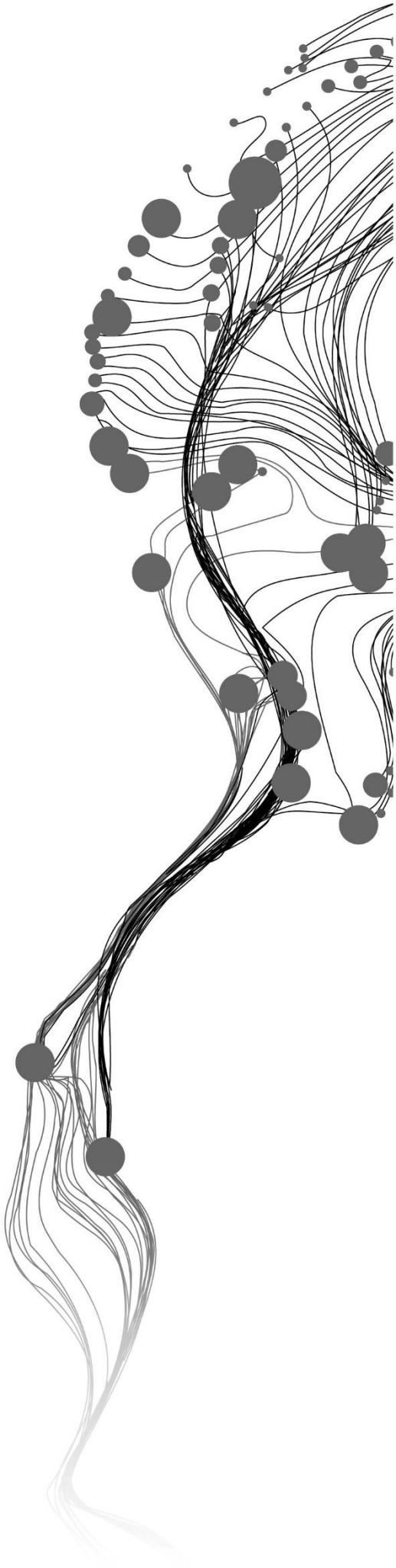
THE INTEGRATION OF THERMAL AND VISUAL IMAGES FOR HUMAN DETECTION USING DEEP LEARNING

QIAO REN
Feb, 2019

SUPERVISORS:
Dr. S. Hosseinyalamdary
Dr. F. Nex



THE INTEGRATION OF THERMAL AND VISUAL IMAGES FOR HUMAN DETECTION USING DEEP LEARNING

QIAO REN
Enschede, The Netherlands, Feb, 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. S. Hosseinyalamdary
Dr. F. Nex

THESIS ASSESSMENT BOARD:

ABSTRACT

Human detection is important in a wide range of applications nowadays. This research aims at finding the best deep learning approach in human detection by using thermal and visual images. This objective has been achieved by answering three research questions: What is the best human detection approach; How to integrate multiple sensors in the state-of-art; How much does multi-sensor approach improve the single-sensor approach.

The convolutional neural network RetinaNet has been found as a state-of-art approach. Its accuracy surpasses all the other deep learning methods, because of its innovative focal loss function and pyramid feature network.

This research implements two multi-sensor fusion models and four single sensor models, based on RetinaNet. Multi-sensor fusion models are based on two approaches: early-fusion and late-fusion. The early fusion model integrates thermal and visual images on pixel level. The late fusion model makes predictions by integrating the decisions that are provided by single-sensor models. The single-sensor models are trained on either thermal images or visual images. The single-sensor models are named as kaistT, cocokittikaistT, kaistRGB and cocokittikaistRGB.

The result shows that late fusion model has significantly improves the performance of single sensor models. The improvement made by late fusion model is 7.2 %, 3.9%, 11.2% and 8.2%, compared with kaist T, cocokittikaistT, kaist RGB and cocokittikaist RGB respectively. The improvements are large. Because late fusion model integrates the extracted feature from both thermal and visual images. Early fusion model does make improvement. Because there is information loss in image processing, before training RetinaNet.

Therefore, it is concluded that late-fusion model with RetinaNet is the best approach in human detection.

Keywords
Multi-sensor integration, deep learning, human detection, thermal images, visual images

ACKNOWLEDGEMENTS

First of all, I would like to express my thanks to my supervisor Dr S. Hosseinyalamdary (Siavash). Siavash is a very supportive supervisor. In the beginning of this research, Siavash gave me a series of lessons on deep learning algorithms. I appreciate that he taught me from scratch. With his guidance, I learned neural network step by step. I made a big progress in understanding the theory. When doing the research, I encountered a lot of difficulties, ranging from algorithms to coding, and from literature research to academic writing. Siavash is willing to give detailed guidance. Without his help, I won't move forward. During our discussion, he gave me a lot of inspirations. Furthermore, I also appreciate that Siavash takes care on not only my thesis but also my future career after graduation. He gave me a lot of useful advice on internship applications, job searching and academia applications.

Secondly, I would like to thank Dr. F. Nex (Fransceco). Fransceco gave me useful feedbacks on my thesis. His response is very quick.

Thanks to the committee members for giving me questions during the proposal defense and midterm defense.

Thanks to my friends and my classmates. Robert helped me with understanding the most difficult part of the code in RetinaNet. My classmates who also work on deep learning thesis, Yiwen Wang, Shan Huang, Ying Ao and Li Liu, gave me inspirations on this research. I enjoy the discussion that we had on convolutional neural networks.

Thanks to my family members. My mother and father gave me a strong support, mentally and financially.

TABLE OF CONTENTS

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation And Problem Satement

1.1.1. Motivation

Font Garamond 11. Human detection is essential for various applications (Liu, Zhang, Wang, & Metaxas, 2016). It is important in disasters management, the autonomous driving systems, automated surveillance and human-robotics interaction (Brunetti, Buongiorno, Trotta, & Bevilacqua, 2018):

1) when a disaster, such as earthquake and flooding, occurs the automatic human detection is able to aid the rescue work (Bharathi.V.S, 2005 ; Niels Gerlif, 2013). The location of victims can be detected and sent to the rescue team. An effective rescue scenario can be planned based on the location of the detected survivors. Without human detection, it is difficult for the rescue team to search for all the victims in the sophisticated disastrous regions.
2) In the autonomous driving systems, the human detection technique supports safety of the platform (Balani, Deshpande, Nair, & Rane, 2015). In order to ensure the safety of the pedestrians, the adjacent people should be detected, and their location should be estimated. If their distance is closer than a critical distance, the vehicle should slow down, detour, or halt.
3) Human detection plays a key role in automated surveillance (D, Manjunath, & Abirami, 2012 ; Moore, 2003). The access of human should be limited in certain areas, such as runways of airports museums in closed hours. In addition, automatic human detection algorithms can alert the owner of a house about the stranger intrusion.
4) In interaction with mobile robotics, robotics could provide better service to human customers if it detects the presence of human and his location (Moore, 2003). For example, in a seamless assistance system, the robot should be able to detect the location of the user and interact efficiently with him.

In conclusion, human detection is of importance in many applications and it is an ongoing research topic.

1.1.2 Problem Statement

This section aims at explaining the following questions. Why it is difficult to detect human? What are the drawbacks of visual and thermal images? Why it is necessary to integrate thermal and visual images for human detection? What are the challenges in sensor fusion? Why deep learning is better than other approaches in human detection?

Human detection is a challenging task (J. Liu et al., 2016). Because, people have different ages, genders, body shapes, positions, appearances. In addition, human can be captured from different views (Figure 1). Moreover, human may be occluded, which makes it difficult to detect human. (Figure 2)

Visual images and thermal images are the two major information sources used in human detection researches (Fan, Xu, Zhang, & Chen, 2008). Visual images are in RGB channel (Reyes-Ortiz, Oneto, Samà, Parra, & Anguita, 2016). Thermal images are visual displays of the amount of infrared energy emitted, transmitted, and reflected by an object (Correa, Hermosilla, Verschae, & Ruiz-del-Solar, 2012). As the temperature of an object increases, the amount of radiation emitted by the object increases. With the help of thermal images, warm objects like human become easily visible against the cool environment. An example of visual and thermal images in day and night is shown in Figure 3 (Hwang, Park, Kim, Choi, & Kweon, 2015).

Both visual images and thermal images have drawbacks (Kim, Hong, & Park, 2017). The drawback of visual images is that they are sensitive to illumination changes. Therefore, they are easily underexposed or overexposed in the sudden changes of illumination. In addition, they require sufficient illumination. Consequently, the quality of visual images deteriorates when the illumination is insufficient, such as at night, at dusk, in the shadow regions, and in foggy weather.

The disadvantage of thermal images is that, when the temperature of the background is high, for instance above 36 Celsius degrees, human detection in thermal images can easily be disturbed by the non-human environment (Kim et al., 2017; Baek, Hong, Kim, & Kim, 2017). Because, the difference between the temperature of human and non-human is small. This situation can happen in the daytime during hot summer. Besides, the resolution of thermal images is low (Fan et al., 2008), which causes some difficulties in human detection.

Sensor fusion is applied in human detection, in order to overcome the drawback of individual sensors (Baltrušaitis, Ahuja, & Morency, 2017). Sensor fusion means to integrate data generated by two different sensors (Baltrušaitis et al., 2017). In human detection, sensor fusion means to integrate visual and thermal images. These two types of images provide complementary detection decisions (Afsar, Cortez, & Santos, 2015). When illumination is sufficient, such as in a daytime, it is easy to detect human by visual images. When illumination is insufficient, such as during the night or dusk, it is easy to detect human by thermal images. The combination of visual and thermal images provides a robust algorithm and gives a higher accuracy in human detection, compared with using a single type of image (Wagner, Fischer, Herman, & Behnke, 2016).

There are challenges in sensor fusion. Firstly, images extracted by different sensors (thermal and visual camera) may have different size and resolution. It is necessary to convert both images to the same size and resolution. Secondly, thermal and visual images have different properties. Thirdly, occlusion makes it difficult to detect human. Fourthly, in convolutional neural network, which sensor should have larger weight.

The approach to implement sensor fusion is decided to be deep learning. Because deep learning is a highly efficient method, compared with other traditional methods like HOG (histogram of gradient) (Zhu et al., 2017). Traditional methods use a dictionary to store all the human in the training data (Wagner et al., 2016). It has two drawbacks (J. Liu et al., 2016). Firstly, its speed is slow, because it compares with all the objects in the dictionary with the unknown object in the image. Secondly, if a human in an image does not exist in the dictionary, then the traditional approach is not able to detect it. In contrast, deep learning, such as convolutional neural network, is able to capture the core of human features.

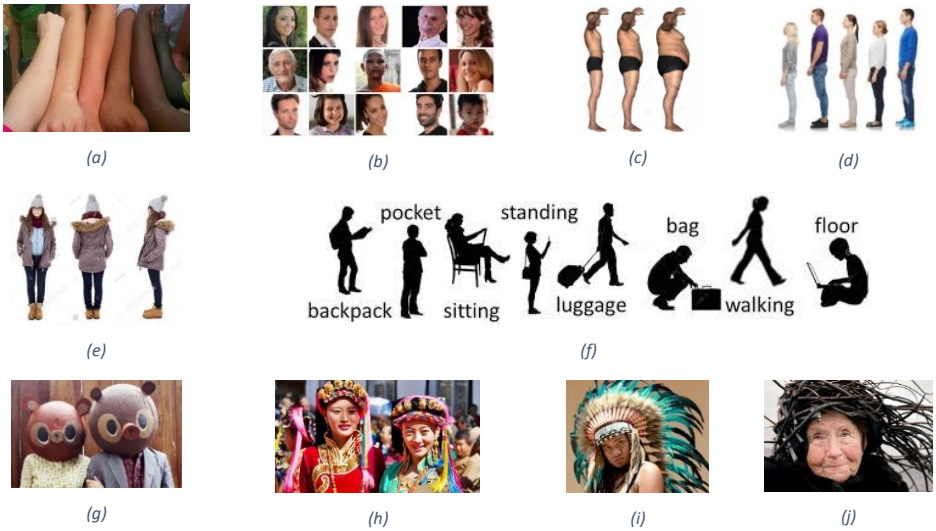


Figure 1 (a) human with different skin colors[1] (b) human with different genders and ages[2] (c) fat and slim people[3] (d) short and tall people[4] (e) the front, back and side view of a people captured by a camera[5] (f) different positions of human[6]. (g)[7] (h[8]) (i)[9] (j)[10] show people's clothes and headwear are in different colors, styles and can even be exaggerated.

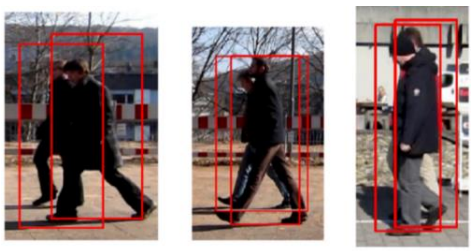


Figure 2 people who are occluded by other people[11]

[1] https://www.quora.com/Is-the-variation-in-human-skin-color-another-example-of-evolutions-natural-selection
[2] https://www.123rf.com/photo_24176012_collage-of-many-different-human-faces.html
[3] https://pt.dreamstime.com/illustration/homem-gordo-e-magro.html
[4] https://www.sciencedaily.com/releases/2017/12/171205115936.html
[5] https://www.123rf.com/photo_24138754_businesswoman-front-back-side-view-isolated.html
[6] http://mbaservicesllc.com/statement-of-position-2015-its-time-to-hit-your-stride/
[7] https://qzhyxx.com/tupian/%E5%8D%A1%E9%80%9A%E5%A4%B4%E5%A5%97.html
[8] https://www.pinterest.ca/pin/332914597439335503/
[9] https://www.etsy.com/sg-en/listing/208818837/feather-headdress-indian-style-green
[10] https://www.kickstarter.com/projects/eyesasbigasplates/eyes-as-big-as-plates/posts/1636340
[11] https://ps.is.tuebingen.mpg.de/publications/tangijcv

Figure 3 Image (a) and (b) are captured in daytime on the same scene(Hwang et al., 2015). Image (c) and (d) are captured at night time on the same scene(Hwang et al., 2015). (a) and (c) are visual images. (b) and (d) are thermal images.

