

Exploring the Impact of Frequency Components on Adversarial Patch Attacks against an Image Classifier Model

Aran Chindaudom
Division of Transdisciplinary Science
JAIST
Nomi, Japan
s2150008@jaist.ac.jp

Prarinya Siritanawan
School of Information Science
JAIST
Nomi, Japan
prarinya@jaist.ac.jp

Kazunori Kotani
Division of Transdisciplinary Science
JAIST
Nomi, Japan
ikko@jaist.ac.jp

Abstract—Exceptional advancements in various computer vision tasks, such as identifying and categorizing objects, have been realized through the use of deep learning models, with a particular emphasis on convolutional neural networks (CNNs). Yet, while these models deliver outstanding results, they remain vulnerable to adversarial examples, thereby raising questions about their safety and dependability. In this paper, we investigate the influence of the image characteristics on the efficacy of adversarial patch attack against an image classifier model. We analyzed such characteristics in the frequency domain, where the frequencies indicate the periodicity and information density that contribute to the efficacy of adversarial patches. Our results showed that low-frequency components had significant contribution to the effectiveness of adversarial patch attacks.

Index Terms—Adversarial Patch, Fourier Analysis, Convolutional Neural Networks, Adversarial Examples

I. INTRODUCTION

Exceptional advancements in various computer vision tasks, such as identifying and categorizing objects, have been realized through the use of deep learning models, with a particular emphasis on convolutional neural networks (CNNs). However, despite their impressive performance, these models are susceptible to adversarial attacks, posing security and reliability concerns. Among various attack strategies, adversarial patches have gained significant attention due to their practical applicability and effectiveness in fooling deep learning models.

While previous studies have primarily focused on the design and evaluation of adversarial patches themselves, the impact of the image characteristics in which they are applied remains under-explored. By providing a comprehensive analysis of this relationship, we aim to deepen the understanding of adversarial patch effectiveness and contribute to the development of more robust deep learning systems.

II. RELATED WORKS

Deep learning represents a distinct category within machine learning methodologies, empowering intelligent systems to discern high-level semantics and vital patterns from extensive data sets. Traditional machine learning methods often faced difficulties in feature extraction due to their incapacity to manage high-dimensional input data [1] and computational limitations [2]. On the other hand, Convolutional Neural Networks (CNNs), a specific type of deep learning, managed to overcome these hurdles by defining

complex associations among a multitude of fundamental visual elements using artificial neurons. These connections between edges and distinct structural features are traced through several hidden layers. Therefore, the efficacy of Convolutional Neural Networks (CNNs) deep learning systems sees a significant boost when provided with ample training data.

Neural networks are built upon intricate networks of interrelations among numerous layers of neural units. Every layer consists of several artificial neurons that employ an activation function to transform high-dimensional input into one or more output values. A conventional neural network can be expressed mathematically in the following manner:

$$g(\mathbf{X}) = g^{(k)}(\dots g^{(2)}(g^{(1)}(\mathbf{X}))) \quad (1)$$

In this equation, \mathbf{X} signifies an input image, while $g^{(i)}$ represents a function associated with the i^{th} layer of the network, where $i = 1, 2, \dots, k$.

Models such as LeNet[3], AlexNet[4], VGG[5], ResNet[6], and Inception[7][8][9] find frequent application in deep learning tasks associated with computer vision. Often becoming the focus of adversarial examples, these models are commonly examined using prominent datasets in computer vision such as MNIST, CIFAR-10, and ImageNet. Handwritten digits form the collection of MNIST [10], while CIFAR-10 and ImageNet are employed for classification tasks. With its 60,000 images in 10 categories [11], CIFAR serves as a rich resource, while ImageNet's 14,196,122 images span across 1,000 classes [12]. Given ImageNet's sheer volume, studies on adversarial examples are often restricted to its smaller subsets.

The discussion around adversarial examples in machine learning has persisted for more than ten years, frequently revolving around systems that use handcrafted characteristics, such as intrusion detection and spam filters. The concept of adversarial examples as a face-off between adversaries and classifiers was introduced by Dalvi et al. [13]. Later, the L-BFGS method for creating adversarial examples was proposed by Szegedy et al. [14], but the approach was viewed as being computationally demanding and not practical due to its reliance on linear search. A more efficient alternative, the Fast Gradient Sign Method (FGSM), was put forward by Goodfellow et al. [15]. They proposed adding a gradient

sign of the cost function to each pixel. This process can be expressed as:

$$\eta = \epsilon \text{sign}(\nabla_{\mathbf{X}} J(\theta, \mathbf{X}, \hat{y})) \quad (2)$$

In this equation, η signifies the perturbation's extent, calculated through gradients obtained during the back-propagation process. Using an original image \mathbf{X} , the manipulated adversarial example \mathbf{X}' can be produced by introducing the perturbation η to the image: $\mathbf{X}' = \mathbf{X} + \eta$.

