International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

Generalizability of Object Detection Algorithms Trained on RGB Images Over Thermal Images

Aryamaan Sen
High School Student
The Shri Ram School, Aravali
Gurgaon, India

***Abstract*— Object detection, a common application of computer vision has tremendous usage in fields like surveillance, autonomous driving, Face-ID, anomaly detection, traffic management, agriculture and more. To overcome problems faced by regular RGB cameras such as weather disturbances and visibility in the dark, the use of thermal images has increased drastically. However, the lack of large scale publicly available annotated thermal image datasets due to the expensive and domain-specific nature of the cameras greatly limits the ability to train such algorithms. This paper tests how well an object detection model trained on RGB images generalizes to thermal images and experiments with different amounts of thermal training data to find improvement in performance in order to find an optimal balance between accuracy and requirement for thermal image training data. Unsurprisingly, results show that the accuracy of the model increased as the amount of thermal training data was increased, plateauing near the end.**

Keywords*— Object detection, thermal images, Generalizability

I. INTRODUCTION

Object detection is a task within computer vision that involves determining whether an image contains objects of certain specified classes as well as pointing out where in the image they are located. This has several applications in the real world such as automated surveillance cameras, self-driving cars, Face-ID and much more.

Over the past few years there has been unprecedented success in the field of object detection with newer, more efficient models being created frequently[1]. However, in most cases object detection is performed using RGB images or videos since they are more commonly used and have a wider usage. Although RGB images are able to cover most use cases very well, in some situations they may not be able to do so. For example, RGB cameras for surveillance or in self driving cars do not work well in low-light situations or bad weather. Moreover, with the COVID-19 pandemic, temperature detection has become essential and thermal imaging works well in these situations and has therefore been adopted by big companies[2].

However, at the time of writing this paper, there are very little publicly available datasets or research that has been done in the area, make in challenging for developers to create and train such models. Considering this information, this paper is going to find how well an object detection model trained using RGB images generalizes over thermal images. The paper will also experiment to find how it can be fine-tuned to accommodate for differences in interpretation of the 2 types of images by the model (difference in cross-modality).

RGB and thermal images have some distinct differences in their cross-modality[3] which makes this task challenging but very important. The difference is because an RGB image has 3 channels of information whereas a thermal image has only one, making it difficult for a model to read and analyze the image properly.

To solve this issue of differences in cross-modality and to increase generalizability, there exist many different methods. Most conventional approaches aim use feature alignment. Some of the newer techniques include data augmentation, meta-learning, and domain alignment. This paper uses data augmentation as it is the simplest and most efficient method for developers to implement. This paper uses a pretrained model and fine-tunes it by training it on different amounts of thermal data to see how much the model improves.

This paper will also look at augmentation of the training data to increase this generalizability. The results of the paper give a conclusion on how RGB image datasets can be used to train object detection models designed for thermal images. Although previous research has experimented with different techniques for generalization, this is the first paper to compare the difference created by using data augmentation in order to find the right balance between thermal data required and accuracy.

II. THE OBJECT DETECTION MODEL

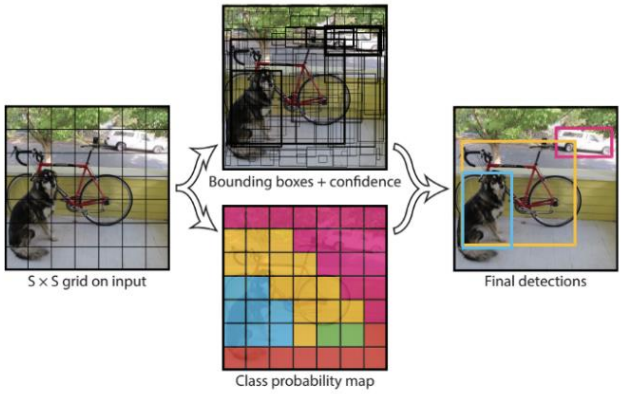
A simple and naïve way to conduct object detection would be to take a random region of interest within an image and use a convolutional neural network (CNN) to check for the presence of an object from the specified classes within the region. This process would continue until the entire image is covered and checked for objects. Initially, this method seems to be efficient and accurate; however, objects can be of different sizes, shapes and aspect rations, having different spatial locations within the image. This would require a very large number of regions of several sizes to be checked to make sure the entire scope for objects to be present is covered. This would be a computational nightmare for large images with a lot of objects. To solve this problem, there have been several algorithms that have been developed to conduct object detection efficiently. Some of these are as follows:

i) Region based Convolutional Neural Network (R-CNN)
ii) Fast R-CNN

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

iii) Faster R-CNN
iv) You-Only-Look-Once (YOLO)