1.2. Research Identification

1.2.1. Research Objectives

The main objective of this study is to find out the best architecture with the approach of convolutional neural network (CNN) applied in human detection. The architecture takes thermal and visual images as input and it detects the human in these images. We aim to test different architectures and provide the most accurate architecture in human detection.

1.2.2. Research Questions

The main objective is achieved by answering the following research questions:

Research Question 1: What is the best human detection approach?
Research Question 2: How to integrate multiple sensors in the state of art?
Research Question 3: How much does multi-sensor approach improve the single sensor approach?

1.2.3. Innovation Aimed At

The current state of art in object detection is RetinaNet. It is a robust one-stage object detector (Lin, Goyal, Girshick, He, & Piotr Dollar, 2018). RetinaNet has a special loss function to solve imbalance between foreground and background classes. Previous studies use Fast RCNN, Faster RCNN and ACF+T+THOG in human detection. However, the state-of-the-art approach surpasses the accuracy of those approaches (Lin et al., 2018). Therefore, we anticipate that our human detection approach using RetinaNet and multiple sensors outperform the previous human detection approaches.

Based on my knowledge, it will be the first time to adapt RetinaNet to multi-sensor integration on human detection. This is the first time that the RetinaNet is incorporated with different architectures of sensor fusion: early fusion level and late fusion.

2. LITERATURE REVIEW

Font Garamond 11. Various methods have been applied on human detection. There are two ways to categorize human detection approaches. From the aspect of sensor, methods can be classified by using single sensor and using multiple sensor. From the aspect of the accuracy and computation expenses, methods can be grouped by traditional approaches and deep learning approaches. An overview of this literature research is shown in Table 1

Using single sensor	Traditional approaches		HOG
			THOG
			Standard ACF
	Deep learning approaches	Two stage	R-CNN
			Fast R-CNN
			Faster R-CNN
		One stage	RatinaNet
			SSD[12]
			YOLO[13]
Using multiple sensors	Traditional approaches		ACF+T
			ACF+T+TM+TO
			ACF+T+THOG
			ACF+C+T
	Deep learning approaches	Multimodal applications	Applications in object detection
			Applications in human detection
		Fusion types	Pixel-level fusion
			Feature-level fusion
			Decision level fusion

Table 1 an overview of the literature research

2.1 Human Detection Using Single Sensor

2.1.1 Traditional Approach

One traditional approach applied in human detection is Histogram of oriented gradients (HOG) (Dalal & Triggs, 2005). HOG extracts the distribution of local intensity gradients along all the possible edge directions, in order to detect the appearance and shape of human. HOG takes visual images as input while THOG takes thermal images as input. In (Baek et al., 2017), thermal-position-intensity-histogram of oriented gradient (TPIHOG or TπHOG) has been proposed. TPIHOG or TπHOG improves nighttime pedestrian detection performance of HOG by incorporating thermal gradient information and its locations and thermal intensities. Currently, standard aggregated channel feature detector (ACF) is widely used as a basis algorithm on KAIST dataset (Dollar, Appel, Belongie, & Perona, 2014; Yang, Yan, Lei, & Stan Z. Li, 2014 ; Nam, Dollar, & Hee Han, 2014). Standard ACF uses color images as input. Standard ACF consists of 10 augmented channels, including color channels, gradient magnitude and gradient histograms (Hwang et al., 2015). This approach decreases computational costs substantially (Zhang, Bauckhage, & Cremers, 2014 ; Paisitkriangkrai, Shen, & Hengel, 2014).

[12] (Du, El-khamy, Lee, & Davis, 2017)
[13] (Redmon, Divvala, Girshick, & Farhadi, 2016)

2.1.2 Deep Learning

A bunch of deep learning approaches have been designed for human detection, including R-CNN, Fast R-CNN, Faster RCNN and RatinaNet.

The novelty of R-CNN (region-based convolutional neural network method) is to apply a region proposal (Girshick, Donahue, Darrell, Malik, & Berkeley, 2013) (Uijlings, Sande, Sande, & Smeulders, 2012). R-CNN selects candidate object locations and then a convolutional neural network (CNN) is implemented on each of the candidate locations. Classification and localization are finally be computed as output. The drawback of R-CNN is that it is slow at training-time (Girshick et al., 2013). Because it needs to run full process of CNN for watch region proposal. The other drawback is that CNN features are not updated in response to regressors (Girshick et al., 2013).

Compared with R-CNN, Fast R-CNN (Girshick, 2015 ; Hosang, Omran, Benenson, & Schiele, 2015 ; Li et al., 2017) improves training speed and detection accuracy. Li et al. (2016) proposed Fast R-CNN in pedestrian detection. Fast R-CNN processes the whole input image in CNN, generates a high-resolution feature map. Region proposal method is then implemented on the feature map. Fast R-CNN is faster than R-CNN during training time. However, computing region proposals still takes long time. Furthermore, Fast R-CNN is not applicable for real time detection, because the test time for each image is slow (Girshick, 2015).

Faster R-CNN incorporates Fast R-CNN network and a region proposal network (RPN) (Ren, He, Girshick, & Sun, 2015) (J. Liu et al., 2016). RPN is trained to produce high quality region proposals. So there is no need to do external independent region proposals, compared with R-CNN and Fast R-CNN (Hosang et al., 2015; Li et al., 2017). After RPN is trained, the process is the same as Fast R-CNN. The computation speed of Faster R-CNN is much faster than Fast R-CNN. Liu et al. (2016) applied Faster R-CNN on color image and visual image separately for human detection.

RetinaNet, introduced by Lin et al. (2018), achieves a higher accuracy compared with aforementioned neural networks on COCO dataset. RetinaNet, as a one stage detector, is different from previous two stage detectors. It consists a backbone network and two subnetworks, one for classification and one for generating bounding box. This research proposed a new loss function in order to address the imbalance between foreground and background classes. RetinaNet-101 is the benchmark of this research.

2.2 Human Detection Using Multiple Sensor

2.2.1 Traditional Approaches

Many traditional multimodal approaches in human detection are extensions of the standard ACF, including ACF+T (Hwang et al., 2015) , ACF+T+TM+TO (Hwang et al., 2015), ACF+T+THOG (Hwang et al., 2015 ; Baek et al., 2017), and ACF+C+T (J. Liu et al., 2016 ; Wagner et al., 2016). ACF+T incorporates thermal intensity as an additional channel (Hwang et al., 2015). ACF+T+TM+TO adds three channels to Standard ACF: thermal intensity, normalized gradient magnitude of thermal images and the histogram of oriented gradients of thermal images. ACF+T+THOG (Hwang et al., 2015 ; Baek et al., 2017) incorporates a contrast enhanced version of the thermal images and HOG features of thermal image as channels. ACF+C+T detector used 10-channel aggregated features to fuse color and thermal images (J. Liu et al., 2016). Generally speaking, The extensions of ACF surpass the accuracy of standard ACF (Hwang et al., 2015). ACF+T+THOG gives the highest accuracy among these extension approaches (Hwang et al., 2015).

2.2.2 Deep Learning

Multi-sensor fusion approaches can be divided into three categories: pixel-level fusion, feature-level fusion and decision-level fusion (Zhu et al., 2017; Baltrušaitis, Ahuja, & Morency, 2017 ; Wagner et al., 2016). Pixel-level fusion means to integrate the information contained in multiple images of the same scene into one image (Y. Liu et al., 2018). Feature level fusion integrates features immediately after they are extracted and it learns the correlation and interactions between features (Baltrušaitis et al., 2017; Zhu et al., 2017). It is implemented on convolutional layers. Decision-level fusion integrates the outputs of multiple sub-neural-networks features (Baltrušaitis et al., 2017; Zhu et al., 2017). It is implemented on the last fully connected layers.

Multimodal fusion by deep neural network has been widely used in object detection. For instance, Images and texts are merged for visual questions answering (Visual QA) (Wu, Shen, Wang, Dick, & Van Den Hengel, 2016). Audios and videos are integrated for speech recognition (Mitra et al., 2016; Ngiam et al., 2011). Images and videos are fused for Action Recognition (Karpathy et al., 2014 ; Simonyan & Zisserman, 2014). The integration of RGB visual images and dense depth images (RGB-D) is used for 3D geometric reconstruction of static and dynamic scenes (Zollhöfer et al., 2018). Human pose detection and hand gesture recognition are two examples that apply RGB-D (Zollhöfer et al., 2018). Besides, multimodal fusion are also used in medical imaging, digital photography, remote sensing and video surveillance (Y. Liu et al., 2018).

Human detection is a sub-branch of object detection. Two sensors, visual and thermal detector, are integrated in human detection (Wagner et al., 2016). This is denoted by RGB-T. In previous researches, different deep neural networks are blended with various levels of fusion. Wagner et al. (2016) applied R-CNN in pixel level fusion and decision Level fusion. J. Liu et al. (2016) implemented Faster R-CNN on Feature level fusion and decision level fusion. This research suggests that halfway fusion on feature level gives the highest accuracy. Guan et al. (2018) introduced a deep neural network which is able to be aware of illumination so that it learns human-related feature under different illumination conditions (day and night).

Previous studies have shown that multi-sensor fusion gives higher accuracy than single-sensor applied in object detection. Deep learning approaches are faster and more robust than traditional approaches. Among all the deep learning approaches, RetinaNet gives the highest accuracy.

3. DATA PREPARATION

Three datasets have been used in used in this research: Kaist dataset, Coco dataset and Kitti dataset. The descriptions are in the following sections: 3.1, 3.2 and 3.3.

3.1. Kaist dataset

Data used in this study is provided by KAIST (Korea Advanced Institute of Science and Technology) which is a research university in South Korea. The KAIST Multispectral Pedestrian Dataset consists of 95,000 colour-thermal pairs. They are captured by a vehicle which carries a colour camera and a thermal camera. The images are captured during day and night time. As an example, a pair of visual and thermal

images with annotation is shown in Figure 1. The visual images are in size of about 200 kb while the thermal images are in size of about 77 kb. All the images have the same dimension 640 pixels * 512 pixels and the same resolution 96 dpi * 96 dpi. Human in all the pairs have been labelled. In total, there are 53293 annotated pedestrians in training dataset. There are 2742 annotated pedestrians in testing dataset. The data is available online[14]. Kaist dataset only has two classes: persons and others.

The testing dataset uses the corrected annotations (table x). Because there some errors in the original annotations of testing dataset (Appendix 1). The errors mainly occur in the testing dataset. In order to eliminate those mistakes, this research uses the annotations which has been corrected[15]. They are corrected by Liu.

Kaist dataset is very challenging. It includes images in day and at night. It includes persons with different postures, such as cycling, walking, standing and sitting. It includes persons in different locations, such as in a pedestrian and on a highway road. It also includes persons in occlusion, in a far distance and in a shadow.

Figure 4 Two images in the KAIST dataset. Image (a) is a visual image. Image (b) is a thermal image. Image (a) and (b) were captured in the same location at the same time. Annotations are shown in red bounding box. In image (a) and (b), there are two annotated persons. The annotations are provided by KAIST dataset.

	Time Period	Locations	Training dataset	Testing dataset
Scenes	Day	Campus	Set 00 : 17,498 images	Set 06 : 648 images
	Day	Road	Set 01 : 8,035 images	Set 07 : 406 images
	Day	Downtown	Set 02 : 7,866 images	Set 08 : 401 images
	Night	Campus	Set 03 : 6,668 images	Set 09 : 175 images
	Night	Road	Set 04 : 7,200 images	Set 10 : 444 images
	Night	Downtown	Set 05 : 2,920 images	Set 11 : 178 images
Total number of images			**50000 images**	**2252 images**
Total number of annotated persons			**53293 bounding boxes**	**2742 bounding boxes**

[14] https://sites.google.com/site/pedestrianbenchmark/home
[15] http://paul.rutgers.edu/~jl1322/resource/annotations_KAIST_testset.tar

3.2. Coco dataset

Coco dataset [16] has 91 object classes[17] (Appendix 2). Person is one of the 91 classes. Coco provides segmentation masks for every object instance. Coco data is downloaded from this webpage[18]. Coco dataset has 330K images. 250,000 people are annotated. Figure 6 shows the examples[19] of segmentation instances in Coco dataset.

Figure 6 example images in coco dataset

3.3. Kitti dataset

Kitti dataset has 8 object classes: pedestrian, cyclist, car, van, truck, sitter, tram and misc. Images are captured by two cameras: a color and a grayscale camera. The locations to capture images include rural areas and highways in the city Karlsruhe, Germany. The annotation is provided by Velodyne laser scanner and a GPS localization system. Kitti dataset is downloaded from this website[20]. An example[21] is shown in Figure 7.

