ is cosine similarity$\}$
    $L_{kl} = KL(soft(S_m(R_n(\bar{x}^i_{sa}))), soft(S_m(R_n(\bar{x}^i_s))))$ $\{KL$ is KL divergence$\}$
    $L_{sc} = \left\|R_n(\bar{x}^i_s) - x^i_s\right\|_1 + \left\|R_n(\bar{x}^i_{sa}) - x^i_s\right\|_1$
    $L(R^\theta_n) = -\lambda_1 L_{cs} + \lambda_2 L_{kl} + \lambda_3 L_{sc}$
    Update $R^\theta_n$ by minimizing $L(R^\theta_n)$ using Adam Optimizer
  **end for**
**end for**
$\hat{B}_m \leftarrow$ concatenate($WNR(.,\hat{k})$, $R^{\theta^*}_n$, $B_m$) {The black box model $\hat{B}_m$ is used by attacker}
**return** $\hat{B}_m$

Table 10. Attack parameters for different adversarial attacks: BIM, PGD and Auto Attack

| Dataset | Attack Parameters | Adversarial Attacks | | |
|---|---|---|---|---|
| | | BIM | PGD | Auto Attack |
| CIFAR-10 | $\epsilon$ | 8/255 | 8/255 | 8/255 |
| | $\epsilon_{step}$ | 0.00156 ($\epsilon$/no of iterations) | 2/255 | – |
| | no of iterations | 20 | 20 | – |
| SVHN | $\epsilon$ | 4/255 | 4/255 | 4/255 |
| | $\epsilon_{step}$ | 2/255 | 2/255 | – |
| | no of iterations | 20 | 20 | – |

## 12. Adversarial Attack Parameters and Training Details

We evaluate the performance of black box model ($B_m$) on three adversarial attacks, PGD, BIM and Auto Attack. Parameters used for each attack are summarized in Table 10.

**Training details of Regenerator Network** ($R_n$): The regenerator network is trained with Adam optimizer with learning rate of $0.0002$ for 300 epochs. Learning rate is decayed using linear scheduler, where we keep the learning rate fixed for 100 epochs and then linearly decay the rate to zero. Batch size is set to 128.

## References


[1] Antonio Barbalau, Adrian Cosma, Radu Tudor Ionescu, and Marius Popescu. Black-box ripper: Copying black-box models using generative evolutionary algorithms. *Advances in Neural Information Processing Systems*, 33:20120–20129, 2020. 1

[2] Gregory Beylkin, Ronald R. Coifman, and Vladimir Rokhlin. Fast wavelet transforms and numerical algorithms i. *Communications on Pure and Applied Mathematics*, 44:141–183, 1991. 1

[3] Albert Cohen, Ingrid Daubechies, and J. C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 45:485–560, 1992. 1

[4] Ingrid Daubechies. Ten lectures on wavelets. *Computers in Physics*, 6:697–697, 1992. 1

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[6] Sunandini Sanyal, Sravanti Addepalli, and R Venkatesh Babu. Towards data-free model stealing in a hard label setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15284–15293, 2022. 1

[7] Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. Data-free model extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4771–4780, 2021. 1

[8] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019. 5