In most methods, models localize images and predict object within those regions. The YOLO model on the other hand looks at the complete image and uses just one Convolutional Neural Network to predict the bounding boxes and the class probabilities. In this model, the image is split into an SxS grid with each grid box having 'm' bounding boxes. For each bounding box, YOLO outputs a class probability and offset values for the bounding box. The boxes having the class probability above a threshold value are selected and used to locate the objects in the image. A pictorial representation of its functioning is below. YOLO is currently more efficient that most other object detection algorithms[4].



YOLO

This paper uses a YOLOv2 model specifically trained for vehicle detection. This is because the dataset used contains primarily images of traffic and using a multi-class detector would add unnecessary computational complexity without providing much benefit in terms of the results. The YOLO model was chosen because it is faster compared to most of its alternatives and is also more efficient. A summary of the model used is as follows:

	Name	Type	Activa...	Learnables
1	input 128×128×3 images	Image Input	128×128×3	-
2	conv_1 16 3×3 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	128×128×16	Weights 3×3×3×16 Bias 1×1×16
3	BN1 Batch normalization	Batch Normalization	128×128×16	Offset 1×1×16 Scale 1×1×16
4	relu_1 ReLU	ReLU	128×128×16	-
5	maxpool1 2×2 max pooling with stride [2 2] and padding [0 0 0 0]	Max Pooling	64×64×16	-
6	conv_2 32 3×3 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	64×64×32	Weights 3×3×16×32 Bias 1×1×32
7	BN2 Batch normalization	Batch Normalization	64×64×32	Offset 1×1×32 Scale 1×1×32
8	relu_2 ReLU	ReLU	64×64×32	-
9	maxpool2 2×2 max pooling with stride [2 2] and padding [0 0 0 0]	Max Pooling	32×32×32	-
10	conv_3 64 3×3 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	32×32×64	Weights 3×3×32×64 Bias 1×1×64
11	BN3 Batch normalization	Batch Normalization	32×32×64	Offset 1×1×64 Scale 1×1×64
12	relu_3 ReLU	ReLU	32×32×64	-
13	maxpool3 2×2 max pooling with stride [2 2] and padding [0 0 0 0]	Max Pooling	16×16×64	-
14	conv_4 128 3×3 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	16×16×128	Weights 3×3×64×128 Bias 1×1×128
15	BN4 Batch normalization	Batch Normalization	16×16×128	Offset 1×1×128 Scale 1×1×128
16	relu_4 ReLU	ReLU	16×16×128	-
17	yolov2Conv1 128 3×3 convolutions with stride [1 1] and padding 'same'	Convolution	16×16×128	Weights 3×3×128×128 Bias 1×1×128
18	yolov2Batch1 Batch normalization	Batch Normalization	16×16×128	Offset 1×1×128 Scale 1×1×128
19	yolov2Relu1 ReLU	ReLU	16×16×128	-
20	yolov2Conv2 128 3×3 convolutions with stride [1 1] and padding 'same'	Convolution	16×16×128	Weights 3×3×128×128 Bias 1×1×128
21	yolov2Batch2 Batch normalization	Batch Normalization	16×16×128	Offset 1×1×128 Scale 1×1×128
22	yolov2Relu2 ReLU	ReLU	16×16×128	-
23	yolov2ClassConv 24 1×1 convolutions with stride [1 1] and padding [0 0 0 0]	Convolution	16×16×24	Weights 1×1×128×24 Bias 1×1×24
24	yolov2Transform YOLO v2 Transform Layer with 4 anchors.	YOLO v2 Transfor...	16×16×24	-
25	yolov2OutputLayer YOLO v2 Output with 4 anchors.	YOLO v2 Output	16×16×24	-

III. DATASET

As the introduction states, there are very little publicly available, annotated datasets with thermal images. For the purpose of this paper, a dataset with thermal and RGB images where the 2 were aligned and annotated was required. However, there is no such dataset that was large enough to suit the needs. Therefore, we used the FLIR ADAS dataset.

