PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Investigating robustness of adversarial camouflage (AC) for naval vessels

Løkken, Kristin Hammarstrøm, Aurdal, Lars, Brattli, Alvin, Palm, Hans Christian

Kristin Hammarstrøm Løkken, Lars Aurdal, Alvin Brattli, Hans Christian Palm, "Investigating robustness of adversarial camouflage (AC) for naval vessels," Proc. SPIE 11543, Artificial Intelligence and Machine Learning in Defense Applications II, 115430G (20 September 2020); doi: 10.1117/12.2573676

SPIE.

Event: SPIE Security + Defence, 2020, Online Only

Investigating robustness of adversarial camouflage (AC) for naval vessels

Kristin Hammarstrøm Løkken[a], Lars Aurdal[b], Alvin Brattli[a], and Hans Christian Palm[a]

[a]Norwegian Defence Research Establishment (FFI), P.O. Box 25, 2027 Kjeller, Norway
[b]Independent researcher

ABSTRACT

The use of camouflage is widespread in the biological domain, and has also been used extensively by armed forces around the world in order to make visual detection and classification of objects of military interest more difficult. The recent advent of ever more autonomous military agents raises the questions of whether camouflage can have a similar effect on autonomous agents as it has on human agents, and if so, what kind of camouflage will be effective against such adversaries.

In previous works, we have shown that image classifiers based on deep neural networks can be confused by patterns generated by generative adversarial networks (GANs). Specifically, we trained a classifier to distinguish between two ship types, military and civilian. We then used a GAN to generate patterns that, when overlaid on parts of military vessels (frigates), made the classifier confuse the modified frigates with civilian vessels. We termed such patterns "adversarial camouflage" (AC) since these patterns effectively camouflage the frigates with respect to the classifier.

The type of adversarial attack described in our previous work is a so-called white box attack. This term describes adversarial attacks that are devised given full knowledge of the classifier under attack. This is as opposed to black box attacks, which describe attacks on unknown classifiers. In our context, the ultimate goal is to design a GAN that is capable of black box attacks, in other words: a GAN that will generate AC that has effect across a wide range of neural network classifiers.

In the current work, we study techniques to improve the robustness of our GAN-based approach by investigating whether a GAN can be trained to fool a selection of several neural network-based classifiers, or reduce the confidence of the classifications to a degree which makes them unreliable. Our results indicate that it is indeed possible to weaken a wider range of neural network classifiers by training the generator on several classifiers.

Keywords: naval vessel, camouflage, artificial intelligence, neural network, robustness

1. INTRODUCTION

Different types of imaging systems are frequently employed for detection, tracking or classification of naval vessels. Such systems may include one or more imaging sensors combined with one or more image processing platforms running the required algorithms. A number of countermeasure techniques are currently employed against such imaging systems. Depending on the observation spectrum employed by the sensor system they can be:

- Signature reduction aimed at simply reducing the signal from the vessel.
- Camouflage in the form of shape or colour changing approaches that aim at modifying the appearance of the vessel.
- Flares/decoys that generate smoke screens or generate artificial targets.
- Active countermeasures such as laser illumination aimed at blinding or confusing the sensors.

Further author information:
Kristin Hammarstrøm Løkken: E-mail: Kristin-Hammarstrom.Lokken@ffi.no

Artificial Intelligence and Machine Learning in Defense Applications II, edited by Judith Dijk, Proc. of SPIE Vol. 11543, 115430G · © 2020 SPIE
CCC code: 0277-786X/20/$21 · doi: 10.1117/12.2573676

With the advent of ever more sensitive and sophisticated imaging sensors combined with steadily improving processing platforms running more and more advanced algorithms, it becomes increasingly difficult to avoid detection, tracking and classification. In recent years, progress in neural networks and machine learning, often described as deep learning, has led to a performance leap for image analysis algorithms, empowering image analysis algorithms that exceed even human performance. Knowing that the vessels' visual and infrared signatures cannot be reduced to zero, and assuming that flares and decoys cannot mask a vessel for more than a short period of time, the question becomes what to do in order to render detection, tracking or classification as hard as possible.

One approach that we will explore further in this work is that of using adversarial camouflage (AC). Under this paradigm, the appearance of the vessel we wish to protect is structured in such a way that it confuses the software analyzing the images of the vessel.

Recent works in the domain of deep learning have shown that deep learning based algorithms for image analysis can be sensitive to surprisingly small changes in the images they analyze. Such techniques, typically described as adversarial techniques, have shown a considerable potential for fooling neural networks in a number of recent works.

In a previous work,[1] we have shown that a careful structuring of the visual appearance in grayscale images of a naval vessel can confuse deep learning based vessel classification algorithms to a very substantial degree. In particular, we showed how even relatively small patches of very specific visual patterns, displayed on parts of a naval vessel, will render classification of that vessel much harder for a deep learning based vessel classifier.

Our further investigations[2] showed that the patches were more robust against added noise or clutter than to degradations of contrast or resolution.

In this paper we will study the effectiveness of adversarial camouflage against a wider range of deep neural network discriminators. In this article we use the terms "discriminator" and "classifier" interchangeably.

