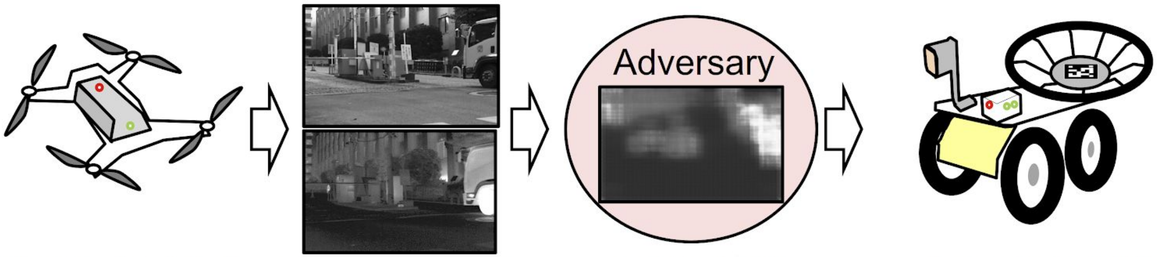
Adversarial Thermal Deep Fakes

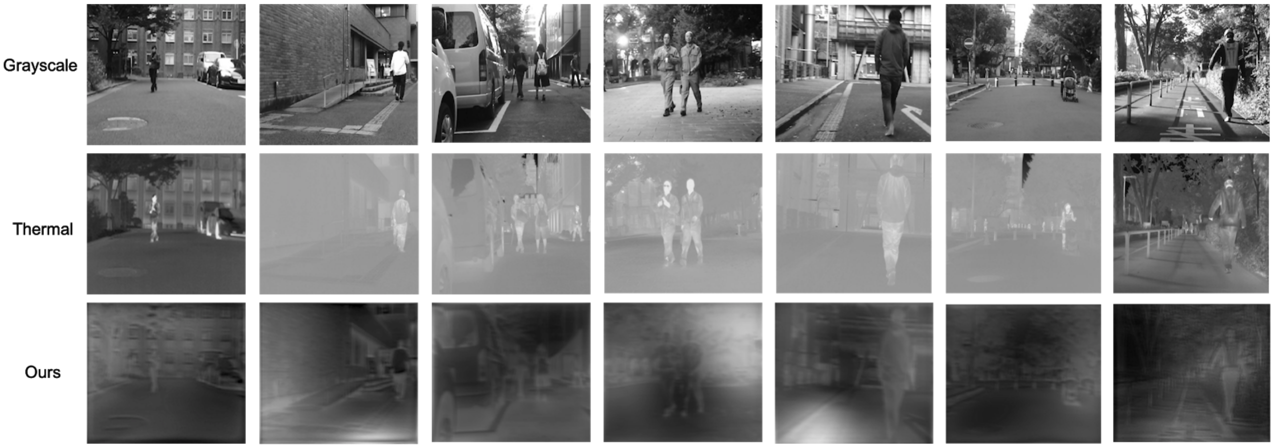
University of Florida

Student Team: Brevin Tilmon, Sumanth Aluri, Gavin St. John

***Faculty: Dr. Sanjeev J. Koppal Email:* sjkoppal@ece.ufl.edu**



1. Primary real sensors in the field
2. Real Grayscale and Thermal Images
3. ThermalGAN Grayscale **to Fake Thermal**
4. Secondary systems compromised

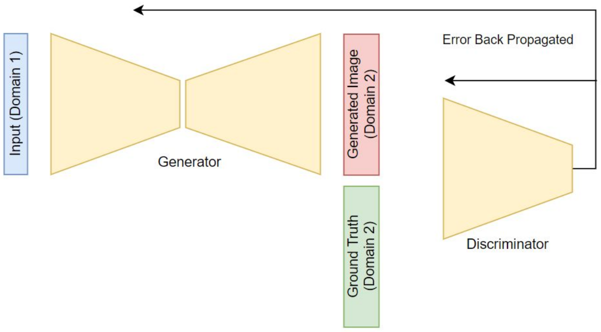


In the two figures above we outline a major vulnerability, where thermal imagery is transferred between robotic and semi-autonomous systems in the field. The top figure shows the attack model, where a *thermal deep fake* (i.e. simulating sensing in the **8 to 14 micron** range) can be created with one of an ensemble of Generative Adversarial Network (GAN) methods we discuss. In the second figure, **we show output of our software, with examples converting a grayscale image into a thermal-like deep fake.** Compare the deep fake in the third row to the real thermal image in the second. Humans or a neural network on secondary systems can be compromised by such data.