[16] introduced the notion of Adversarial patch attacks to create universal adversarial examples suitable for practical real-world applications, particularly in image classifiers. They crafted these patches as robust, universal adversarial samples which could be physically adhered to any real-world environment using printable materials. This, in turn, caused classifiers to make erroneous predictions for any target class. Image classifiers are typically programmed to concentrate on the most prominent patterns within an image. Capitalizing on this aspect, adversarial patches infuse real-world input images with smaller, yet considerably more noticeable patterns.

[17] primarily investigated adversarial examples, demonstrated that a model can maintain high accuracy when utilizing only low-frequency information in the Fourier domain. This finding is relevant to our work [18]. As shown in Fig. 1, we have discovered that image classes with more low-frequency information, such as the Monarch butterfly, are more resistant to adversarial patch attacks compared to image classes with background-like features, like Valleys and Forelands.

Index	Class	Winrate (%)	Sample Images
979	Valley	95	
976	Foreland	94	
978	Seashore	89.30	
885	Velvet	87.65	
975	Lakeside	85.45	
..
..
323	Monarch	5.25	
685	Odometer	4.86	

Fig. 1: Attack success rate (winrate) of the patch against different types of classes. Monarch class images are much harder to fool than Valley class

[19] aimed to counter adversarial patches by employing a sliding window to extract a set of ally patches based on the information content of the input image. The classifier's labeling is determined by voting through multiple ally patches. However, this method exhibits a limitation when dealing with background-like images, as the ally patches with true labels have much lower information, causing the adversarial patch to remain dominant. [19] acknowledged this limitation, stating that in monotonous background images, the adversarial patch will be dominant, and most non-adversarial patches will fail to fulfill the constraint.

[20] investigated the concept of frequency regularization to increase adversarial robustness. When comparing natural

and adversarial inputs, it became apparent that the majority of differences lay in the low-frequency region. As a result, [20] proposed frequency regularization (FR), a process that balances the outputs of both natural and adversarial inputs within the frequency domain, which in turn, results in an improved robust accuracy of the model. This work provides additional support for the assertion that image classes possessing substantial low-frequency content are inherently more resilient to adversarial patch attacks.

Two works, [21] and [22], had proposed optimizations related to adversarial patches. [21] had investigated the significance of shape information on deep neural networks' robustness in the context of adversarial attacks, demonstrating that their optimized shapes performed better at smaller scales. However, they did not consider the significance of background images when adversarial efficacy was concerned, as they chose random images for their testing. [22] had proposed a method to simultaneously optimize the position and perturbation for an adversarial patch, obtaining high attack success rates in the black-box setting. They fixed the scale and shape of the patch, which left room for further investigation into the impact of scale optimization on adversarial efficacy.

In this paper, our paper seeks to fill this knowledge gap by systematically examining the image characteristics in frequency domain, where the frequencies indicate the periodicity and information density that contribute to the efficacy of adversarial patches. Our findings will not only provide valuable insights into the vulnerability of deep learning models to adversarial patches in various visual context but also contribute to the development of more robust and reliable defense mechanisms against such attacks.

III. METHODOLOGY

A. Adversarial Patch

The adversarial patch \mathbf{P}' is computed by maximizing the expectation of the [16] function as follows:

$$\mathbf{P}' = \underset{\mathbf{P}}{\operatorname{argmax}} E_{\mathbf{X} \sim \mathbb{X}, o \sim \mathbb{O}, l \sim \mathbb{L}} [\log \Pr(\hat{y} | a(\mathbf{P}, \mathbf{X}, o, l))] \quad (3)$$

where \hat{y} represents the target image classifier model's confidence of the training image being the target class, where the patch operator a is applied over the distribution of \mathbb{X} training set images. The patch location l and the randomized patch rotated at random orientation o are also varied over the distribution of locations \mathbb{L} and orientations \mathbb{O} to improve the patch's resiliency against varied placement locations and angles. \mathbf{X} represents an image from the training set \mathbb{X} . \mathbf{P} represents the adversarial patch that, when added to the original image \mathbf{X} , is designed to cause a machine learning model to misclassify the image.