The FLIR ADAS dataset contains annotated thermal imagery and non-annotated RGB images where the images are synced (taken at the same time) but not aligned. There are over 14,000 images with over 10,000 short video segments and random image samples. This paper uses only the images from the dataset. The dataset has 80 classes of objects with the majority being images of persons, cars, trucks or bicycles.

This dataset provides a large repository of synced thermal and RGB images, however, the images were not paired and the RGB images did not have ground truth bounding boxes of their own so they could not be used to test the accuracy of the pre-trained YOLOv2 model on the RGB images. Therefore, the images were cropped to best fit the thermal images and the RGB images were then pre-processed to be of the sane aspect ratio. Due to slight misalignments in the paired images, the bounding box sizes were increased to 135% of their original bounds to factor for the misalignment.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

IV. METHODOLOGY

To test how well an object detection model trained on RGB images generalizes on thermal images, this paper tests the accuracy of a pre-trained object detection model on the images of the FLIR ADAS dataset. Average precision was calculated along with the precision-recall curves for each experiment.

This paper was initially working with a multi-class YOLOv3 detector trained on the MS COCO dataset; However, this was unnecessary for two main reasons. First, a multi-class classification is more computationally complex than a single class classifier and for the purpose of this research, various classes of objects are not necessarily required. Second of all, it was found that the YOLOv3 multi-class classifier was mostly detecting vehicles anyway since that was the primary class in the set of thermal images of the FLIR ADAS dataset. Therefore, a single class vehicle detection model was used to make the research in this paper more efficient. A pre-trained YOLOv2 vehicle detection model was used. However, testing other models and multi-class detectors could provide further insight into this data.

Once the model was decided, the dataset was processed. This FLIR ADAS dataset's thermal images are of size 640x512 which was too large to work with. Furthermore, the model could only take square images as its input. Therefore, all images were cropped and resized to 128x128 pixel images. To match this the bounding box sizes and positions were scaled down as well. The pre-processing mentioned in the 'Dataset' section of this paper was carried out on the thermal images too.

Now, with the model and data ready, the pre-trained weights of the model were tested on the test set of the dataset and acted as a baseline test. This test was used as a reference point for the rest of our experiments. Th model was then trained using thermal images from the training set of the FLIR ADAS dataset and the accuracy of the model was tested again. This was done in stages. First, 1000 thermal images were used to train the model, then 2000, then 3000 and so forth till the entire dataset of 8861 thermal images was exhausted. The model was trained using the formal training options:

1) *Learning rate- 0.001*
2) *Mini-batch size- 16*
3) *Epochs- 30*

In the future, other techniques such as domain alignment and meta-learning can be used to further enhance the findings from this paper. The purpose of this paper is to find the right balance between usage of thermal images in the training data and accuracy of the model achieved. This will benefit developers working on object detection using thermal images and possibly other computer vision tasks as it would give insights into how RGB data can be used and how the model can be molded to increase accuracy and reduce the requirement for thermal images which are scarcely available.

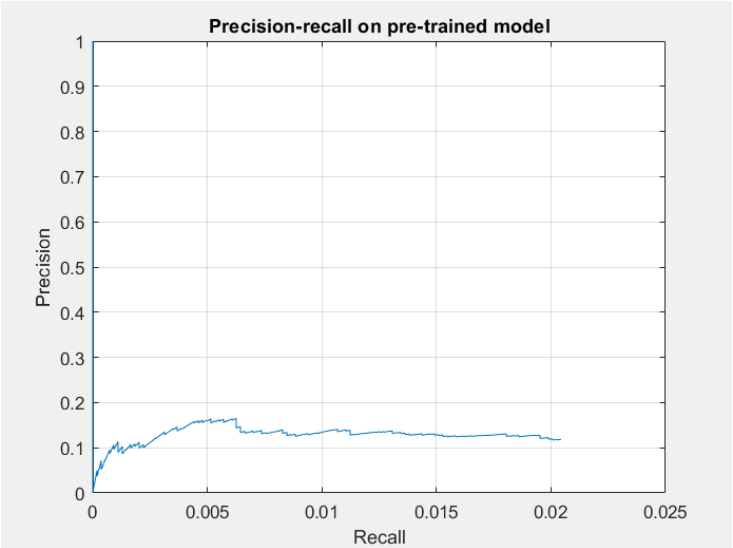
V. RESULTS

As stated earlier, this paper experiments with different amounts of thermal training data to help understand the benefit that increasing the amount of training data provides which can help developers understand how much data they required to reach the approximate accuracy levels they need.