In Section 2 we will give a short and very brief historical introduction to military vessel camouflage. We also provide an overview of the existing body of work related to techniques aimed at confusing neural networks, so called *adversarial* techniques. In Section 3 we describe the architectures of our neural networks, and go into detail about how they are trained and tested. In Section 4 we present and discuss the results we have obtained and in Section 5 we conclude.

2. RELATED WORKS

2.1 Naval vessel camouflage

In times of warfare, misleading the enemy, also called military deception, is of the essence. "Deception" can be defined as the act of causing someone to accept as true or valid what is false or invalid.[3] Camouflage is regarded as a means to this end. At sea, camouflage can be divided into two categories:[4–6]

- Concealment or signature reduction – measures taken to blend in with the background.
- Disruptive type – artifices designed to deceive enemy sensor systems, rendering identification or targeting more difficult by making the size, range, speed, heading or class difficult to determine.

From a military perspective, different types of camouflage are cost-effective means of increasing survivability and combat persistence and have been used by armed forces all over the world throughout history. The interested reader is referred to excellent literature on camouflage inspired from a zoological perspective[7] and military camouflage through history.[4,5,8]

Experiments with variants of dazzle camouflage[4,5,8] were carried out during the two World Wars. An example of dazzle camouflage is shown in figure 1. An interesting aspect of dazzle camouflage was that it did *not* necessarily aim at making the vessel harder to detect, but rather to make it harder for an observer to obtain good bearing, speed and distance estimates. As such it bears a certain resemblance to the type of camouflage we develop in the work reported here.

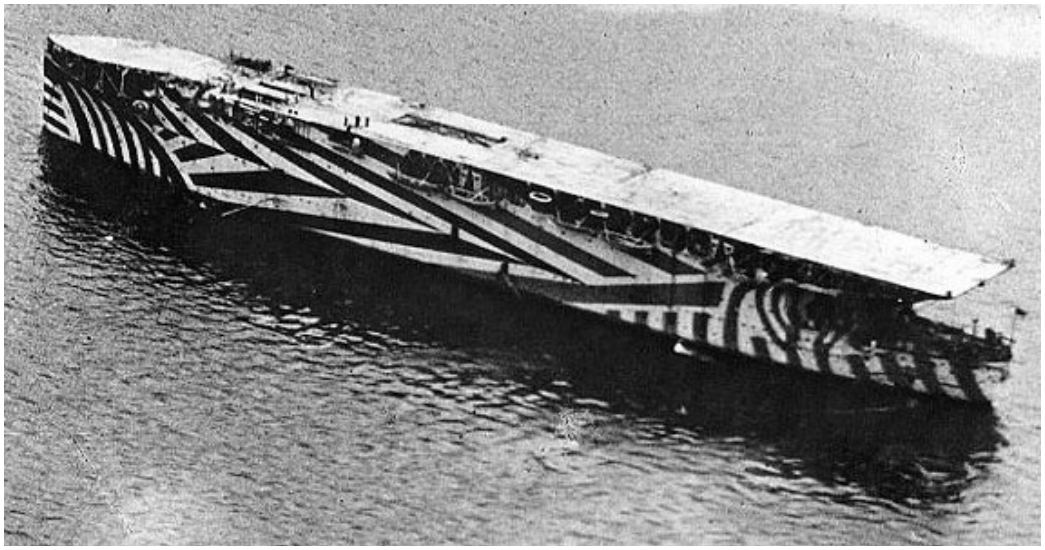


Figure 1: An example of dazzle camouflage. HMS Argus painted with dazzle camouflage in 1918. Photograph from wikipedia.org.[9]

2.2 Neural networks

Recent developments have shown the value of different types of neural networks for a number of complex applications in image processing, see for instance Goodfellow (2016).[10] In particular, variants of Convolutional Neural Networks (CNNs) have revolutionized the performance of automatic systems for object classification based on images. Today, performance of such systems is often on par with or even better than human performance (ibid). An interesting recent development in research related to neural networks is the introduction of adversarial attacks aimed to thwart the performance of a neural network.[11–17]

Recent work shows that small – undetectable to a human observer – changes in an image can be enough to fool a neural network discriminator, causing it to misclassify an image in a very dramatic fashion.[11–13,17,18] A recent study also produces images that are unrecognizable to human beings, but are classified as specific objects by state-of-the-art neural networks, with very high confidence.[19] This weakness in neural networks is shown to be due to the linear features of the neural networks.[13]

A further development of adversarial attacks is the concept of adversarial patches[15] and physical world attacks[14,20] on neural networks. It is shown that small patches with very specific patterns can make an image processing neural network misdetect or misclassify an object in the image. Such patches may fool the neural network even if they cover a relatively small part of the image, and regardless of whether or not they cover any of the features of the original object. Recent work has shown promising results using real world attacks making modifications that look natural to the human eye, but which confuses a neural network classifier.[21]

In our previous works[1,2] we used generative adversarial networks to create certain patterns for the adversarial camouflage. Our approach consisted in developing specific patches that can be applied to naval vessels in order to make classification networks misclassify them. The question was whether it was possible to alter a military vessel (with paint for instance), in a way that would fool a neural network discriminator into misclassifying it as civilian. We demonstrated that it was possible, and we have also investigated briefly the robustness of the patches. We have seen that the patches could be degraded somewhat and still fool the discriminator.