B. Adversarial Patch Training

We initiated the adversarial patch, denoted as \mathbf{P}' , as a square patch with a width equivalent to 0.3 of the InceptionV3's input size ($299 \times 299 \times 3$), resulting in a patch size of $89 \times 89 \times 3$. We trained these patches by incorporating both the concepts of random location and random rotation.

We selected a subset of the ImageNet validation dataset, denoted as \mathbb{X} , which consisted of 10 correctly classified

images for each of the 1000 ImageNet classes. Hence, the total number of images in the set \mathbb{X} was 10,000.

The training step involved forward propagation where we overlaid the adversarial patch \mathbf{P}' on 4 images at a time, from the set \mathbb{X} , at the random location l , and rotated at orientation o . We optimize the patch minimizing the following loss function:

$$\mathcal{L} = -\log(p_{\hat{y}}) \quad (4)$$

Where $p_{\hat{y}}$ is the predicted probability of the classifier model for the input sample belonging to the target class \hat{y} .

The averaged gradient across the four inputs was then back-propagated to update the patch. Since our prior experiments indicated further iterations beyond 15,000 resulted in negligible improvement, We repeated these steps until we exhausted all images in the set 12 times, this allowed us to achieve 30,000 iterations of updates.

We replicated these steps for each of the 1,000 target classes, thereby creating 1,000 distinct adversarial patches.

C. Adversarial Patch Testing

Let \mathbb{P} be the set of adversarial patches, where each patch $\mathbf{P}_{\hat{y}}$ is trained to fool the classifier into predicting a specific target class \hat{y} . Let $m(\cdot)$ be the pre-trained InceptionV3 model, and $\hat{\mathbb{X}}$ be the set of images from the ImageNet validation dataset to be tested. The images from $\hat{\mathbb{X}}$ had been excluded \mathbb{P} 's training process, are appropriately prepared for model input, where the original class of each image is correctly predicted by the model $m(\cdot)$ without the patch.

For each image $\mathbf{X} \in \hat{\mathbb{X}}$ and each patch $\mathbf{P}_{\hat{y}} \in \mathbb{P}$, we:

- 1) Overlay the patch $\mathbf{P}_{\hat{y}}$ onto the image \mathbf{X} at a random location, creating the attacked image $\mathbf{X}' = o(\mathbf{X}, \mathbf{P}_{\hat{y}})$, where $o(\cdot)$ is the overlay function.
- 2) Use $m(\cdot)$ to classify \mathbf{X}' , obtaining the predicted class $y_{\text{pred}} = m(\mathbf{X}')$.

We then assess the effectiveness of each target class \hat{y} :

- Success: The classifier is fooled by the patch and predicts the target class, represented as

$$s(\hat{y}) = \left| \{ \mathbf{X} \in \hat{\mathbb{X}} \mid m(o(\mathbf{X}, \mathbf{P}_{\hat{y}})) = \hat{y} \} \right| \quad (5)$$

where $|\cdot|$ denotes the cardinality (size) of the set.

- Failure: The classifier is not fooled by the patch and either predicts the original class of the image or some other class, represented as

$$d(\hat{y}) = \left| \{ \mathbf{X} \in \hat{\mathbb{X}} \mid m(o(\mathbf{X}, \mathbf{P}_{\hat{y}})) \neq \hat{y} \} \right| \quad (6)$$

This process is repeated for all images in $\hat{\mathbb{X}}$ and all patches in \mathbb{P} , producing a success and failure count for each target class.

The adversarial efficacy in the form attack success rate, E , for each target class \hat{y} , can be defined as:

$$E_{\hat{y}} = \frac{s(\hat{y})}{s(\hat{y}) + d(\hat{y})} \times 100 \quad (7)$$

This gives the effectiveness as a percentage, where $s(\hat{y})$ is the number of successful attacks, and $s(\hat{y}) + d(\hat{y})$ is the total number of attacks for the target class \hat{y} . A higher $E_{\hat{y}}$ value

indicates that the adversarial patch $\mathbf{P}_{\hat{y}}$ is more effective at fooling the classifier m into predicting the target class \hat{y} .