For this research, the Intersection over Union metric was used to measure the accuracy of the model in the different experiments. The Intersection over Union metric (IoU) is a way of measuring the extent of overlap of two bounding boxes. The greater is the overlap between the ground truth and predicted bounding boxes, the greater is the IoU. Precision-recall curves for each experiment were also created for each experiment which allowed us to see how much the precision drops as the recall is increased. However, these curves are not shown in this paper as it is not essential to the purpose of the paper.

A. *Baseline Test*

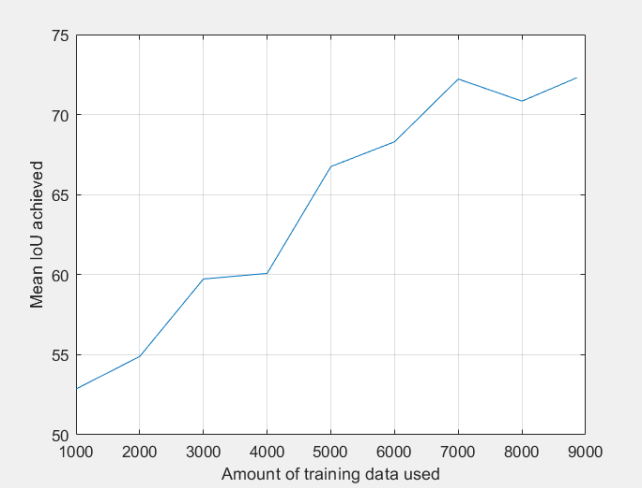
The accuracy of the pre-trained YOLOv2 vehicle detection model on the thermal images of the test set of the FLIR ADAS dataset was used as the baseline test. Unsurprisingly, a comparatively low mean IoU of 46.35% was achieved. The precision-recall graph for the test is plotted below.



B. *Experiments:*

In this paper, experiments were conducted on the amount of thermal data used in re-training the YOLOv2 model. First, the model was trained on 1000 thermal images, then 2000, then 3000 and so forth till all 8861 images were used to train the model in stages. The precision-recall curves for each experiment were plotted and the mean IoU for each experiment was calculated too. A graphical representation of the mean IoU across the experiments is given below

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022



From the experiments, it is clear that the model benefits greatly from the training on the thermal images. However, it also shows that the model need not be initially trained on thermal images and can be trained on RGB images. The plot of the Mean IoU vs amount of training data graph allows us to visualize the benefit of the additional training and the exact benefit provided by each increment.

As is visible in the graph, the growth initially is a lot, with the greatest increase in Mean IoU coming when the amount of thermal training images was increased from 4000 to 5000 images. Beyond this, the growth rate starts to reduce and somewhat plateaus. There is also an exception to the trend when the number of images is increased from 7000 to 8000. This was regarded as an exception due to computational error and not considered deeply as the general trend of the graph was still upward.

Therefore, for purposes that do not need near perfect detection accuracies, a good balance can be achieved at the point where the Mean IoU achieved starts plateauing. This is because adding large amounts of training data beyond this point adds minimal accuracy to the model and a good accuracy can be obtained without the need for tremendously large amounts of thermal training data.

VI. FUTURE RESEARCH DIRECTIONS

So far, we have seen how well an object detection model trained primarily on RGB images generalizes on thermal images. We have discussed how augmenting the training data to contain varying amounts of thermal images impacts the accuracy of the model. However, there are several other methods for domain generalization which was not within the scope of this paper. This includes testing different object detection models, newer versions of these models, techniques such as ensemble learning, self-supervised learning, domain alignment and domain adaptation[5].

1) Testing different models: Generalization of an object detection model could be vastly different for different models and different generaltions of models as well based on how they work. Conducting a comparative analysis of how different models generalize would be a promising direction to investigate.