3. METHOD

The goal of our research is to investigate the robustness of adversarial camouflage. We simulate real world attacks by generating patterns that may be painted onto military vessels. By "attack" we mean the attempt to fool the discriminator neural network by using adversarial camouflage. Our neural network image generators are trained

to generate patterns that make image classifiers misclassify frigates as civilian vessels. We use grayscale images in the visual spectral range in our trainings and tests.

We use two types of deep neural networks:

- Discriminators
- Generators

The discriminators are based on standard discriminator networks, ResNet[22] discriminators with depths 18, 34, 50 and 101, and the generators are trained using generative adversarial networks (GANs) against one or more of the discriminators.

3.1 Generator

In order to train the generators, we have selected a total of 500 images of frigates (vessels clearly belonging to the class military). In each of these images we have manually designated an area within which the generator may modify the images. Images of six frigates and the designated masks are shown in figure 2. The masks are slightly blurred (prior to use the borders of the masks are blurred by a Gaussian filter with parameter $\sigma = 3.0$) and the masks have their opacity set to 90%. Both these steps are taken in order to simulate a situation where the adversarial pattern is actually painted onto the vessel.

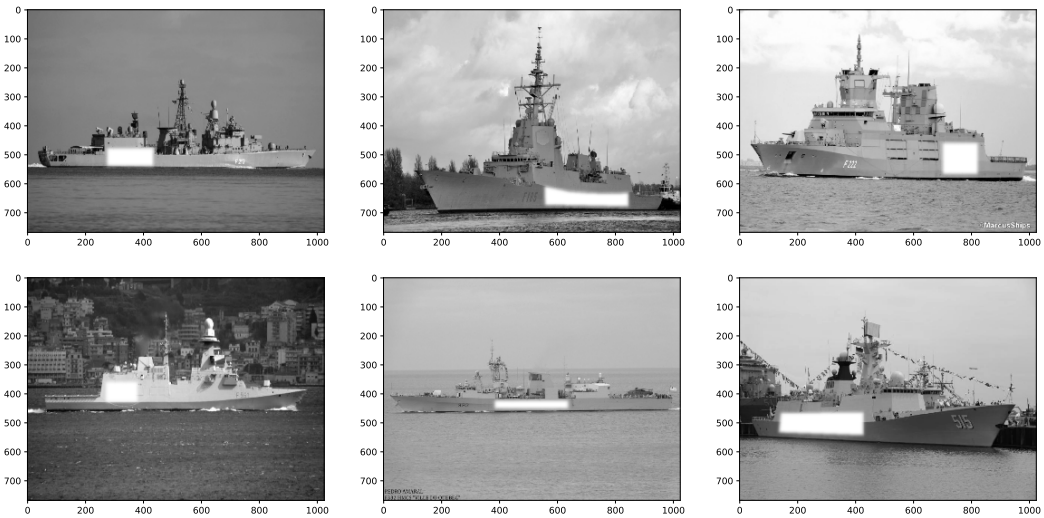


Figure 2: Images of six frigates with designated mask areas. All original photographs from shipspotting.com. Photographers (in reading order): *D173457Q Brian*, *Marcel and Ruud Coster*, *Marcus-S*, *Tomasello Letterio*, *Pedro Amaral* and *Ulf Kornfeld*.

We have trained the generators using GANs. The aim of the generators that are trained on single neural network discriminators is to produce a pattern that, when mixed into the frigate image within the area designated by the mask, will fool that discriminator into misclassifying the image of the frigate as a civilian vessel. The generator that is trained on multiple discriminators is designed to make the opposing discriminators less confident.

3.2 Discriminator

We have trained several convolutional deep neural network discriminators based on standard architectures, ResNet with depth 18, 34, 50 and 101. The frigate-cruiseship (frig-cru) discriminators are trained on frigate and cruiseship images only. The military-civilian (mil-civ) discriminators are trained on a much larger set of vessel images, but with only two possible classes: Military or civilian. Both types of discriminators are trained on grayscale images.

With these two dataset modes, we have trained a total of seven image discriminator networks:

- ResNet18 (frig-cru and mil-civ)
- ResNet34 (frig-cru and mil-civ)
- ResNet50 (frig-cru and mil-civ)
- ResNet101 (frig-cru only)

Our neural networks are implemented in PyTorch,[23] using some fastai[24] infrastructure.

The mil-civ discriminator networks were trained on a total of 93.000 images; 68.000 civilian and 25.000 military. The frig-cru discriminators were trained on a total of 3.600 images; 1.800 cruiseships and 1.800 frigates. The discriminators were trained for 25-27 epochs. Upon termination of training, the discriminators were tested on a testset consisting of 91 images of cruiseships and 100 images of frigates. The test was passed if the discriminator made 3 errors or less on the testset. All of our discriminators passed this test after 25 epochs, except for the ResNet101, which was trained for 27 epochs before passing the test.