D. Frequency-Filtered Patches

In this experiment, we are testing the attack efficacy of adversarial patches under the influence of different types of frequency filters: low-pass, high-pass, and band-elimination filters. We apply these filters with different cutoff frequencies r to the adversarial patches before overlaying them onto the original images.

To perform operations in the frequency domain, the image must be shifted from the spatial domain using the Discrete Fourier Transform (DFT). Consider $\mathbf{X}[m, n]$ as a single channel image where (m, n) represent spatial coordinates. The mathematical definition of DFT is as follows:

$$\mathcal{F}(u, v) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathbf{X}[m, n] \cdot e^{-i2\pi\left(\frac{um}{M} + \frac{vn}{N}\right)} \quad (8)$$

where (u, v) are the frequencies in the frequency domain and M and N are the width and height of the image respectively.

The DFT produces a complex-valued output; The amplitude function $A(u, v)$ gives the magnitude of each frequency component, and the phase function $\phi(u, v)$ gives the phase shift of each frequency component.

$$A(u, v) = \sqrt{\text{Re}[\mathcal{F}(u, v)]^2 + \text{Im}[\mathcal{F}(u, v)]^2} \quad (9)$$

$$\phi(u, v) = \arctan\left(\frac{\text{Im}[\mathcal{F}(u, v)]}{\text{Re}[\mathcal{F}(u, v)]}\right) \quad (10)$$

For an adversarial patch \mathbf{P} , we obtain its frequency domain representation by performing a Fourier Transform \mathcal{F} on each color channel of the patch:

$$\mathcal{F}(\mathbf{P}) = \hat{A}(u, v) \cdot e^{i\phi'(u, v)} \quad (11)$$

Where $\hat{A}(u, v)$ gives the amplitude of frequency (u, v) in a color channel of patch \mathbf{P} , and $\phi'(u, v)$ gives the phase of frequency (u, v) in a color channel of patch \mathbf{P} .

We then apply the filter function f to the amplitude function of each color channel of the patch. This is represented by:

$$A'(u, v) = (f \circ \hat{A})(u, v) \quad (12)$$

In this equation, $A'(u, v)$ is the filtered amplitude function of the patch.

Low-pass filter: For the low-pass filter, we allow frequencies within a certain square range around the DC component to pass.

$$f(A(u, v)) = \begin{cases} A(u, v), & \text{if } |u| \leq r \wedge |v| \leq r \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

Here, r is the cutoff frequency for the low-pass filter.

High-pass filter: For the high-pass filter function with cutoff frequency r , we also allow frequencies DC component to pass. The filter function is defined as:

$$f(A(u, v)) = \begin{cases} A(u, v), & \text{if } |u| > r \wedge |v| > r \\ A(u, v), & \text{if } u = 0 \wedge v = 0 \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

Band-Elimination filter: We eliminate frequencies within a certain range from the DC component. The filter function is defined as:

$$f(A(u, v)) = \begin{cases} 0, & \text{if } |u| > r \wedge |v| > r \\ 0, & \text{if } |u| \leq r + \omega \wedge |v| \leq r + \omega \\ A(u, v), & \text{otherwise} \end{cases} \quad (15)$$

Here, r denotes the lower-cutoff frequency, and ω is the width of the elimination band.

After the application of filters, we convert \mathbf{P} back to the spatial domain using the inverse Fourier transform F^{-1} to obtain the filtered patch:

$$\mathbf{P} = \mathcal{F}^{-1}(A'(u, v) \cdot e^{i\phi'(u, v)}) \quad (16)$$

Attack success rate E_r for each filtered patch $\mathbf{P}_{\hat{y}}$ with varying r cut-off frequencies, is then calculated using the same method as Section III-C.

IV. EXPERIMENTAL RESULTS

Despite maintaining consistency in the training parameters and data, the success rate of adversarial attacks showed a broad range from 96% to 0%, even under optimal conditions. This varied depending on the specific class targeted by the patch. Fig. 2 presents the $E_{\hat{y}}$ in a ranked order for all of the 1000 potential targeted class in the adversarial patch attack against the ImageNet classification dataset.

From the results, over 400 of the targeted classes displayed an attack success rate of less than 1%. Therefore, it is evident that the choice of target classes for the adversarial patch plays a crucial role in determining the overall effectiveness of the attack.