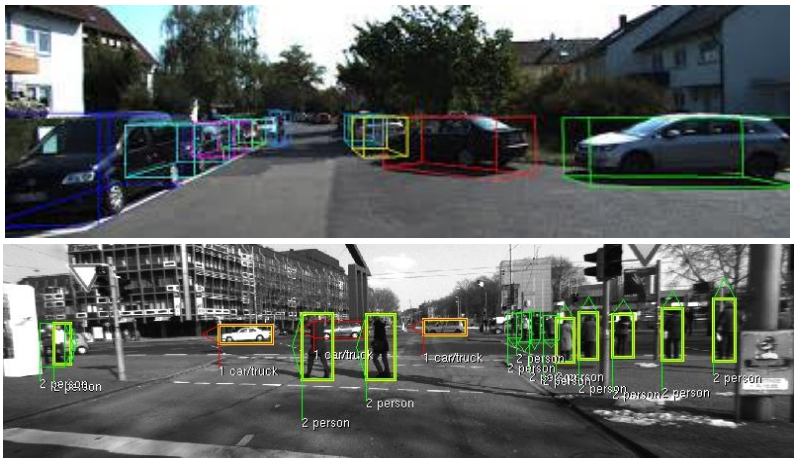


Figure 7 example images in kitti dataset

[16] https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/
[17] https://tech.amikelive.com/node-718/what-object-categories-labels-are-in-coco-dataset/
[18] http://cocodataset.org/#download
[19] https://github.com/nightrome/cocostuff
[20] http://www.cvlibs.net/datasets/kitti/
[21] http://www.cvlibs.net/datasets/karlsruhe_objects/

4. METHODOLOGY

4.1. Method Overview

The workflow of this research is demonstrated in Figure 9. Six models have been trained and evaluated: an early-fusion model, a late-fusion and four single sensor models.

Input of fusion models are pairs of thermal and visual images. Input of single-sensor models are either thermal or visual images. The training process of all six models are based on a convolutional neural network (CNN) RetinaNet. Trained models are tested on test dataset by detecting the location of human.

Output of each model contains two parts: bounding boxes and corresponding scores. A Bounding box is a rectangle with the smallest perimeter within which the whole human lies. A bounding box is defined by four parameters: coordinates of the left top corner of the rectangle, a height and a width (Figure 8). Each bounding box has a corresponding score. Score is the degree of certainty to which there is a human on a certain location. The range of a score is from 0 to 1.

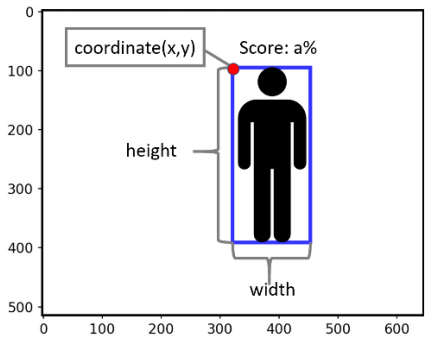


Figure 8 output of a human detection model

Finally, the quality of the human detection will be evaluated. By comparing the quality of multi-sensor and single-sensor models, the model which provides the highest quality will be chosen as the best model. The next two sections will give a detailed explanation.

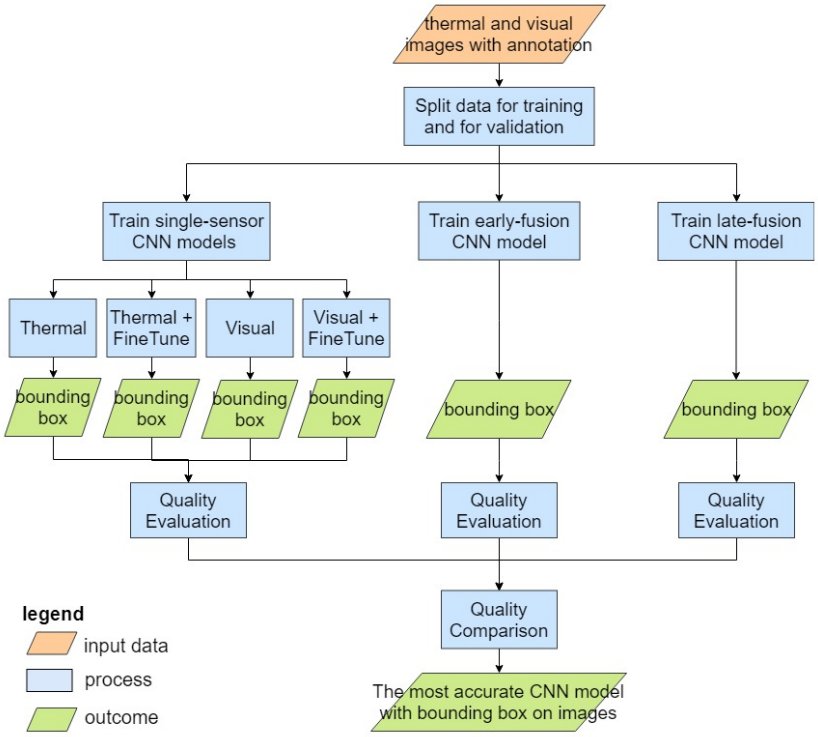


Figure 9 overview of this research

Fusion of sensors can be done in different layers of a deep network. Sensor fusion is categorized into early, halfway, and late fusion.

Early fusion is also called pixel-level fusion. It integrates images from different sensors on pixel level in image processing. In early fusion, the output of multiple sensors should be consistent. For instance, the thermal and visual images can be integrated in pixel level, if the resolution of two sensors are the same. The image manipulation can be applied to adjust two sensor data, but the image manipulation loses information and adds artifacts in the images. Figure 10a shows the early fusion when the two sensors are integrated in the very early layer of the network.

In the late fusion, also known as decision-level fusion, the two sensors are independently applied to achieve the desired output. At the end, the output of these independent network is integrated based on the output of each sensor. For instance, several cameras may look at the scene from different perspectives for scene understanding, the images of each camera are applied to classify the objects in the scene and a voting scheme is utilized to integrate the classified objects of each camera and make final decision. Figure 10b shows the late fusion and two output results of the network are integrated in the last layers.

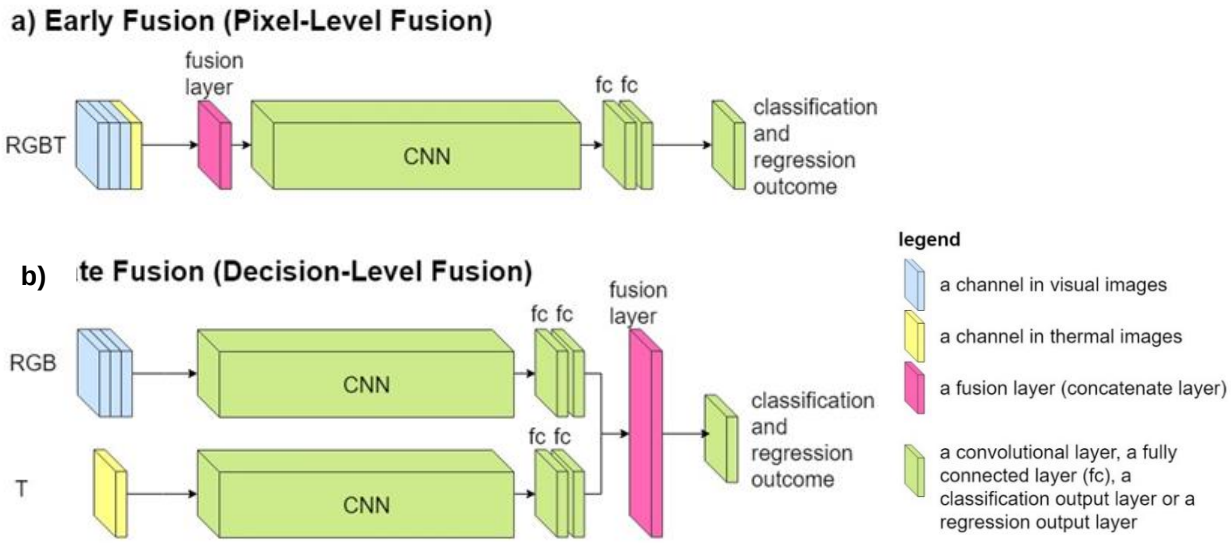


Figure 10 early fusion and late fusion models

4.2. Assumptions

- It is assumed that the time of capturing visual images and thermal images has been synchronized. If the visual images and thermal images are not captured simultaneously, then it is assumed that the time has been calibrated. Otherwise, it is out of the scope of this study.
- It is assumed that the location of visual camera and thermal camera are at the same location. If they are at different location, then it is assumed that the location has been calibrated.
- It is assumed that the labeled ground truth is perfect and correct. This means that all the rectangular bounding box of labeled human are in correct position and with correct size. There is no object which is not a human but labeled as a human. There is no object which actually is a human but not labeled.
- It is assumed that if only a part of human present in an image, such as a finger, a hand or a foot, it is unnecessary to detect it as a human. Because it is incomplete.

4.3. Why to use RetinaNet

RetinaNet is chosen as the state-of-art CNN, because of two reasons. RetinaNet is innovative in solving two problems.

1) using Focal Loss to solve class imbalance problem. Firstly, RetinaNet is able to solve the problem class imbalance. Class imbalance means the objects of interesting classes (such as human) occurs much rare than the objects of non-interesting classes (such as road, vehicles, trees). Most of the locations contribute to no-useful information. Training is inefficient. RetinaNet solves this problem by using a special loss function, called focal loss. Focal loss reduces the contribution of outliers. Outliers are the samples which are hard to be detected. "The novel focal loss focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the detector during training." [22]

[22] https://medium.com/@14prakash/the-intuition-behind-retinanet-eb636755607d

2) Using Feature Pyramid Network to solve information loss problem in feature extraction:
RetinaNet is able to solve the problem that some feature information are lost in pooling layers. Pooling layers are responsible for down sampling the image. It causes the problem that high-resolution maps[23] in early layers contributes to weak features, while the low-resolution maps in late layers contribute to strong features (Figure 11). This is harmful to object detection. To solve this problem, RetinaNet uses a feature pyramid network (Figure 12). "It combines low-resolution, semantically strong features with high-resolution, semantically weak features. "[24] Therefore, this feature pyramid network has rich semantics at all scales. Layers are interconnected. Each layer has its own classification head and regression head.

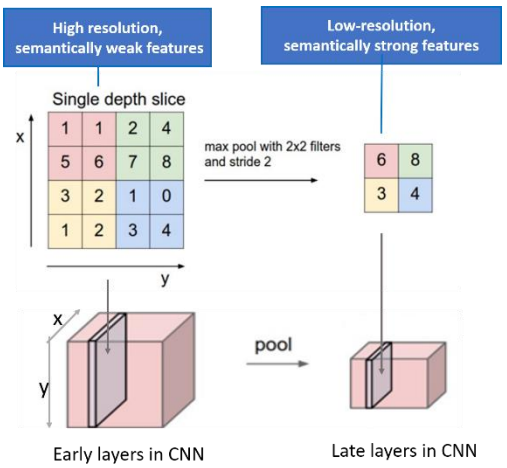
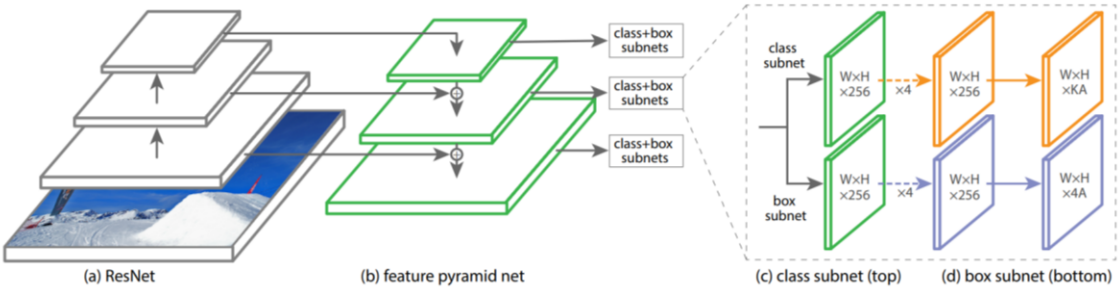


Figure 11 high-resolution maps in early layers contributes to weak features, while the low-resolution maps in late layers contribute to strong features



The one-stage **RetinaNet** network architecture uses a Feature Pyramid Network (FPN) [20] backbone on top of a feedforward ResNet architecture [16] (a) to generate a rich, multi-scale convolutional feature pyramid (b). To this backbone RetinaNet attaches two subnetworks, one for classifying anchor boxes (c) and one for regressing from anchor boxes to ground-truth object boxes (d). The network design is intentionally simple, which enables this work to focus on a novel focal loss function that eliminates the accuracy gap between our one-stage detector and state-of-the-art two-stage detectors like Faster R-CNN with FPN [20] while running at faster speeds.

Figure 12 feature pyramid network

[23] https://medium.com/@14prakash/the-intuition-behind-retinanet-eb636755607d
[24] https://towardsdatascience.com/review-fpn-feature-pyramid-network-object-detection-262fc7482610

4.4. Single Sensor Models

Single sensor models are the models which are trained with either thermal or RGB images. There are two purposes to use single sensor models in this research. Firstly, they are used to compare with multi-sensor fusion models. Secondly, the best single sensor models are the input of late-fusion model.

An important approach applied in single sensor models is fine tuning. Fine-tuning is the process in which parameters of a model is adjusted very precisely in order to fit with certain observations. Fine tuning requires a model to learn features from a broad domain in order to help learning features from a specific domain. For example, training a model on animal images helps the model to learn features on human images. The advantage of fine-tuning is to speed up the training process and to overcome small dataset size. Detailed explanation is in the section Implementation.