3.3 Data

In order to train the networks involved in our experiments we need large numbers of images of maritime vessels, both civilian and military. One source for such images is the excellent web site `shipspotting.com`. This is a web site for shipspotters all over the world. As a user one can upload ship images along with the location of the shot, the name and type of the vessel, etc. There are millions of images available and a large fraction of the images show military vessels. The range of vessels is enormous.

We have downloaded a large number of images from shipspotting.com in the form of RGB images. These show vessels in all aspect angles as well as in all kinds of operating scenarios: in ports, in harbours, close to land, at sea, etc.

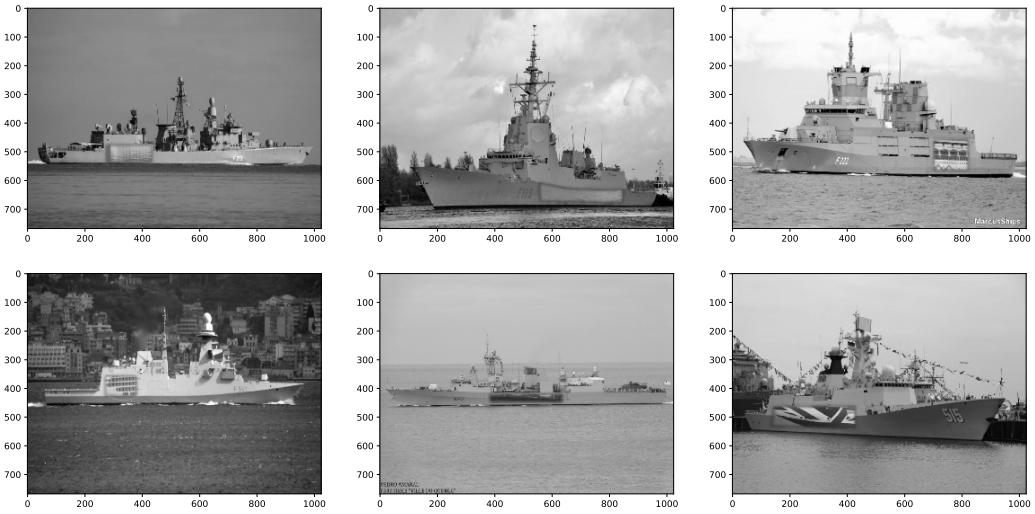
Our data was downloaded in September 2017. As mentioned above, the training set for the mil-civ discriminator contains 68.000 civilian and 25.000 military ship images. The number of classes in the civilian dataset is almost identical to the dataset that we downloaded. However, the military dataset has been reduced so as to contain mostly ships that are currently in use, and to exclude ship classes that may also be used in civilian shipping. The ship classes used in our military dataset are:

- Battleships
- Corvettes
- Cruisers
- Fast Attack Craft
- Frigates
- Landing Ships

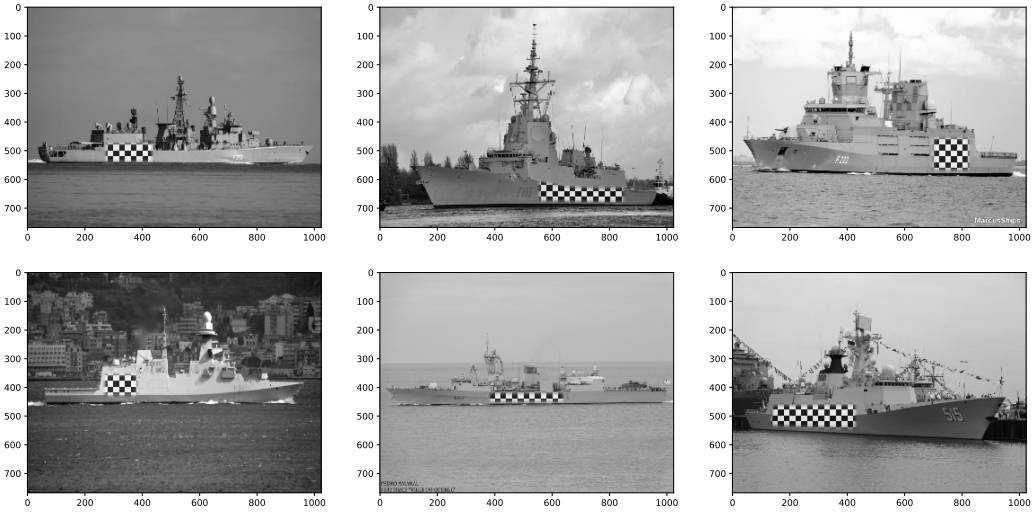
3.4 Training and testing the generators

When the discriminators are properly trained, our work on manufacturing the generators begins. We set up GANs with various combinations of discriminators. We train two generators against single discriminators and one generator against multiple discriminators. The discriminators are frozen (not trained) during the training of the generators.

The generators are tested by taking the output from the generator – a frigate image with the generated pattern included in the designated mask area – and feeding it to a discriminator. Before conducting these tests, we make sure that the discriminators are up to a certain standard. First, the discriminators are presented with the original frigate images. The discriminators are expected to clearly classify the test images as military when they are presented without the patch. Second, we have inserted parts of cruiseship images in the patch area to see whether this would make discriminators any less reliable. Third, the same test was performed with a checkerboard pattern, see figure 3. All of our discriminators have passed these tests before being exposed to the adversarial patches, confidently classifying the frigates as military vessels.