2) Ensemble Learning: Ensemble learning typically learns multiple copies of the same model, initialized on different weights and/or using different splits of training data and then uses an ensemble of these copies to make its prediction. This method is very effective in enhancing the performance of a model and to increase generalizability of a model to cover a wider range of applications[6].

3) Self-supervised learning: Self-supervised learning is a technique which leverages the underlying structure in the data. The general process is to predict any unobserved or hidden properties of the input from any observed part of it. This process allows the model to learn more generic features, irrespective of the task at hand and prevents overfitting to domain specific features that exist. This could be another promising area to explore with the potential to produce positive results.

4) Domain alignment: This is a conventional technique of trying to minimize the difference between the 2 domains, in our case thermal and RGB images, which allows the model to work efficiently over both domains. This technique works by recognizing features invariant to the domains which will allow any target domain shift to be relatively less challenging for the model to overcome.

Investigating these areas could be very beneficial to getting positive results. Conducting research in these areas and building on the results of this paper will make the research completer and more structured with definitive results for how well object detection models trained on RGB images can generalize to thermal images and a tabulated comparison between different techniques applied on different models.

VII. CONCLUSION

Due to the rise in usage of thermal imaging for object detection purposes and the shortage of research and data, this paper explores the field of thermal object detection. The main goal of this paper was to understand how well a neural network trained for object detection using RGB images generalizes over thermal images. This paper aims to find how the model can be fine-tuned by training the model on different amounts of thermal images and how much benefit this provides to us. This would benefit developers looking to create efficient models without having to use tremendous amounts of thermal data.

As is evident from the experiments, the Mean Intersection over Union has a general increasing trend as the number of thermal images in the training data increases. Overall, promising results were achieved as a massive 27.41% improvement from the baseline tests was observed.

This shows that it is plausible to use detection models trained on RGB images for thermal object detection. This will be very beneficial to developers as it reduces the need for extensive thermal datasets to get accurate results.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

While this paper was able to get some insights to how a model generalizes over the 2 distinct domains of RGB and thermal images, further investigating other methods stated in the 'Future Research Directions' section will be extremely useful to further the findings of this paper. These results will help find the perfect balance between RGB and thermal data as well as adaptation techniques that are required to reach a target accuracy of a particular model. Therefore, it is critical to explore these areas to get results that could potentially have major impacts in the world of computer vision.

VIII. REFERENCES

[1] "R-CNN, Fast R-CNN, Faster R-CNN, YOLO — Object Detection Algorithms | by Rohith Gandhi | Towards Data Science." Accessed August 15, 2021. https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e.

[2] Hwang, Soonmin. *KAIST Multispectral Pedestrian Detection Benchmark*. https://github.com/SoonminHwang/rgbt-ped-detection.

[3] Wang, Guan'an, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. "RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment." *ArXiv:1910.05839 [Cs]*, October 28, 2019. http://arxiv.org/abs/1910.05839.

[4] "Object Detection Using YOLO v3 Deep Learning - MATLAB & Simulink - MathWorks India." Accessed August 15, 2021. https://in.mathworks.com/help/vision/ug/object-detection-using-yolo-v3-deep-learning.html.

[5] Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. "Domain Generalization in Vision: A Survey." *ArXiv:2103.02503 [Cs]*, July 18, 2021. http://arxiv.org/abs/2103.02503.

[6] Blanchard, Gilles, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. "Domain Generalization by Marginal Transfer Learning." *ArXiv:1711.07910 [Stat]*, January 6, 2021. http://arxiv.org/abs/1711.07910.

[7] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection." *ArXiv:2004.10934 [Cs, Eess]*, April 22, 2020. http://arxiv.org/abs/2004.10934.

[8] Torralba, Antonio, and Alexei A. Efros. "Unbiased Look at Dataset Bias." In *CVPR 2011*, 1521–28. Colorado Springs, CO, USA: IEEE, 2011. https://doi.org/10.1109/CVPR.2011.5995347.

[9] Tzelepi, Maria, and Anastasios Tefas. "Graph Embedded Convolutional Neural Networks in Human Crowd Detection for Drone Flight Safety." *IEEE Transactions on Emerging Topics in Computational Intelligence* PP (March 4, 2019): 1–14. https://doi.org/10.1109/TETCI.2019.2897815.