4.5. Early Fusion Architecture

Early fusion means to integrate thermal and visual images on pixel level before training CNN (Figure 6 a). In this research, the proposed method is to convert RGBT to HST (Hue, Saturation and Thermal) (Figure 7). Firstly, RGB is converted to HSL. Then the illumination band (L) in HSL is replaced by thermal band (T). This process is shown in Figure 13 and Figure 14. The mathematical algorithm that converts RGB to HLS is in appendix 3 [25] . The removed band is illumination, because this research aims at that providing a CNN such that illumination has no influence on the prediction outcome. The reason that RGBT has to be converted to a 3-band image is that the maximum amount of input bands in RetinaNet is three. It is impossible to feed a 4-band image (R+G+B+T). HST solves this problem.

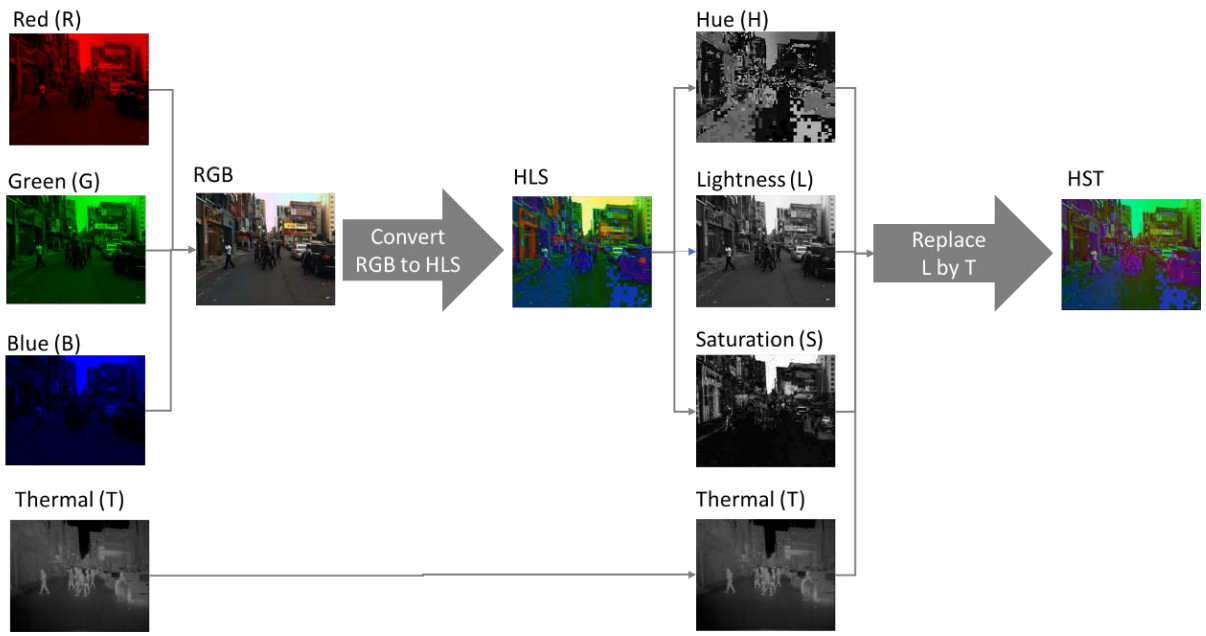


Figure 13 the approach to convert RGBT bands to HST bands

[25] https://docs.opencv.org/3.1.0/de/d25/imgproc_color_conversions.html#color_convert_rgb_hls

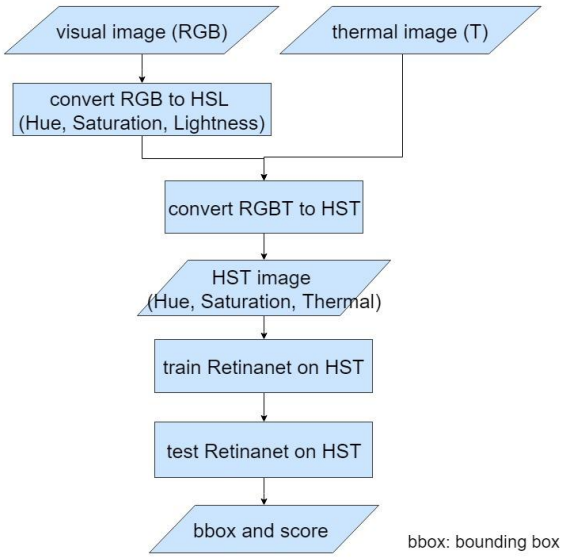


Figure 14 workflow of building up an early fusion model

4.6. Late Fusion Architecture

Late fusion aims at integrating the decisions which has been provided by thermal model and RGB model. The workflow of late fusion is demonstrated in Figure 19.

Step 1: use single-sensor models to predict bounding boxes. For each pair of thermal and visual images, there are bounding boxes which are predicted by a thermal model and bounding boxes which are predicted by an RGB model.

Step 2: **decision integration**. The following two cases need to be analysed in order to integrating the bounding boxes from thermal model and RGB model.

The first case (Figure 15) is that there is no overlap between a T bounding box and an RGB bounding box. In this case, both of bounding boxes are generated as the final output. Because thermal and RGB models extract different features. A person who is detected by thermal model may not be detected by RGB model and vice versa. Generating bounding boxes from both models integrates the competences of both models.

The second case (Figure 16) is that there is an overlap between a T bounding box and an RGB bounding box. Then the question is whether both of them should be the output or only one of them should be the output. The algorithm to solve this problem is the following. If the overlap between two boxes is small, then the two bounding boxes are interpreted as detecting different persons. If the overlap is large, then the two bounding boxes are interpreted as detecting the same person. Because it is very likely to happen that there is a small difference between the locations which are detected by two different models. Overlap is defined as IOU (intersection over union) between thermal and RGB bounding boxes. The way to distinguish whether the overlap is large or small is to set a threshold of IOU. This threshold is called as merge threshold. If IOU is smaller than the merge threshold, then both of the bounding boxes should be generated in the output. If IOU is larger than the merge threshold, then only one bounding box should be generated. This brings the next question: which bounding box is more trustworthy than the other? This is

found by comparing the scores of the thermal and RGB boxes. The bounding box with larger score is more trustworthy. Therefore, this bounding box is the output.

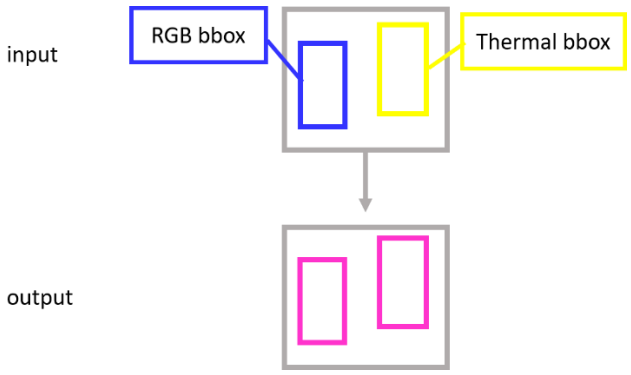


Figure 15 in case there is no overlap between Thermal and RGB bounding boxes

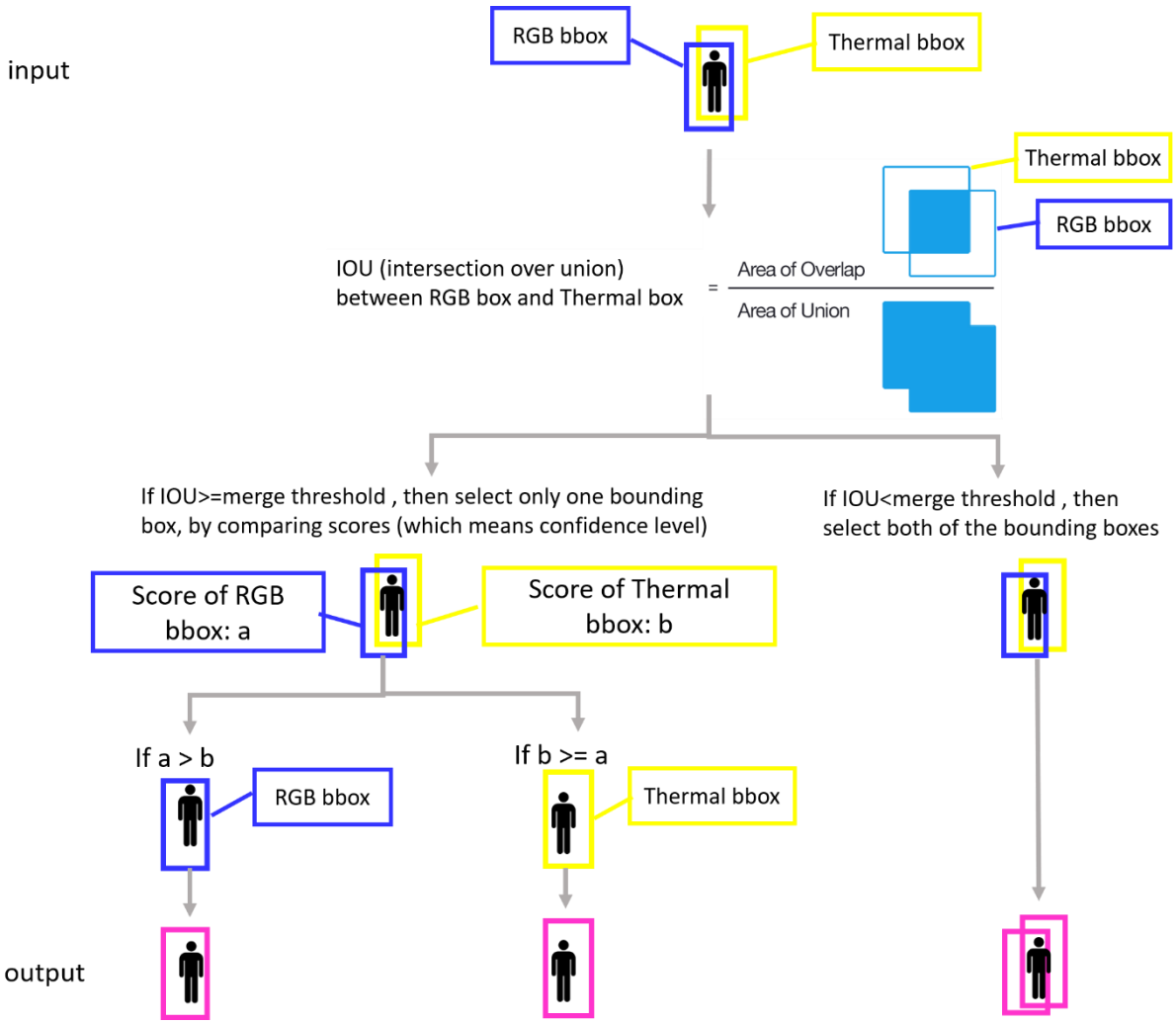


Figure 16 in case there is an overlap between Thermal and RGB bounding boxes

4.7. Quality Assessment

The approach of quality assessment is to evaluate three indexes under the condition of different score thresholds. The three indexes are recall, precision and average IOU. This approach is based on the literature review (Han, Zhang, Cheng, Liu, & Xu, 2018). The formulas in quality assessment are the following equations.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$average\ IOU = \frac{IOU\ of\ all\ the\ predicted\ bounding\ boxes}{total\ number\ of\ predicted\ bounding\ boxes}$$

$$F1\ score = \left(\frac{recall^{-1} + precision^{-1}}{2}\right)^{-1} = 2\frac{\text{precision} * \text{recall}}{precision + recall}$$

The terms in formulas are explained in this paragraph. True positive (TP) is the number of human who are correctly predicted by a model. False positive (FP) is the number of non-human who are incorrectly predicted as human. False negative (FN) is the number of human who are incorrectly predicted as non-human. Precision tells how many of the selected objects were correct. It is a measure of completeness (Powers, 2007) [26]. Recall tells how many of the objects that should have been predicted are actually predicted. It is a measure of exactness (Powers, 2007) [27]. IOU is intersection over union between two bounding boxes. It measures to what degree that a predicted bounding box and an annotated bounding box are overlap. This is illustrated in Figure 17. IOU ranges from 0 to 1. IOU is 1 when these two bounding boxes are completely overlap. Average IOU measures the average difference between predicted human locations and annotated locations, in the whole testing dataset.

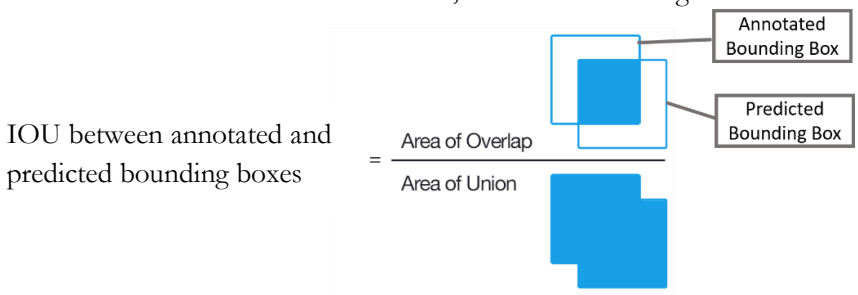


Figure 17 IOU between annotated and predicted bounding boxes

The criteria to define a true positive is to set a threshold of IOU. The threshold is called as IOU threshold or true positive threshold. In this research, the true positive threshold is set as 0.7. This means that if

[26] http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf

[27] http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf

IOU is larger than this threshold, then this predicted bounding box is a true positive. Otherwise, it is a false positive.