(a) Cruiseship patches.



(b) Checkerboard patches

Figure 3: Test images with part of cruiseship images (top) and checkerboard pattern (bottom) inserted in the masked area. All original photographs from shipspotting.com. Photographers (in reading order): *D173457Q Brian*, *Marcel and Ruud Coster*, *Marcus-S*, *Tomasello Letterio*, *Pedro Amaral* and *Ulf Kornfeld*.

3.4.1 Generators trained on single discriminators

We have trained two generators using single discriminators:

- Mil-civ ResNet18 – hereafter called mil-civ generator
- Frig-cru ResNet18 – hereafter called frig-cru generator

These generators are tested against both frig-cru and mil-civ ResNet discriminators with depths 18, 34 and 50. Thus both series of tests included one white box attack and five black box/gray box attacks. In this article we use the term "gray box" when the generator has access to the architecture of the discriminator *or* the dataset with which it was trained, but not both.

3.4.2 Generator trained on four different ResNets (4ResNet)

We have trained a generator using ResNet18, ResNet34, ResNet50 and ResNet101 frig-cru discriminators, referred to as the 4ResNet frig-cru discriminator.

In our previous tests, we have trained generators on single discriminators, aiming at causing this one discriminator to fail by classifying frigates as civilian vessels. The idea behind the 4ResNet test is that the user of a mil-civ discriminator would want it to be very confident about its classifications before making further decisions (e.g. letting the vessel pass or raising an alarm). If by simple means one could make such discriminators less confident or more confused, this is an interesting result for prospective users of such discriminators.

Thus, it is interesting to put the confidence of a wider range of discriminators to the test. Our goal is to reduce the confidence of a wide range of discriminators rather than completely fooling one single discriminator.

To accomplish this, the loss function of the generator is tuned so that its objective is no longer to completely fool all of the four discriminators, but rather make them less confident when classifying the frigate images as military. We set the goal for the generator to make the discriminators less than 25% confident that the frigates in the training images are military.

The resulting generated patches were then tested against the four ResNets individually.

4. RESULTS

The three generators produce three different patterns that may be added to the mask area in the test images. These are presented in figure 4.

(a) Camouflage pattern generated by the ResNet18 mil-civ generator.

(b) Camouflage pattern generated by the ResNet18 frig-cru generator.

(c) Camouflage pattern generated by the 4ResNet frig-cru generator.

Figure 4: Adversarial camouflage produced by the three generators. Original photograph from shipspotting.com. Photographer: *Marcel and Ruud Coster*.

4.1 Generators trained on single discriminators

The generators trained on single discriminators (a and b in figure 4) both generate striped patterns. Figure 5 shows the results from the ResNet18 mil-civ generator tested against six discriminators. Each test was performed with six test images, all displaying frigates, as shown in the figure. The return value from a discriminator is composed of two numbers: The confidence that the ship is civilian, and the confidence that it is military, respectively. These two numbers add up to one, and are shown as blue (civilian) and red (military) bars in the figure, labeled as CIV and MIL, respectively.

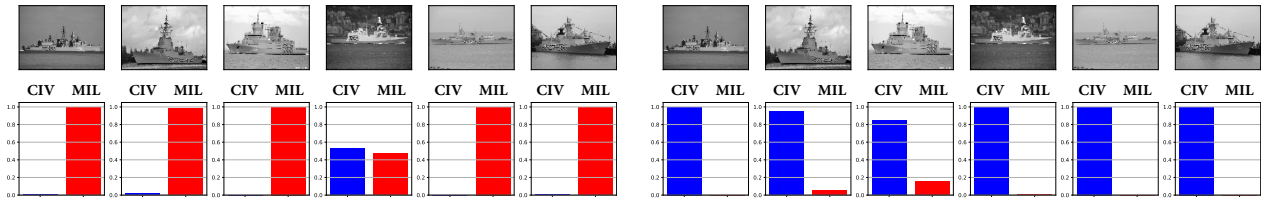
The results on the left are from tests against frig-cru discriminators (black box), while the results on the right are from tests against mil-civ discriminators. The top right result is from the ResNet18 mil-civ discriminator which was used to train the generator. This is the white box attack in this test. The other results on the right are trained on the same dataset, but with different discriminator architectures (gray box).

The white box attack (top right) clearly shows that the discriminator on which the generator was trained is easily fooled.

One would expect the gray box attacks on the right to render the discriminators less confident (show more blue bars) than the black box attacks on the left. The differences between the ResNet34 discriminators are insignificant. The confidence of the ResNet50 gray box attack is indeed reduced in one of five test images. As shown below (figure 6), this discriminator is not fooled at all by the frig-cru black box attack, demonstrating that the gray box attack impedes the confidence of the discriminator. However the ResNet50 frig-cru discriminator seems to have been fooled to a larger degree than the ResNet50 mil-civ discriminator. This is unexpected, and more tests will be needed to look further into these results.