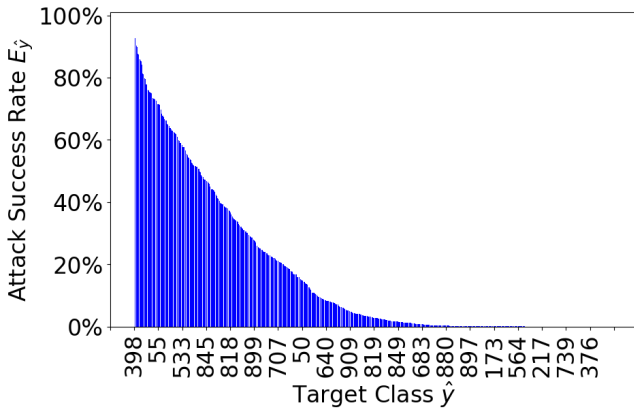


Fig. 2: $E_{\hat{y}}$ for all 1000 potential targeted classes in the adversarial patch attack against the ImageNet classification dataset. Over 400 of these targeted classes displayed an attack success rate of less than 1%.

Next, we then test for E_r by corrupting the adversarial patches through various cutoff frequencies r .

Since our patches are trained with a width and height of 89×89 , which is a 0.3 scale of input size for Inceptionv3, we attempted to corrupt from 0 to 44 Hz with both high-pass and low-pass filters and tested the patch efficacy under different magnitudes of high and low frequency corruptions.

However, for the band-elimination filter, we can only vary r from 0 Hz to 42 Hz. This is because the band-elimination filter cuts out a band of frequencies around the cutoff frequency, and when r is larger than 42, the band extends beyond the boundaries of the frequency space.

Fig. 3, Fig. 4 and Fig. 5 show the spatial domain before and after we have applied the different type of filters to an adversarial patch.

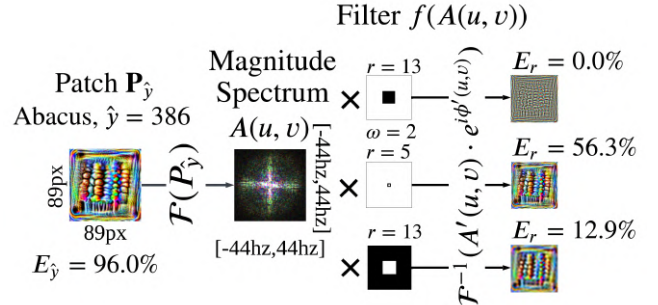


Fig. 3: The spatial domain, attack success rate E_r of the adversarial patch targeting ImageNet class 386 (abacus) before and after applying different filters. The filters were used to corrupt the patch before testing its efficacy.

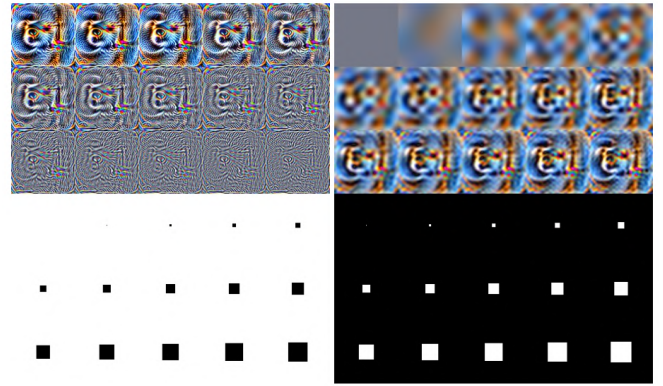


Fig. 4: The spatial domain of the adversarial patch after applying high-pass filters (left), and low-pass filters (right) from $r = 0$ to $r = 15$.

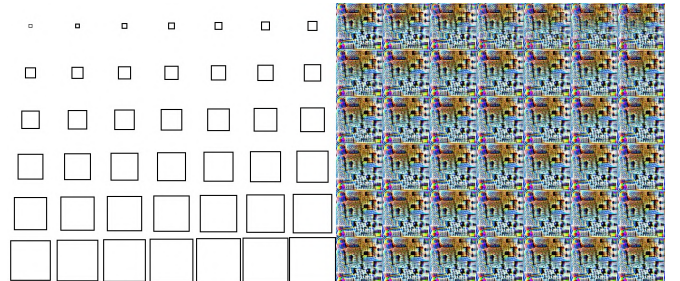


Fig. 5: The spatial domain of the adversarial patch after applying band-elimination filters from $r = 0$ to $r = 42$, changes to spatial domain are visually imperceptible.