A model has different precision and recall, under different score threshold. A score threshold is used to filter out the bounding boxes which has low confidence level. As shown in Figure 14, when the score threshold is set too low (0.1 in Figure 18a), the predicted bounding boxes with low confidence appear (score 38.4%, 12.8%, 21.3% and 31.1%). False positive would increase. A low score threshold causes recall to be high and precision to be low. When the score threshold is too high (Figure 18c), a lot of human are not detected. Because the model is not confidence enough to make that prediction. False negative increases. This causes precision to be high and recall being low.

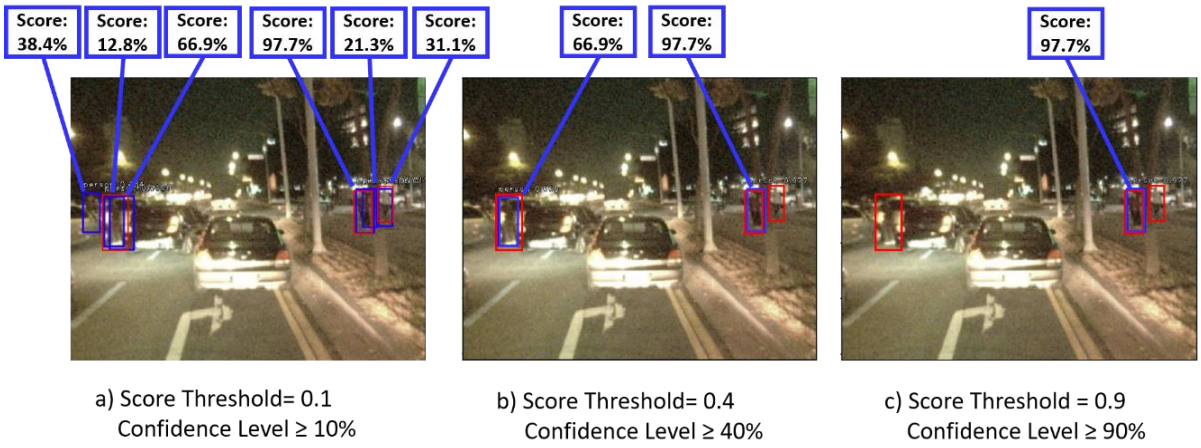


Figure 18 predicted bounding boxes caused by different score threshold

The approach to find out the best performance of a model is to find out the largest F1 score. F1 score is the harmonic average between precision and recall. Both precision and recall are considered into the computation of F1 score. F1 score measures the trade-off between precision and recall. F1 score is 0 in the worst case. F1 score is 1 in the best case. For each model

The workflow of quality assessment is shown in Figure 19. Each model is tested nine times with different score thresholds: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. If the score of a predicted bounding box is smaller than the score threshold, then it is discarded. If it is larger than the score threshold, then it is generated as a prediction. To determine whether a predicted bounding box is true positive or false positive, IOU is compared with a true positive threshold. After the total true positive is counted, precision and recall are calculated. F1 score indicates the best performance of a model.

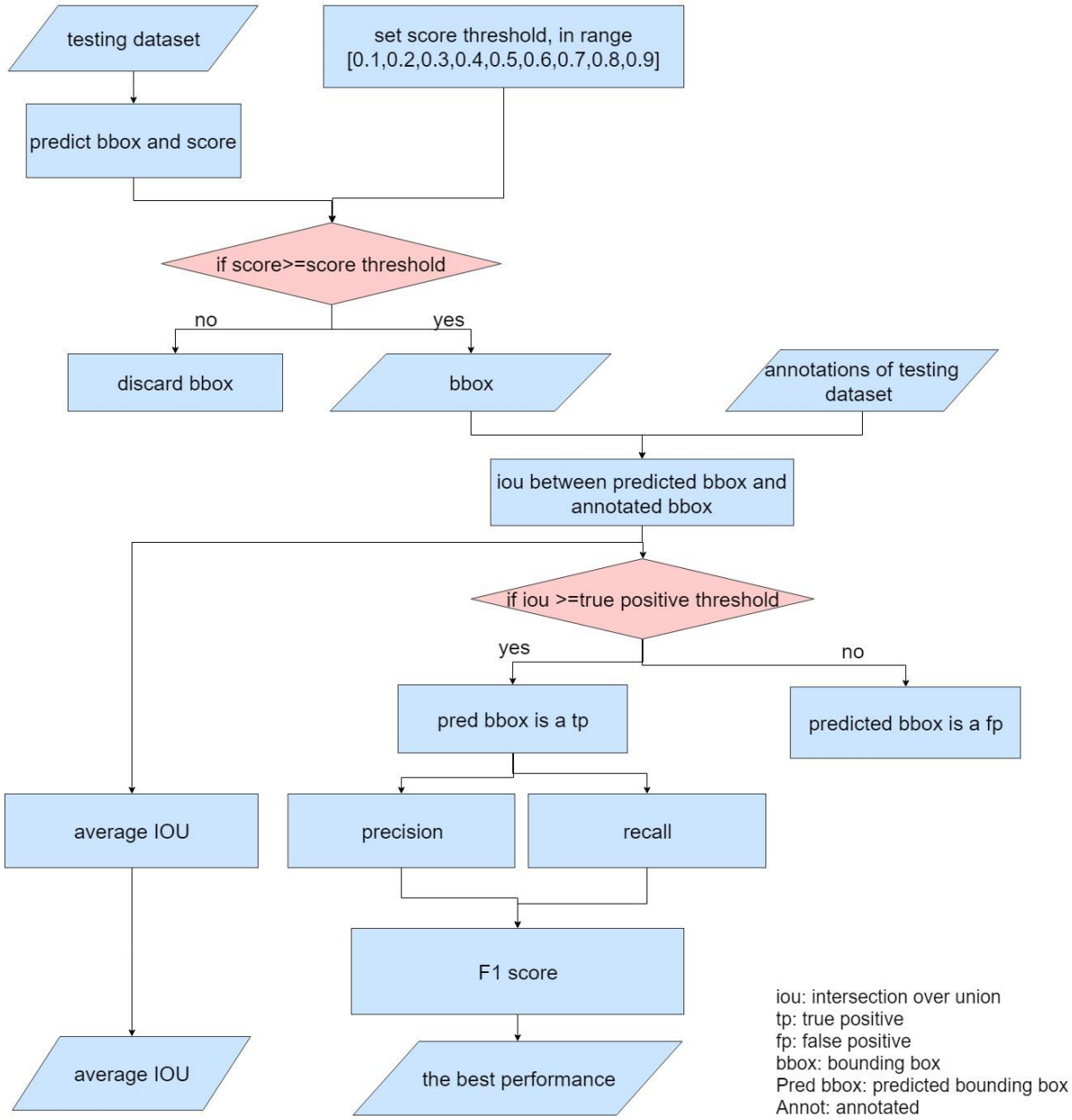


Figure 19 method of qualilty evaluation

5. EXPERIMENTAL SETUP

The following parameters has been set in the training process.

- Backbone: Resnet50. Resnet 50 is a good choice of backbone. Because it has a high accuracy with low operation costs compared with other backbones (Appendix 4).
- Number of epochs: 50. This number is set as default.
- Number of steps per epoch: 10000
- Size of batches: 1. There is a limitation in the hardware. When batch size increases, the computer crashes. So default setting 1 is used.

6. IMPLEMENTATION

This section explains how the single-sensor and multi-sensor models have been trained. The time spent to train a model is about 40 hours. The programming language used in the implementation process is: Python. Programming platform is Keras and Tensoflow.

6.1. Single Sensor Models

RetinaNet has been trained on single-sensor Kaist dataset. The default weights in RetinaNet are the weights that have been trained by ImageNet dataset. Therefore, two models have been generated:

- Model kaistT: a model trained with thermal images. The weights were initialized by default.
- Model kaistRGB: a model trained with visual images. The weights were initialized by default.

Besides that, fine tuning has been implemented. Because fine-tuning improves adjust the weights precisely. The competence of CNN to distinguish human and nonhuman features would be improved. It is anticipated that using broad datasets, such as coco and kitti dataset, to implement finetuning will improve the accuracy of a model. The relation between ImageNet, Coco, Kitti and Kaist datast is shown in Figure 20. The dataset represented by the large circle contains the classes of the dataset represented by the small circle. ImageNet contains more classes than Coco. Coco contains more classes than Kitti. Kitti contains more classes than Kaist. Finetuning requires a model to transfer what it learns from a broad domain to a specific domain.

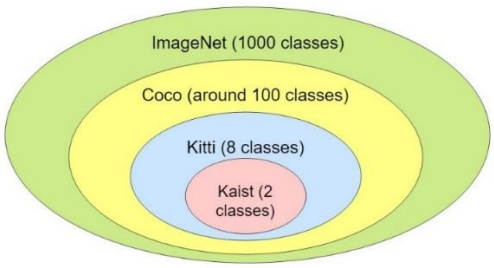


Figure 20 relation between four dataset: ImageNet, Coco, Kitti and Kaist datasests

In order to find out the best approach of implementing finetuning, two processes have been made:

- Train RetinaNet model with only kitti dataset.
- Train RetinaNet model with Coco dataset and Kitti dataset.

After training RetinaNet with these two approaches, the result shows that finetuning with the combination of Coco and Kitti provides higher average precision than finetuning only with kitti. As shown in Table 2, the average precision of pedestrian on CocoKitti model (0.44) is higher than it on Kitti model (0.34). The average precision of pedestrian on CocoKitti model (0.46) is higher than it on Kitti model (0.39).

	firstly finetuned by Coco, then finetuned by kitti	only finetuned by kitti
Average precision of pedestrian	0.44	0.34
Average precision of cyclist	0.46	0.39

Table 2 a comparison between two approaches of finetuning

Therefore, two models with finetuning are generated: cocokittikaistT, cocokittikaistRGB. They were firstly trained on Coco dataset, then trained on Kitti dataset and finally trained on Kaist dataset.

- Model cocokittikaistT: a model trained with thermal images. The model has been finetuned on Coco and Kitti dataset before training on Kaist dataset.
- Model cocokittikaistRGB: a model trained with visual images. The model has been finetuned on Coco and Kitti dataset before training on Kaist dataset.

To sum up, there are four single-sensor models generated in this research: model kaistT, model kaistRGB, model cocokittikaistT and model cocokittikaist-RGB. The workflow is shown in Figure 21.

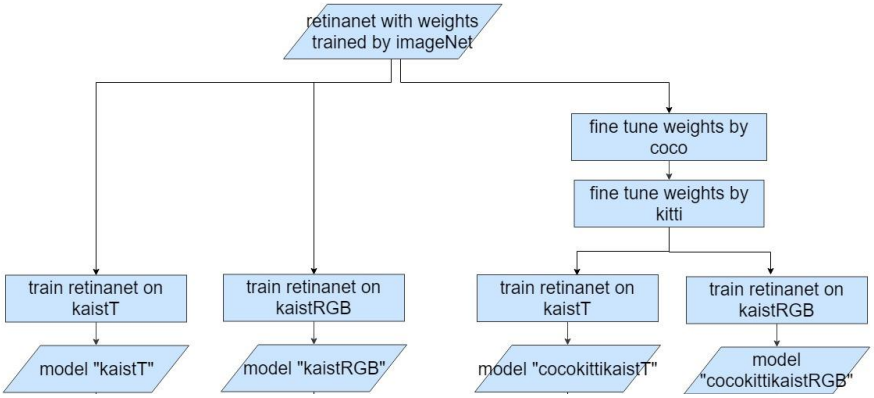


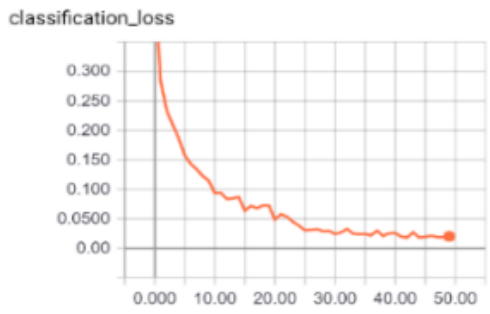
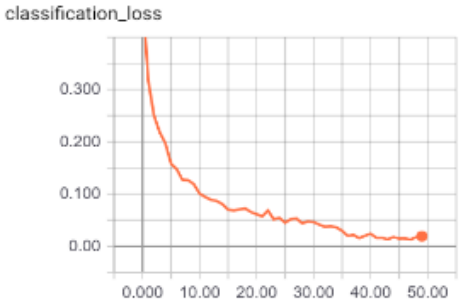
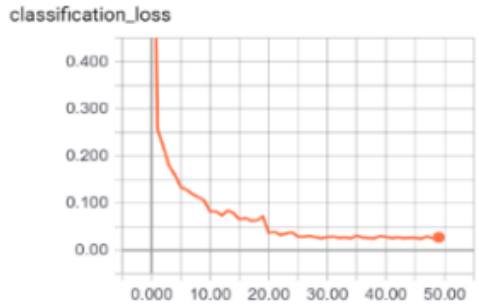
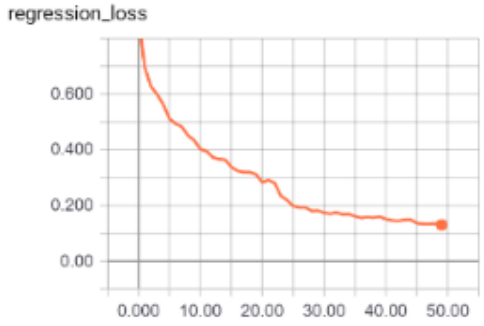
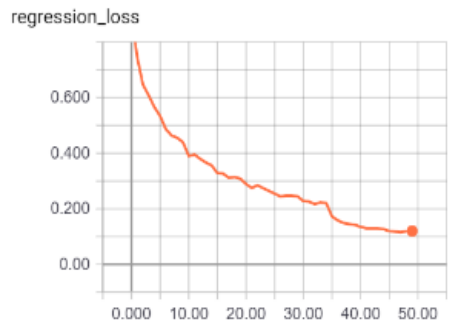
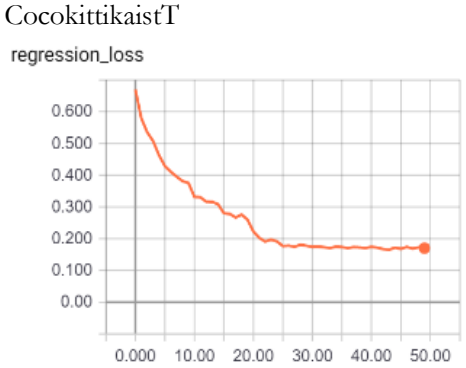
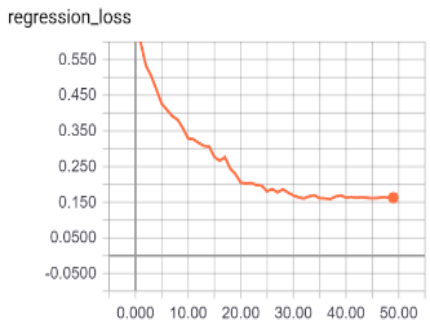
Figure 21 the implementation to train four single-sensor models