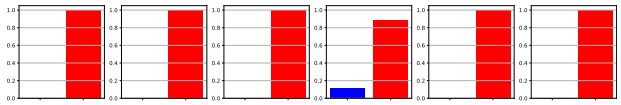
The results from the ResNet18 frig-cru generator tested against the same six discriminators are shown in figure 6. As expected, the white box attack (top left) indicates that the discriminator on which the generator was trained will be fooled by the generated adversarial patch; five of the six test images fool the discriminator. The gray box attacks (frig-cru dataset) yield a somewhat less reliable result than the black box attacks. The ResNet34 frig-cru discriminator is fooled by one of six test images. The ResNet50 frig-cru discriminator also seems to have been made less reliable. However this is the same discriminator as was used in the mil-civ test above, where the results were unexpected. More tests are needed to examine the reason for this behaviour.

The black box attacks are visible on the right. As expected, all the bar diagrams are red, indicating that the mil-civ discriminators were not the least fooled or weakened by the frig-cru generator.

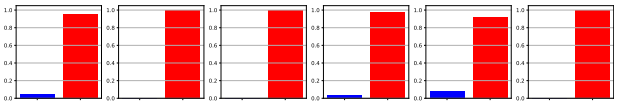


(a) ResNet18 discriminator trained on a frig-cru dataset.

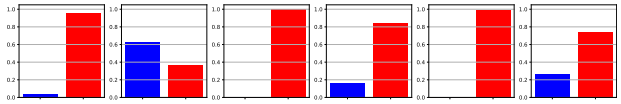
(b) ResNet18 discriminator trained on a mil-civ dataset.



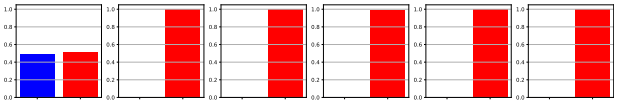
(c) ResNet34 discriminator trained on a frig-cru dataset.



(d) ResNet34 discriminator trained on a mil-civ dataset.



(e) ResNet50 discriminator trained on a frig-cru dataset.



(f) ResNet50 discriminator trained on a mil-civ dataset.

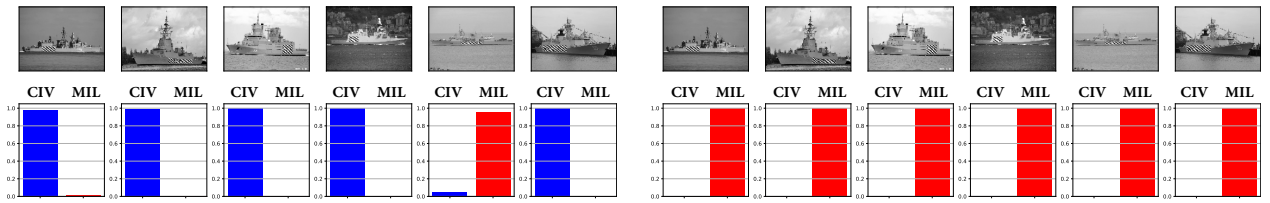
Figure 5: Generator trained on a Resnet18 mil-civ discriminator. Showing results from a white box attack (b) and five black/gray box attacks. All original photographs from shipspotting.com. Photographers (in reading order): *D173457Q Brian*, *Marcel and Ruud Coster*, *Marcus-S*, *Tomasello Letterio*, *Pedro Amaral* and *Ulf Kornfeld*.

4.2 Generator trained on several discriminators

The training of the 4ResNet generator was different from the training of the single discriminator generators in that the goal of the generator was *not* to completely fool all the discriminators. Rather the goal was to make all the discriminators less confident about their classifications. As such the discriminators could still classify the frigate images as military, but they would be less confident.

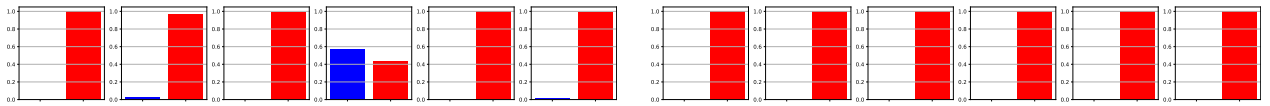
The adversarial pattern made by the 4ResNet generator is shown in the right (c) in figure 4. Figure 7 shows the six test images with this pattern inserted in the respective mask areas.

The generator was tested against the four discriminators on which it was trained, and the results are shown in figure 8. The results show that all of the discriminators classify one or more of the frigates as military, and as such are not completely fooled by the generator. However, all of the discriminators also classify two or more of



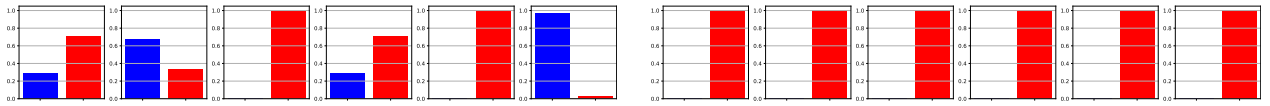
(a) ResNet18 discriminator trained on a frig-cru dataset.

(b) ResNet18 discriminator trained on a mil-civ dataset.



(c) ResNet34 discriminator trained on a frig-cru dataset.

(d) ResNet34 discriminator trained on a mil-civ dataset.