To make the experiment manageable, we have chosen the 20 best performing target classes using the results obtained in Fig. 2. Our selection of the target classes is shown in

table I. Therefore, we have a total of 20 (target classes) \times 2 (filter types: low-pass and high-pass) \times 45 (cutoff frequency variations) + 5 (target classes) \times 1 (filter type: band-elimination) \times 43 (cutoff frequency variations) = 2,015 different patch configurations to test over $\hat{\mathcal{X}}$.

TABLE I: Top 20 ImageNet target classes in Fig. 2, sorted by highest $E_{\hat{y}}$





















\hat{y}	Target Class	$E_{\hat{y}}$	$\mathbf{P}_{\hat{y}}$
398	Abacus	96.27%	
918	Crossword	94.84%	
288	Leopard	92.66%	
944	Artichoke	91.41%	
30	Bullfrog	90.88%	
490	Chainmail	90.26%	
952	Fig	89.93%	
987	Corn	89.80%	
323	Monarch	87.67%	
37	Turtle	87.34%	
971	Bubble	87.19%	
124	Crayfish	86.26%	
397	Pufferfish	86.00%	
752	Racket	85.48%	
791	Cart	85.45%	
805	Soccerball	85.36%	
293	Cheetah	84.13%	
566	Horn	82.53%	
862	Torch	81.35%	
616	Knot	81.16%	

Fig. 6 shows the average attack success rate E_r across 20 top performing target classes after applying high-pass and low-pass filters. From the results, we can remove a large portion of the higher frequency information (with low-pass filter) of the image without significantly affecting the efficacy of the adversarial patch. However, removing low-frequency information (with high-pass filter) significantly decreased adversarial efficacy, even at very low portions.

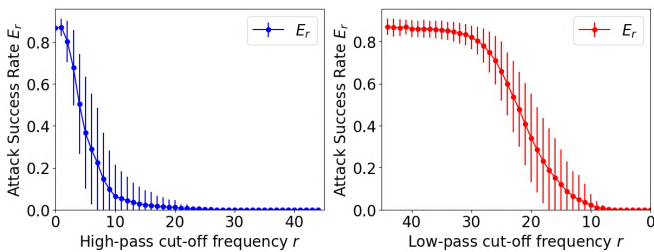


Fig. 6: Average E_r of top 20 patches after corrupting with high-pass filter (left), and low-pass (right) filter.

To identify the band of frequencies that had the most influence on the efficacy of the adversarial patch, we attempted to corrupt our top 20 patches by using band-elimination filters with a bandwidth of 2 Hz, shown in Fig. 5. Unlike high-pass and low-pass filters, the corruption was very small, invisible

to the eye. However, adversarial efficacy showed an almost 20-point drop in average efficacy after filtering out 3 to 5 Hz information at $r = 2$, as shown in Fig. 7.

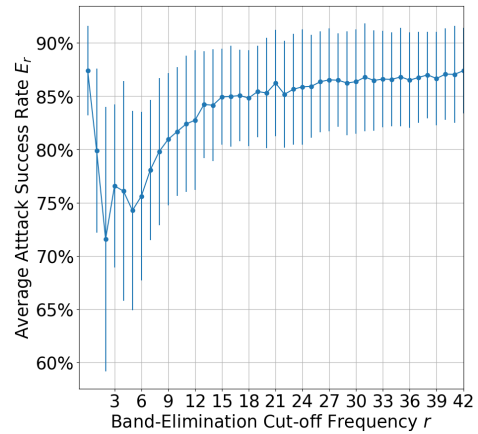


Fig. 7: Average E_r of top 20 patches for varying band-elimination filters at $\omega = 2$.

Additionally, through visual inspection, we have found that the magnitude spectrum of image classes that are most vulnerable to patch attacks has a consistent behavior at low frequencies when compared to classes that are much more resistant, where edge features are very well defined, resulting in a distinct characteristics at lower frequencies, shown in Fig. 8. This suggested that unlike the findings in other forms of adversarial attacks [23], image classification models may be particularly vulnerable to adversarial patch attacks that exploit low-frequency components.