To evaluate the quality of training process, loss functions are applied. Loss function is "a measure of how good a prediction model does in terms of being able to predict the expected outcome"[28] There are three loss functions during training process: classification loss, regression loss and loss (which is total loss). They are defined as:

- Classification Loss: it indicates the difference between predicted classes and the classes annotations. RetinaNet uses Focal Loss to calculate classification loss.
- Regression Loss: it indicates the difference between predicted locations and the locations in ground truth. RetinaNet uses Standard Smooth Loss to calculate regression loss.
- Loss: Total Loss = Classification Loss + Regression Loss.

The loss functions of single-sensor models are shown in Table 3. The horizontal axis is the number of epochs, ranging from 0 to 49. The value of loss function is very large in the first epoch. Because the weights in early layers of a convolutional neural network has not been adjusted to the input dataset. With the increase of epochs, the value the loss function decreases. Because the network is learning features from input dataset. Weights are adjusted to fit the data. Until a certain epoch, the value of loss function reaches a convergence. This means that the prediction that a model makes is very close to the ground truth in dataset. In Table 3, the convergence values in all the models are significantly small. This means that all the models have been well trained.

[28] https://heartbeat.fritz.ai/5-reg KaistRGB ression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0

Classification loss	KaistT 	KaistRGB 
	CocokittikaistT 	CocokittikaistRGB 
Regression loss	KaistT 	KaistRGB 
	CocokittikaistT 	CocokittikaistRGB 
loss	KaistT	KaistRGB

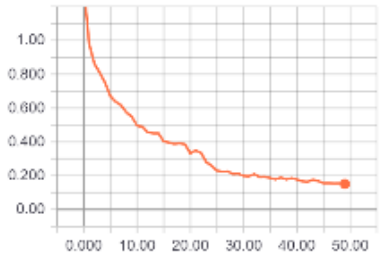
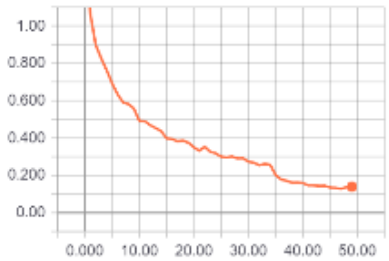
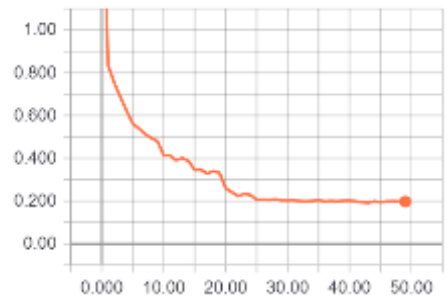
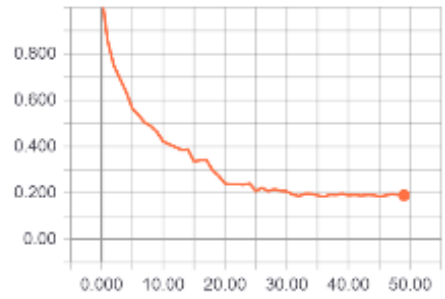
	loss 	loss 
	CocokittikaistT loss 	CocokittikaistRGB loss 

Table 3 classification loss, regression loss and total loss of four single-sensor models

6.2. Early fusion Models

The input data of early fusion model is HST (Hue Saturation Thermal) images, as explained in methodology. They are generated from Kaist dataset set0-set5. The HST images are divided into two parts: training and validation dataset. 90% of set0-set5 is for training dataset. The rest 10% is for validation dataset. Validation images are randomly selected from each set in set0-set5. This ensures that in each scene, 10% of images are in the validation dataset.

The loss function is shown in Table 4. All the loss functions have reached their convergence. The loss value in the final epoch is very small. This means that the prediction made by early-fusion model is significantly close to the ground truth.

Mean average precision (mAP) in early fusion is 0.97, which is very high. Mean average precision (mAP) computes the average value of precision with respect to recall (Han et al., 2018). "mAP is the average of the maximum precisions at different recall values"[29]. The higher the mAP is, the higher the accuracy is. The formula of mAP is in Appendix 5.Table 4, mAP increases with the number of epochs increases. mAP reaches its convergence at epoch 30.

[29] https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173

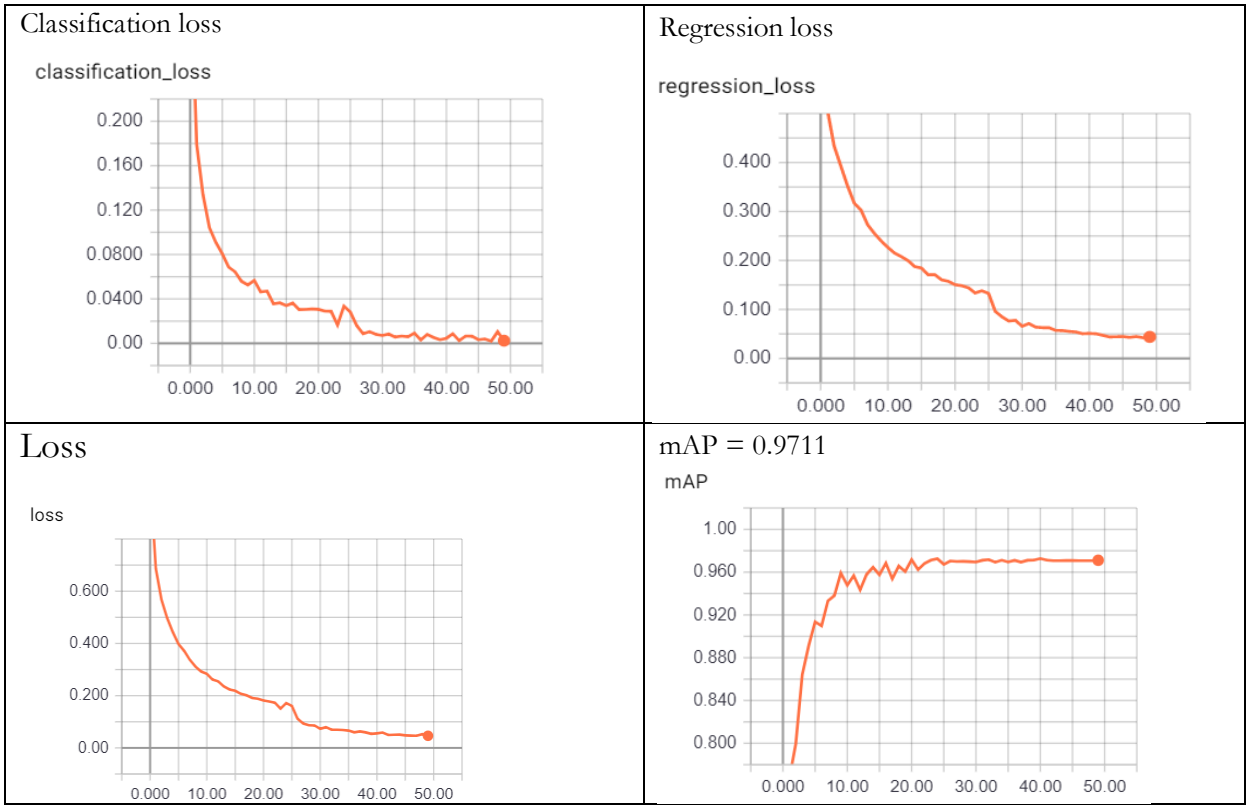


Table 4 classification loss, regression loss, total loss and mean average precision of the early-fusion model

6.3. Late fusion Models

Late fusion model aims at integrating the decision of single-sensor models. The single-sensor models are considered as sub-networks in late fusion models.

The sub-networks of late fusion are the best model on thermal images and the best model on RGB images. In this research, the model Cocokittikaist T with score threshold 0.2 has the best performance on thermal images. The model Cocokittikaist RGB with score threshold 0.2 has the best performance on visual images. The evaluation process is illustrated in the section quality assessment. The best models are found by evaluating the precision and recall of all the score thresholds in four single-sensor models. Each single-sensor model is tested on testing dataset. The best performance of a model is determined by the score threshold which provides the largest F1 score. After that, the best performance of kaistT and cocokittikaistT are compared, in order to find out the best model with the best score threshold on thermal images. Same process goes for two RGB models.

The workflow of late fusion is shown in Figure 22. Each sub-network predicts bounding boxes on images. If T box and RGB box are not overlapped, then both of them are predicted. If they are overlapped, then IOU (intersection over union) needs to be calculated. If IOU is larger than the merge threshold, then the bounding box with larger score will be generated. Otherwise, both of the bounding boxes are generated as the output. This process has been executed for multiple times with different merge threshold, until the best merge threshold has been found. The final output of late fusion model is bounding boxes which is

provided by the best merge threshold. The bounding boxes can be visualized on both thermal and visual images.

An important step in building up the late fusion model is to find out the best threshold to merge T box and RGB box. This is called "merge threshold". It is necessary to set a reasonable merge threshold. Because if the merge threshold is too large, then two bounding boxes which detect the same person will be interpreted detecting different persons. The amount of false positive will increase. If the merge threshold is too small, then the person who has been detected by one of the models will be missed out. False negative will increase. Figure 25 visualizes the output bounding boxes generated by different merge threshold. The input are: T boxes (blue bounding boxes) and RGB boxes (yellow bounding boxes). The different values of merge threshold cause different output bounding boxes (white bounding boxes). When merge threshold is 0.6 (or 60%), only one bounding box is generated on one person. When merge threshold is 0.8 (or 80%), redundant bounding boxes are generated.

The method to optimize the merge threshold is to evaluate the precision and recall. Late fusion model has been ran with different merge threshold. For each candidate merge threshold, a trade-off between precision and recall is evaluated. The best merge threshold is the one which provides the largest F1 score. The candidate merge thresholds are 0.6, 0.7, 0.8 and 0.9. It has been found that the best merge threshold is 0.7. The recall and precision corresponding to each merge threshold is shown in table x. F1 score is calculated based on them. Figure 24 visualizes the relation between precision and recall. The value precision, recall and F1 score is in Table 5. When merge threshold is 0.7, F1 score reaches its maximum, which is 0.3416. Therefore, 0.7 is set as the best merge threshold.