(e) ResNet50 discriminator trained on a frig-cru dataset.

(f) ResNet50 discriminator trained on a mil-civ dataset.

Figure 6: Generator trained on a Resnet18 frig-cru discriminator. Showing results from a white box attack (a) and five black/gray box attacks. All original photographs from shipspotting.com. Photographers (in reading order): *D173457Q Brian*, *Marcel and Ruud Coster*, *Marcus-S*, *Tomasello Letterio*, *Pedro Amaral* and *Ulf Kornfeld*.

the frigates as civilian, and several of the military classifications are uncertain. This indicates that it is possible to train a generator on several discriminator architectures, rendering all of them prone to error against the generator.

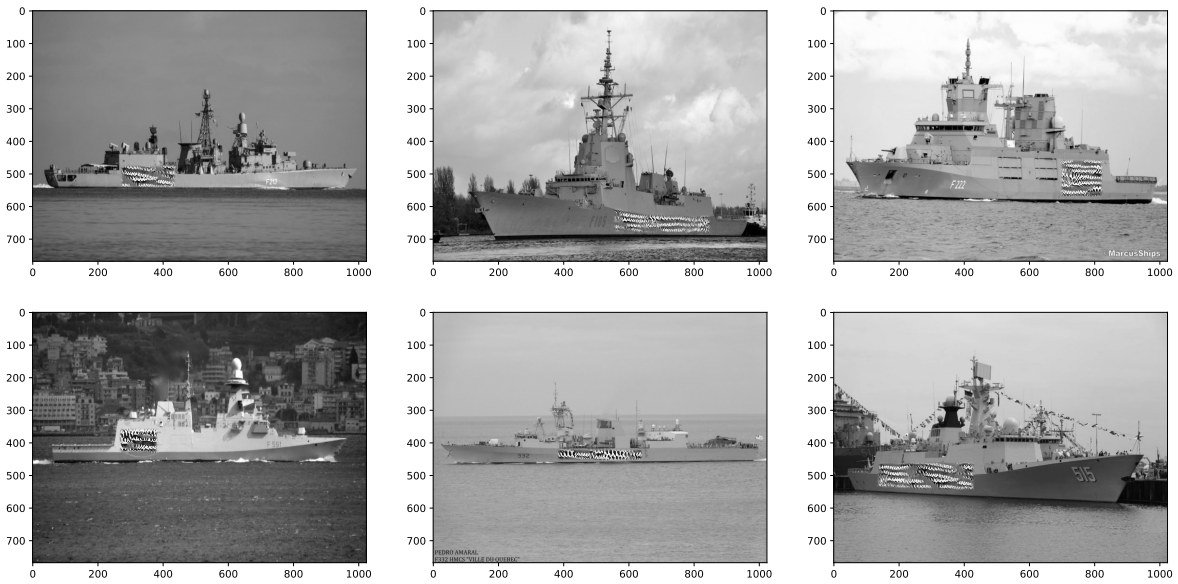
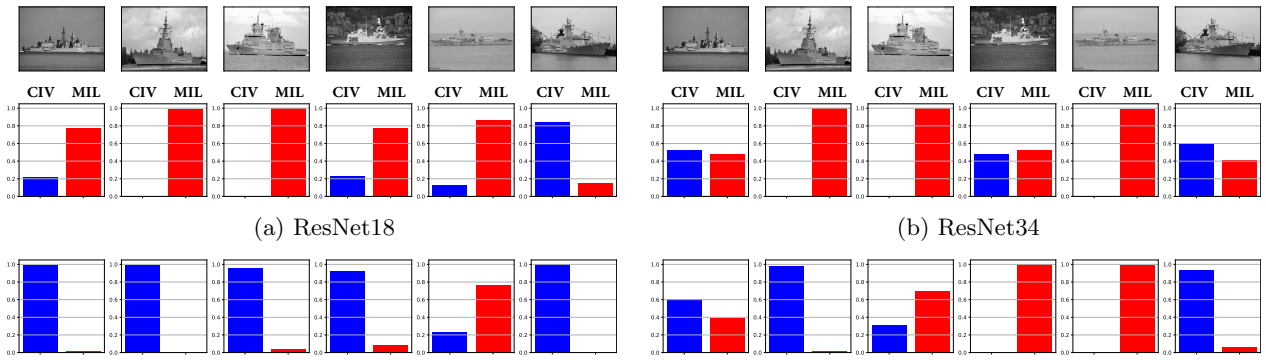


Figure 7: Military vessels modified by inserting adversarial camouflage pattern in the modification mask. All original photographs from shipspotting.com. Photographers (in reading order): *D173457Q Brian*, *Marcel and Ruud Coster*, *Marcus-S*, *Tomasello Letterio*, *Pedro Amaral* and *Ulf Kornfeld*.



(a) ResNet18 (b) ResNet34

(c) ResNet50 (d) ResNet101

Figure 8: Results for the 4ResNet generator tested against the four ResNet discriminators. All original photographs from shipspotting.com. Photographers (in reading order): *D173457Q Brian*, *Marcel and Ruud Coster*, *Marcus-S*, *Tomasello Letterio*, *Pedro Amaral* and *Ulf Kornfeld*.