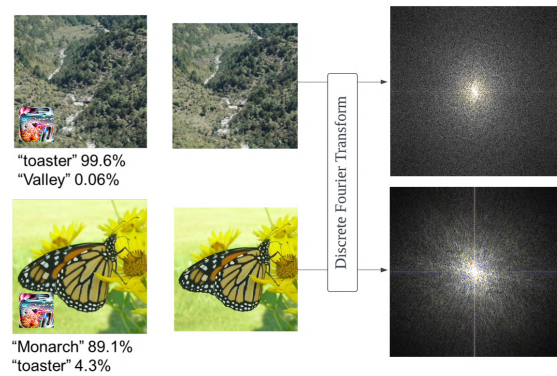


Fig. 8: Magnitude spectrum and InceptionV3 prediction of images that are highly resistive (bottom) and highly vulnerable (top) against adversarial patch attack.

V. CONCLUSION

We have demonstrated that low-frequency components play a pivotal role in determining the effectiveness of adversarial patch attacks. Our analysis of characteristics in the frequency domain revealed that adversarial patches were most effective when they retained their low-frequency components, affirming the low frequencies' dominance in adversarial patch attacks. Filtering out low-frequency information significantly decreased the patch's efficacy, even at very low portions. In contrast, the removal of high-frequency information from the adversarial patch had a relatively minimal

impact on the efficacy of the attack. We were also able to pinpoint a significant drop in adversarial efficacy upon specific information on the adversarial patch.

These findings offer insights for the development of more robust defense against adversarial attacks. It is crucial to note that adversarial attack effectiveness varies with the specific class targeted by the patch, underlining the need to consider the image characteristics when designing and evaluating both adversarial attacks and defenses. Future work should continue to investigate this complex interaction between image characteristics and adversarial attack efficacy. Our research serves as a stepping stone to further exploration of additional factors that could influence the performance of adversarial patches and their interactions with various elements of visual context. Such studies would contribute to the broader goal of building more secure and reliable deep learning systems in the face of adversarial threats.

REFERENCES

- [1] Y. Bengio and Y. Lecun, "Scaling learning algorithms toward ai," in *Large-Scale Kernel Machines*, L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, Eds. MIT Press, 2007, pp. 321–359.
- [2] D. Storcheus, A. Rostamizadeh, and S. Kumar, "A survey of modern questions and challenges in feature extraction," in *International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, vol. 44, PMLR, 2015, pp. 1–18.
- [3] Q. V. Le, M. Ranzato, R. Monga, *et al.*, "Building high-level features using large scale unsupervised learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
- [4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *25th International Conference on Neural Information Processing Systems (NIPS)*, vol. 1, 2012, pp. 1097–1105.
- [5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations, (ICLR)*, 2015.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.
- [7] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going Deeper with Convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in *31st AAAI Conference on Artificial Intelligence*, 2017.
- [10] *MNIST handwritten digit database*, Yann LeCun, Corinna Cortes and Chris Burges. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>.
- [11] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Toronto, Tech. Rep. TR-2009, 2009.
- [12] O. Russakovsky, J. Deng, H. Su, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015, ISSN: 0920-5691. DOI: 10.1007/s11263-015-0816-y.
- [13] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004, pp. 99–108, ISBN: 1581138881. DOI: 10.1145/1014052.1014066.
- [14] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations (ICLR)*, 2014.
- [15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *International Conference on Learning Representations (ICLR)*, 2015.
- [16] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial Patch," in *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.
- [17] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A fourier perspective on model robustness in computer vision," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.
- [18] A. Chindaudom, P. Siritanawan, K. Sumongkayothin, and K. Kotani, "Surreptitious adversarial examples through functioning qr code," *Journal of Imaging*, 2022.
- [19] A. Abdel-Hakim, "Ally patches for spoliation of adversarial patches," *Journal of Big Data*, vol. 6, p. 51, Jun. 2019. DOI: 10.1186/s40537-019-0213-4.
- [20] B. Huang, C. Tao, R. Lin, and N. Wong, *Frequency regularization for improving adversarial robustness*, 2022. arXiv: 2212.12732 [cs.CV].
- [21] Z. Chen, B. Li, S. Wu, J. Xu, S. Ding, and W. Zhang, "Shape matters: Deformable patch attack," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 529–548, ISBN: 978-3-031-19772-7.
- [22] X. Wei, Y. Guo, J. Yu, and B. Zhang, *Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks*, 2022.
- [23] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A fourier perspective on model robustness in computer vision," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.