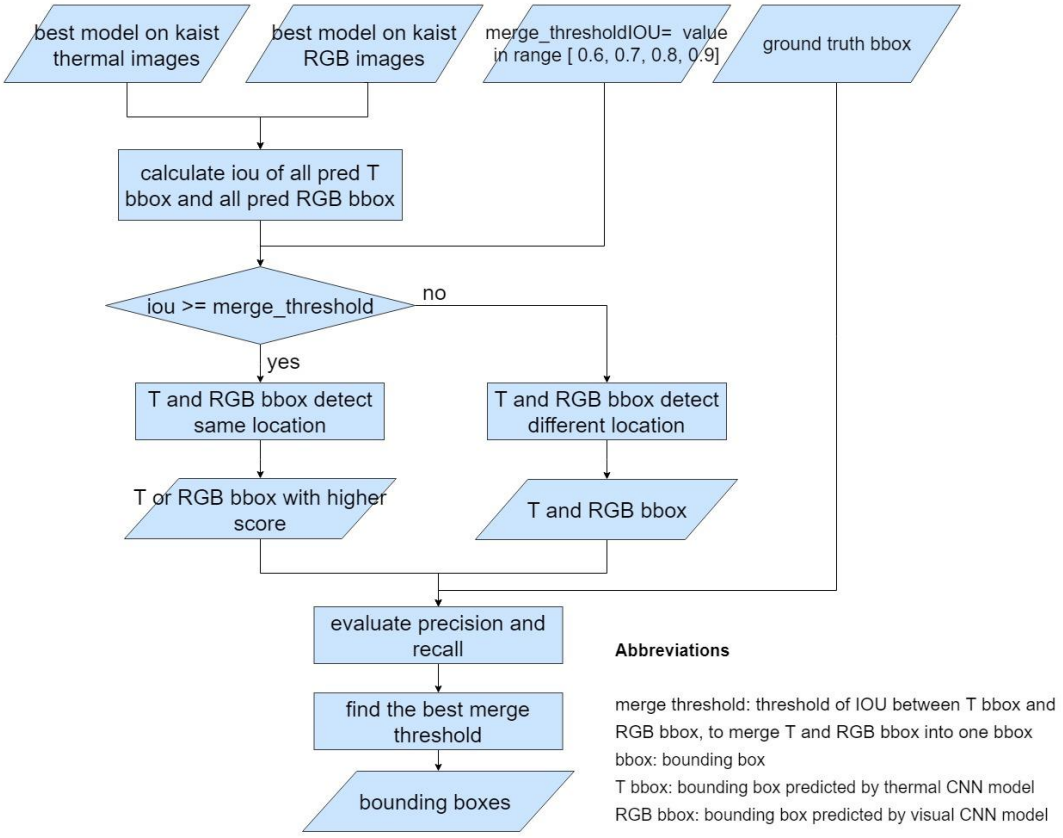


Figure 22 workflow of late fusion

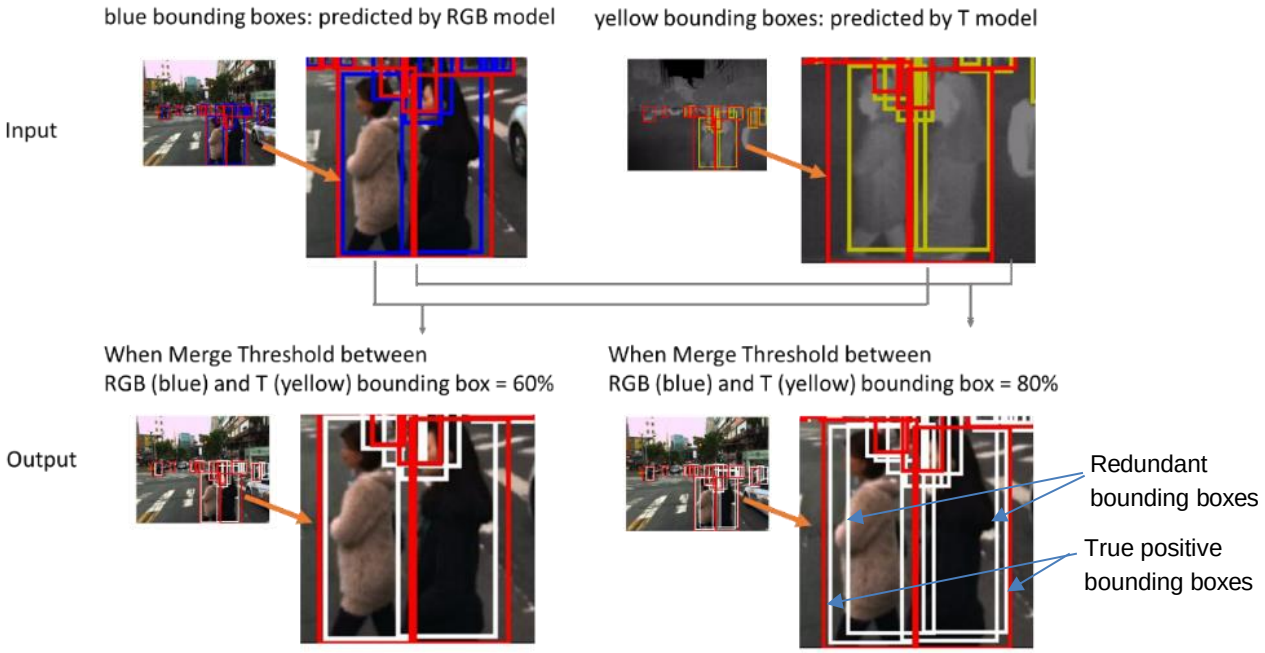


Figure 25 The output bounding boxes are different when merge threshold has different value. In this examples, two merge thresholds are tested: 60% and 80%. White bounding box is the output of late fusion. Red bounding boxes are annotations.

Candidate merge threshold	0.6	0.7	0.8	0.9
Recall	0.2962	0.3112	0.3291	0.3868
Precision	0.3955	0.3787	0.3551	0.3338
F1 score	0.3381	0.3416	0.3415	0.3353

Table 5 presicion and recall of different merge threshold

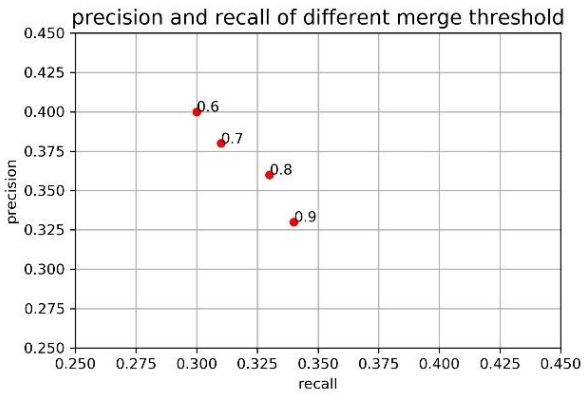


Figure 24 recall and precision of candidate merge thresholds: 0.6, 0.7, 0.8, 0.9

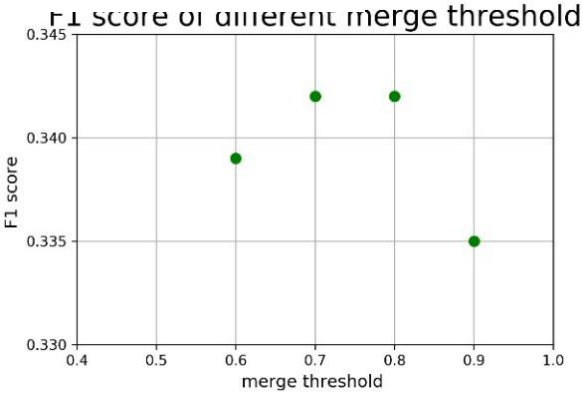


Figure 23 F1 score of different merge threshold

7. RESULTS

The result section provides qualitative analysis (8.1) and quantitative analysis (8.2).

7.1. Qualitative Analysis

Figure 27 and Figure 28 shows the predicted outcome of six models, on day and night

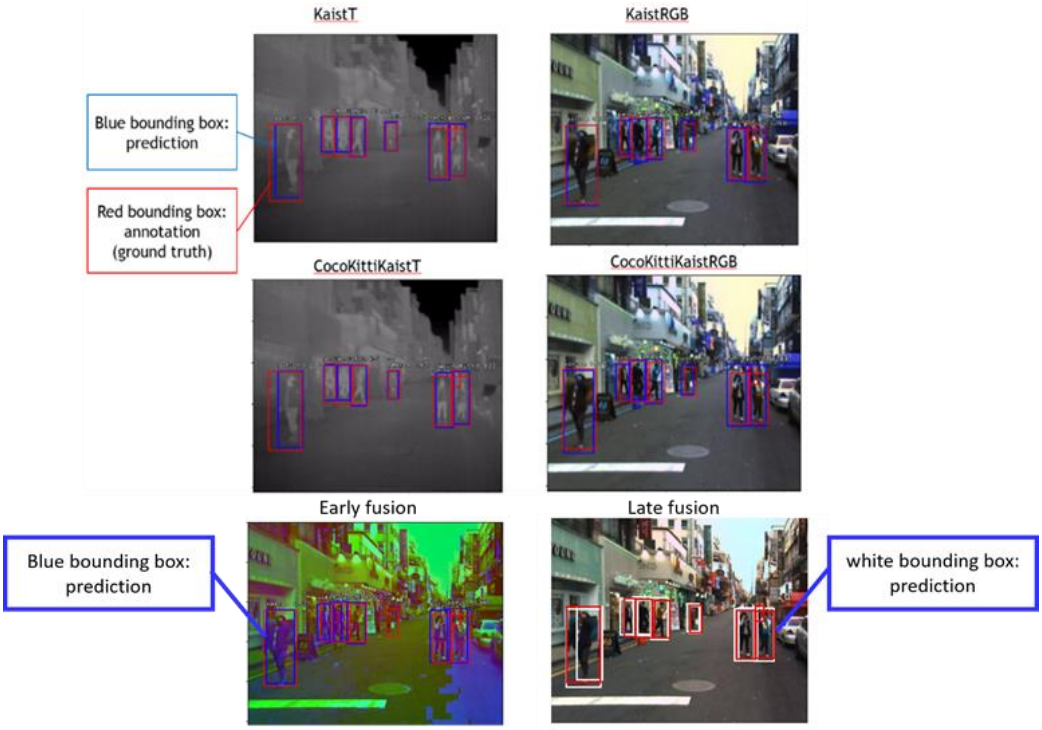


Figure 26 bounding boxes predicted by six models on day images

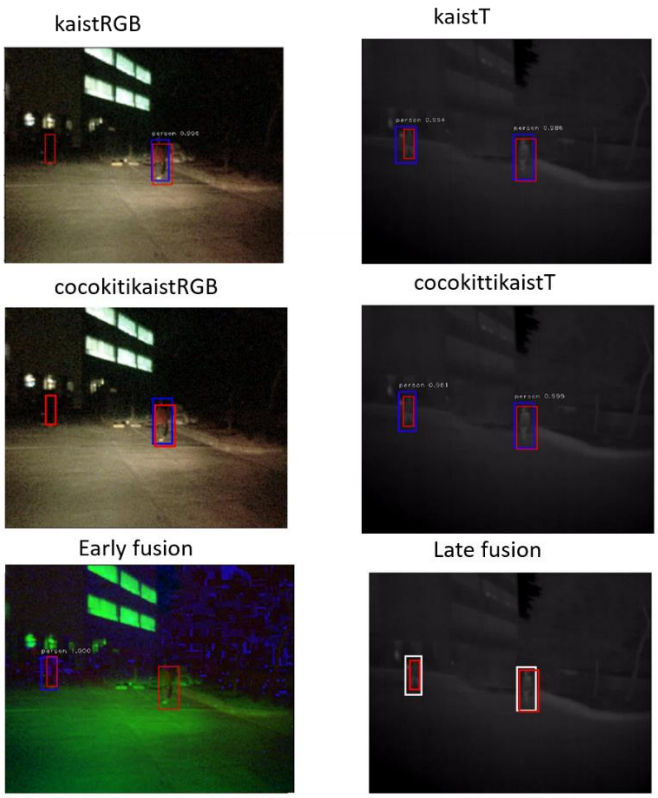


Figure 27 bounding boxes predicted by six models, on night images

The result of late fusion shows that it successfully integrates the decision from RGB model and T model. In Figure x, there is a person that has been detected by RGB model but not T model. In Figure x, there is a person that has been detected by T model but not RGB model. In both of these cases, late fusion model successfully detects the person. This matches with our anticipation. It shows that the late fusion model improves the detection in challenging situations: when a person is in low illumination, in a shadow, is far away from the camera or is occluded.

In the Figures, blue bounding box is predicted by RGB model. Yellow bounding box is predicted by T model. White bounding box is predicted by late fusion model. Annotation is represented by red bounding box.

The output of late fusion is shown in two images. The reason is that the algorithm of decision integration is independent from T or RGB images. Because learning features from images have been completed by single sensor models. Late fusion only focuses on analysing bounding boxes and generating bounding boxes.

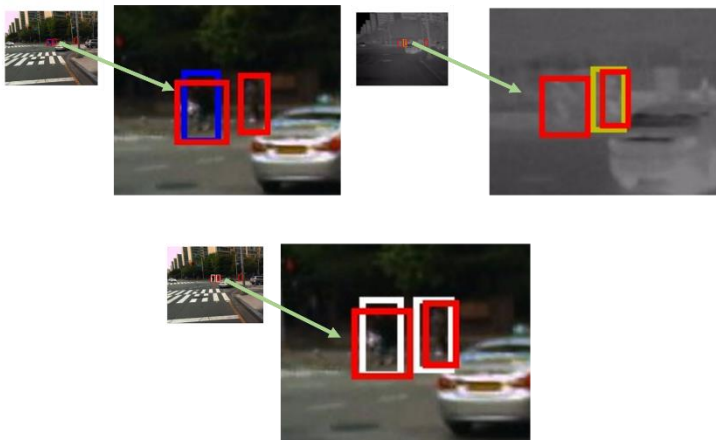


Figure 28 example of late fusion images. The first row are two inputs. The image on the second row is the output.

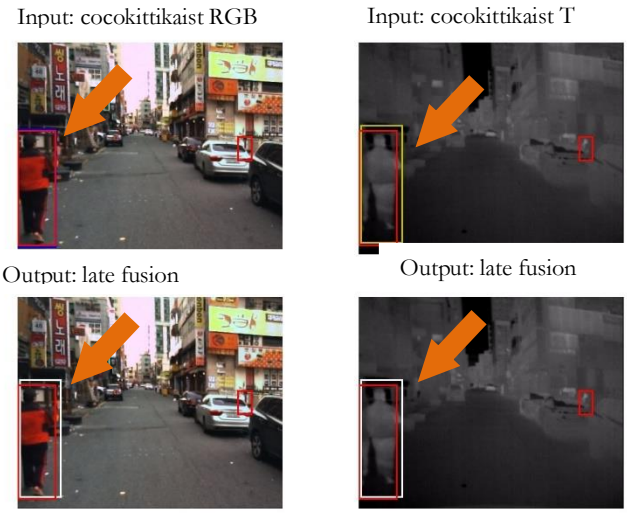


Figure 29 lan example of late fusion outcome

7.2. Quantitative Analysis

Each model is evaluated with nine different score thresholds, ranging from 0.1 to 0.9. Running a model with a score threshold generates three values: precision, recall and average IOU. The numerical result of single-sensor models and multi-sensor fusion models are shown in three tables. Table 6 shows the precision and recall of six models. Table 7**Error! Reference source not found.** shows average IOU of six models. Table 8 shows F1 score of six models.
The best performance of a model is determined by the largest F1 score. Based on the calculation of precision and recall, the best performance has been found for each model. This is shown in Table 9.