Introduction

The basic idea behind Generative Adversarial Networks (GAN) is to have two networks competing with each other. The first is the generator which creates realistic fake

Generative Adversarial Network (GAN)



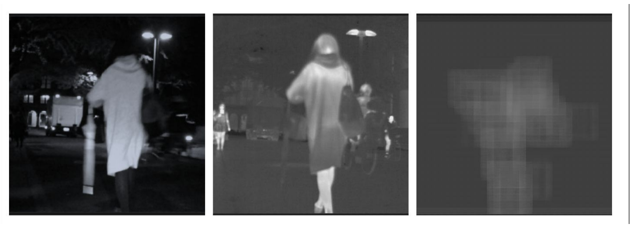
images, and the second is the discriminator which distinguishes between the fake and real images. In our implementation the generator is itself an autoencoder that attempts to convert images from the grayscale domain to thermal. Domain conversion is possible using a simple autoencoder trained by itself , however the discriminator helps to further fine tune the model. This allows the final model perform even better than if it were trained with traditional supervised learning using only ground truth labels.

Vulnerability Model: There is no doubt that thermal sensors are crucial in any battle field. Such devices are especially important for the field of robotics. Thermal sensors are often used in low lighting environments and for detection of humans, both of which can be especially sensitive scenarios. In such conditions it is vital to make sure that the mobile sensory input is not corrupted or attacked with fake signals. Research into detecting and preventing manipulation of thermal data is important to mitigate the risk of damages and mission failures.

Technical Challenges for Thermal Deep Fakes: A major problem we faced was the difficulty in extracting information from grayscale images that could be used to reconstruct thermal ones. To start with, color images and thermal images both measure disjoint sets of the electromagnetic spectrum, meaning that information from one domain does not directly translate into information in the other domain. Furthermore gray scale images are themselves a projection of color images into an even lower dimensional space, making the task of extracting thermal features even more challenging. This is further compounded by the fact that gray scale images perform even more poorly in low lighting conditions where shapes that could otherwise be distinguished meld together.

GAN-based Thermal Deep Fakes

State of the art (simple): To outline how hard the problem actually is, we compare with two alternative methods. In the adjoining figure here, weshow an unsophisticated pure autoencoder. This demonstrates how traditional machine learning algorithms trained using only ground truth have cost functions based directly on the difference between generated and read images. This approach is unsophisticated as it only prioritizes changing vague pixel values to better match the ground truth. By contrast, In a GAN model the generator is tasked with changing higher level features that directly impact how realistic a fake image can be.



State of the art (GAN limited domain): There have been methods to go directly from color/gray imagery to thermal data. In the adjoining figure we show the results from [1], which can be very convincing for the limited domain of cropped imagery which contain human figures from a narrow demographic set. The results, while impressive, do not address our concern for general thermal conversion in

scenes with trees, vehicles, buildings, as well as human figures with large variation in depth and perspectives.

Our Approach: Kniaz [1] modified the classic Unet architecture for their generator. We take a similar approach, except we notice that our result quality was proportional to the size of our generator model. Given this, we experimented with the larger Densenet architecture from Huang [5] due to the large model size and wide success in monocular depth estimation and semantic segmentation. Using the standard Wasserstein GAN from Arjovsky [3] with minimal adjustments, we are able to generate high resolution thermal images. We believe we mostly benefit from the representative power of Densenet architectures, which are not currently used in most thermal GAN applications. This follows work from Pittaluga [6] that uses large refining networks to enhance generator results for reconstructing images from sparse 3D point clouds.