5. CONCLUSIONS

In the work reported here we have investigated the use of generative neural networks in order to produce adversarial patches that will make discriminative neural networks trained to distinguish between civilian and military vessels less confident about their predictions.

In our previous work our goal has been to make the discriminators misclassify the test images, i.e. classify frigates as civilian vessels, mainly using white box attacks. In this article we have conducted black and gray box attacks, and our goal has been to render the discriminators less confident about their predictions.

Our ResNet18 mil-civ generator made the ResNet18 frig-cru discriminator fail on one in six test images. This indicates that some insight into the discriminator architecture may be useful, even though the dataset is different. It is worth mentioning that this result was absent from the opposite test, where the ResNet18 frig-cru generator was tested on the ResNet18 mil-civ discriminator. This is thought to be due to the datasets. In the former test the discriminator is trained on a large dataset which includes the dataset that the generator-training discriminator was trained on. In the latter test this was not the case. Rather the generator was trained on a frig-cru discriminator, and had not been trained on the other classes from the mil-civ dataset that the opposing discriminator was trained on. These results are somewhat inconclusive, and there is a need for more research to assert any further trends.

We have shown that a generator may be trained against several discriminators, making these discriminators less confident about their classifications. Further work on this matter should include training generators on even broader sets of discriminators and performing both white, gray and black box attacks.

REFERENCES

[1] Aurdal, L., Løkken, K. H., Klausen, R. A., Brattli, A., and Palm, H. C., "Adversarial camouflage (ac) for naval vessels," in [*Proceedings of SPIE - The International Society for Optical Engineering*], **11169**, 111690H (09 2019).

[2] Løkken, K. H., Brattli, A., Palm, H. C., Aurdal, L., and Klausen, R. A., "Robustness of adversarial camouflage (ac) for naval vessels," in [*Proceedings of SPIE - The International Society for Optical Engineering*], **11394**, 113940W (04 2020).

[3] Merriam Webster Dictionary, "Deception." https://www.merriam-webster.com/dictionary/deception (2019). [Online: accessed August 8th 2019].

[4] Raven, A., "The Development of Naval Camouflage." http://www.shipcamouflage.com/development_of_naval_camouflage.htm (2019). [Online: accessed August 13th 2019].

[5] USNI News, "Camouflaged Ships: An Illustrated History." https://news.usni.org/2013/03/01/camouflaged-ships-an-illustrated-history (2013). [Online: accessed August 13th 2019].

[6] Merriam Webster Dictionary, "Camouflage." https://www.merriam-webster.com/dictionary/camouflage (2019). [Online: accessed August 8th 2019].

[7] Cuthill, I. C., "Camouflage," *Journal of Zoology, Thomas Henry Huxley Review* **308**, 75–92 (05 2019).

[8] López, M., "Bismarck line drawings and paint schemes." https://www.kbismarck.com/drawings.html (2019). [Online: accessed August 13th 2019].

[9] Wikipedia, "Dazzle camouflage." https://en.wikipedia.org/wiki/Dazzle_camouflage (2019). [Online: accessed June 27th 2019].

[10] Goodfellow, I., Begio, Y., and Courville, A., [*Deep Learning*], The MIT Press, Cambridge, Massachusetts, USA (2016).

[11] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R., "Intriguing properties of neural networks," *arXiv e-prints* , arXiv:1312.6199 (Dec 2013).

[12] Carlini, N. and Wagner, D., "Towards Evaluating the Robustness of Neural Networks," *arXiv e-prints* , arXiv:1608.04644 (Aug 2016).

[13] Goodfellow, I. J., Shlens, J., and Szegedy, C., "Explaining and Harnessing Adversarial Examples," *arXiv e-prints* , arXiv:1412.6572 (Dec 2014).

[14] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D., "Robust Physical-World Attacks on Deep Learning Models," *arXiv e-prints* , arXiv:1707.08945 (Jul 2017).

[15] Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J., "Adversarial Patch," *arXiv e-prints* , arXiv:1712.09665 (Dec 2017).

[16] Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K., "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in [*Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*], *CCS '16*, 1528–1540, ACM, New York, NY, USA (2016).

[17] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Berkay Celik, Z., and Swami, A., "The Limitations of Deep Learning in Adversarial Settings," *arXiv e-prints* , arXiv:1511.07528 (Nov 2015).

[18] Gu, S. and Rigazio, L., "Towards Deep Neural Network Architectures Robust to Adversarial Examples," *arXiv e-prints* , arXiv:1412.5068 (Dec 2014).

[19] Nguyen, A., Yosinski, J., and Clune, J., "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in [*Computer Vision and Pattern recognition (CVPR '15)*], IEEE (2015).

[20] Thys, S., Van Ranst, W., and Goedemé, T., "Fooling automated surveillance cameras: adversarial patches to attack person detection," *arXiv e-prints* , arXiv:1904.08653 (Apr 2019).

[21] Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A. K., and Yang, Y., "Adversarial camouflage: Hiding physical-world attacks with natural styles," in [*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 997–1005 (2020).

[22] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 770–778 (2016).

[23] https://pytorch.org/.

[24] https://www.fast.ai/.