We implement our GAN in the PyTorch framework. We use a custom five layer CNN with batch normalization and Leaky ReLU activations for our discriminator. Following the algorithm from Arjovsky [3], we clamp our discriminator weights between -0.01 and 0.01 and update the discriminator network fives times for every generator update. We use a learning rate of 1e-6, batch size of 8 and obtain best results training for 80 epochs. Training for 80 epochs takes 1.5 hours on a single NVIDIA A100 GPU. We also lower input resolution to 240x320 from 480x640 to lower the parameter search space. We convert our inputs to 3D grayscale to leverage the pretrained Densenet169 architecture that requires 3D inputs.

Defending against Thermal Deep Fakes

Here we have shown that thermal deep fakes from non-thermal (i.e. RGB or grayscale) data is a significant threat to autonomous systems, since processing on these images may result in catastrophic decisions, particularly in night-vision environments where reliance on long wave infrared imagery.

Defense against thermal deep fakes can take two forms: training against known GAN / other deep fake techniques and using information inherent to the environment being captured. For the former, a discriminator network can be trained to differentiate between real thermal images and fake thermal images generated with state of the art techniques. However, the influence of this technique is reduced when defending against novel techniques. Another discriminator should be trained which uses information inherent to the environment being captured. Encoded in thermal images is information about the temperature of the objects imaged. The difficulty in translating RGB or grayscale images to thermal is that the visible spectrum doesn’t accurately carry information about temperature. This means that often the deep fake networks rely on categorizing objects in the scene and then translating each to thermal based on how those objects usually look in thermal. Due to this the deep fake network loses the context of the scene and the temperature data of those objects in relation to each other. This means a network can be trained to recognize when the (fake) thermal image information differs from the context of the background information.

Another possible technique relies on the prior of physics, encoded as large sets of imagery of the relevant scenes. Assuming this system is using images recorded over time then either of the proposed defenses can be adjusted to take image history into account. Not only relying on absolute differences but differences in rate of image change over time. Dependence on ambient temperature (via an onboard thermometer) can be used to further cull the space.

Conclusions

We have demonstrated that thermal deep fakes are a real possibility for scenes of vehicles and persons that could occur on a battlefield. They are designed to fool a neural network into *thinking* they are real, on a non-physical level (i.e. colormap, contrasts, distribution of intensities, etc.). We contend that the lack of overlap between color and thermal imagery mean that the data produced for real scenes, however visually compelling, can result in catastrophe if the measurements are taken to be thermally realistic. We have discussed what steps can be taken to defend against thermal fakes, which include both physical-priors and entering the arms race of GAN-based defense of GAN-based thermal fakes. We believe this area to be teeming with possibilities for both research and applications.

Bibliography

1. V. Kniaz, V. Knyaz, J. Hladuvka, W. G. Kropatsch, V. Mizginov. *"ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification Multispectral Dataset"*. European Conference on Computer Vision (ECCV) 2018.
2. Y. Sun, W. Zuo, M. Liu. *"RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes".* IEEE Robotics and Automation Letters 2019.
3. M. Arjovsky, S. Chintala, L. Bottou. *"Wasserstein Generative Adversarial Networks".* International Conference on Machine Learning (ICML) 2017.
4. I. Goodfellow, J. Pouget, M. Mirza, B. Xu, D. Warde, S. Ozair, A. Courville, Y. Bengio. *"Generative Adversarial Nets".* Advancements in Neural Information Processing Systems (NIPS) 2014.
5. G. Huang, Z. Liu, L. Maaten, K. Weinberger. *"Densely Connected Convolutional Networks".* IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017.
6. F. Pittaluga, S. J. Koppal, S. Kang, S. Sinha. *"Revealing Scenes by Inverting Structure from Motion Reconstructions".* IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019.
7. F. Pittaluga, S. J. Koppal, A. Chakrabarti. *"Learning Privacy Preserving Encodings through Adversarial Training"*. IEEE Winter Conference on Applications of Computer Vision (WACV) 2019.