In all the precision-recall graphs, F1 sore graphs and average IOU graphs, the best performance of a model is represented by a large cirlcle. The non-best performances are represented by small circles. In all the precision and recall graphs, precisions and recall shows a negative correlation. Precision decreases with the recall increases. The reason has been explained in the section quality assessment. The following sections make comparative analysis on the models.

precision and recall of single-sensor and multi-sensor-fusion models

	kaistT		cocokittikaistT		kaistRGB		cocokittikaistRGB		early fusion		late fusion	
score threshold	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall
0.1	30.06%	25.36%	28.07%	30.61%	35.59%	19.18%	32.02%	23.02%	41.35%	11.24%	39.96%	32.84%
0.2	36.85%	23.65%	38.49%	27.74%	39.65%	18.15%	39.62%	21.58%	43.61%	10.91%	39.96%	32.84%
0.3	40.48%	22.43%	43.24%	25.41%	42.08%	16.78%	42.91%	20.47%	45.02%	10.63%	39.96%	32.84%
0.4	42.62%	21.30%	46.36%	23.06%	43.34%	16.97%	45.46%	19.65%	47.00%	10.48%	41.82%	31.17%
0.5	44.66%	20.15%	49.01%	20.92%	44.71%	16.13%	47.67%	18.55%	48.02%	10.25%	44.66%	29.13%
0.6	46.78%	18.95%	52.47%	18.99%	45.60%	15.00%	49.89%	17.47%	48.63%	9.99%	47.75%	26.99%
0.7	48.26%	17.30%	54.86%	16.71%	46.99%	14.10%	51.94%	16.36%	49.82%	9.83%	50.29%	24.61%
0.8	51.53%	15.84%	58.02%	14.03%	49.59%	12.98%	54.22%	14.50%	50.57%	9.45%	53.26%	21.34%
0.9	54.55%	12.69%	61.99%	10.08%	51.73%	10.50%	56.99%	11.50%	52.23%	8.79%	56.46%	16.22%

Table 6 precision and recall of single-sensor models and multi-sensor-fusion models

average IOU of single-sensor models and multi-sensor-fusion models						
score threshold	kaistT	cocokittikaistT	kaistRGB	cocokittikaistRGB	early fusion	late fusion
0.1	55.49%	52.51%	59.75%	57.17%	62.11%	62.09%
0.2	60.21%	59.32%	62.46%	62.44%	63.37%	62.09%
0.3	62.51%	62.43%	63.92%	64.62%	63.93%	62.09%
0.4	63.87%	64.30%	64.61%	66.09%	65.00%	63.40%
0.5	65.11%	66.03%	65.58%	67.32%	65.59%	65.18%
0.6	66.26%	67.90%	66.14%	68.38%	66.06%	66.89%
0.7	67.12%	69.05%	66.80%	69.18%	66.45%	68.04%
0.8	68.42%	70.46%	67.99%	70.30%	66.76%	69.48%
0.9	69.57%	72.15%	69.21%	71.24%	67.48%	70.83%

Table 7 average IOU of of single-sensor models and multi-sensor-fusion models

F1 Score of single-sensor models and multi-sensor-fusion models						
score threshold	kaistT	cocokittikaistT	kaistRGB	cocokittikaistRGB	early fusion	late fusion
0.1	0.275	0.293	0.249	0.269	0.177	0.361
0.2	0.288	0.322	0.249	0.279	0.175	0.361
0.3	0.289	0.32	0.249	0.277	0.172	0.361
0.4	0.284	0.308	0.244	0.274	0.171	0.357
0.5	0.278	0.293	0.237	0.267	0.169	0.353
0.6	0.27	0.279	0.226	0.259	0.166	0.345
0.7	0.255	0.256	0.217	0.249	0.164	0.33
0.8	0.242	0.226	0.206	0.229	0.159	0.305
0.9	0.206	0.173	0.175	0.191	0.151	0.252

Table 8 F1 score of single-sensor models and multi-sensor-fusion models

Model name	Score threshold when a model reaches its largest F1 score	Recall	Precision	F1 score	Average IOU
kaistT	0.3	22.43%	40.48%	0.289	62.51%
cocokittikaistT	0.2	27.74%	38.49%	0.322	59.32%
Kaist RGB	0.3	17.68%	42.08%	0.249	63.92%
Cocokittikaist RGB	0.2	21.58%	39.62%	0.279	62.44%
Early fusion	0.1	11.24%	41.35%	0.177	62.12%
Late fusion	0.3	32.84%	39.96%	0.361	62.09%

Table 9 Best performance of each model

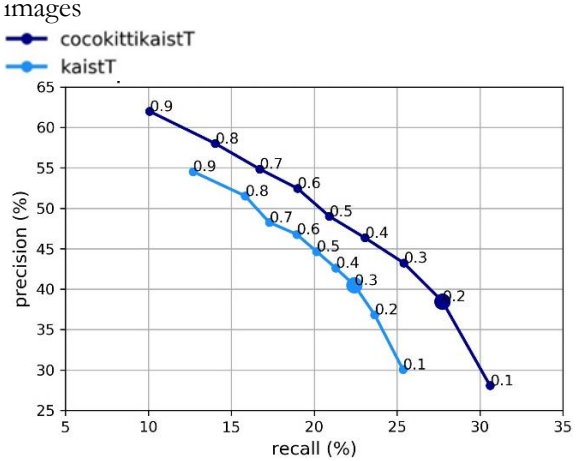
7.3. Single Sensor Models

It is observed that the models fine-tuned on coco kitti dataset have larger F1 score than the models which have only been trained on ImageNet dataset. Figure 30 shows a comparison between model kaistRGB and cocokittikaistRGB. Figure 30 a) shows that the precision of cocokittikaistRGB is larger than kaistRGB, when score threshold is in range 0.2 to 0.9. The recall of cocokittikaistRGB is also larger than kaistRGB. Figure 30 b), with the same score threshold, the F1 score of cocokittkaistRGB model is larger than the F1 score of KaistRGB. The improvement that cocokittikaistRGB has made on kaistRGB is 3%. This is calculated by using the F1 score of the best performance of cocokittikaistRGB minus the the F1 score of the best performance of kaistRGB. Same observation is found on thermal models. The best performance of CocokittikaistT is 3.3% higher than kaistT.

The reason that finetuned model has better performance is that the weights has been precisely adjusted to the true values. A model transfers what it learns from a broad dataset to a specific dataset. In Coco dataset, classes are very broad, including dogs, elephants, cars and bicycles. CNN learns the similarities and differences between human and non-human objects. Besides, human appearances in Coco dataset is broader than Kaist dataset. Coco contains a wide range of human postures: sitting, running, cycling and doing sports. In Kaist dataset, the postures of human are limited: mostly walking and standing. Kitti dataset uses a GPS localization system which is very unique from other datasets. It equips Kitti data with precise locations of human. This is very helpful for training RetinaNet to detect the locations of human.

Models trained on thermal images has better performance than models trained on RGB images. As shown in Figure 33, the best performance of cocokittikaistT (big dark blue dot) has larger F1 score than the best performance of cocokittikaistRGB (big brown dot). the best performance of kaistT has larger F1 score than the best performance of kaistRGB. The reason is that RGB images are very sensitive to the illumination. If a person is in a shadow, under a tree or in a dusk, RGB images only shows black. No useful information is provided to the model. In kaist dataset, thermal and visual images have the same resolution. So resolution does not cause the difference on RGB model and thermal model.

a) Precision and recall of models on thermal images



b) F1 score of models on thermal images

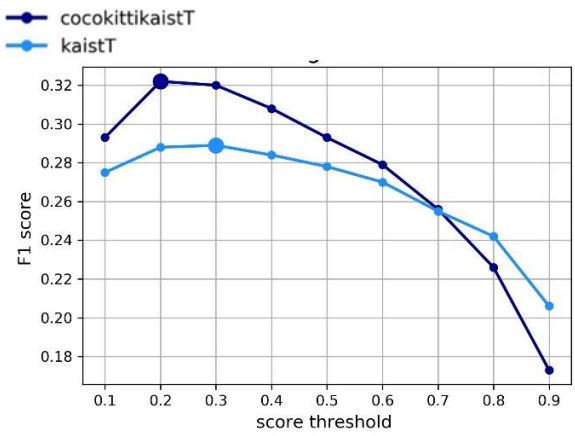
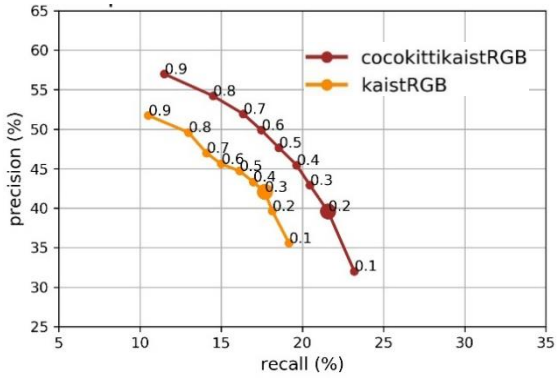


Figure 30 a comparison between model cocokittikaistT and kaistT

a) Precision and recall of models on visual images



b) F1 score of models on visual images

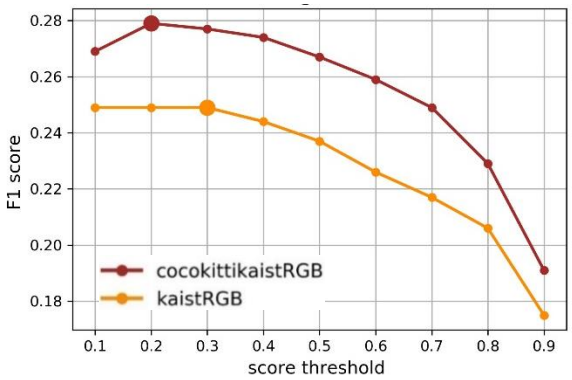


Figure 31 a comparison between cocokittikaistRGB and kaistRGB

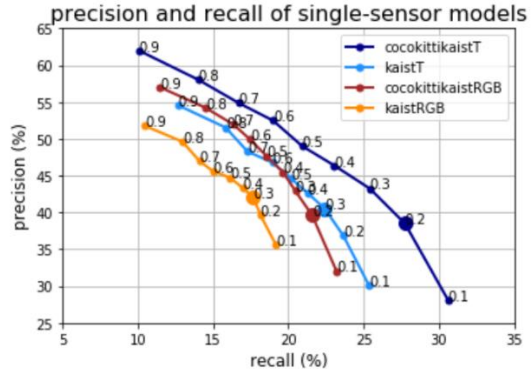


Figure 32 precision and recall of single-sensor models

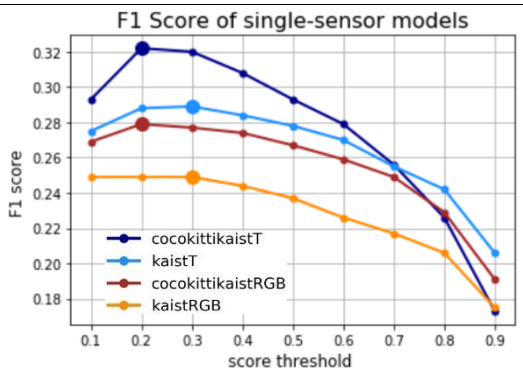


Figure 33 F1 score of single-sensor models

8.2 Comparison between multi-sensor-fusion models and single-sensor models

Figure 36 shows the precision and recall of six models. Figure 35 shows F1 score of six models. Figure 34 shows average IOU of six models. The observations found from these three models are described on this section.

Late fusion has the largest recall among the six models (Figure 36). With the same score threshold, the recall of late fusion is always larger than all the single-sensor models, as shown in Figure 36. This result fits with our anticipation. Because late fusion incorporates bounding boxes predicted from both T and RGB models. It means that late fusion model predicts the larger amount of bounding boxes than single sensor models.

Late fusion improves the performance of single-sensor models. This is shown in Figure 35. Late fusion has the largest F1 score among all the six models. Table 10 shows how much improvement that late fusion has made. Late fusion improves kaist T, cocokittikaistT, kaist RGB and cocokittikaist RGB by 7.2%, 3.9%, 11.2% and 8.2% respectively. The improvement value is calculated by the difference of F1 Score between the best performance of late fusion model and the best performance of other five models.

Compared with Model	Improvement made by late fusion
kaistT	7.2%
cocokittikaistT	3.9%
Kaist RGB	11.2%
Cocokittikaist RGB	8.2%

Table 10 Improvement that the best performance of late fusion has made on the single-sensor models

Early fusion has the worst performance among all the six models. F1 score of early fusion is the lowest among the six models. This is shown in Figure 35. The recall of early fusion is the smallest among the six models (shown in Figure 36). Low recall indicates that early fusion does not predict much bounding boxes. The reason is very likely to be that information loss happened due to the conversion from RGBT to HST. The band L (lightness) in HSL has been removed. This causes information loss in visual images. Therefore, it is difficult to extract sufficient human features in the HST images. Small amounts of bounding boxes are predicted.

Fusion models has higher minimum precision, compared with fusion models. This observation is found in both early and late fusion models. It is observed in Figure 36. In Table 11, the minimum precision value of each model is recorded. When score threshold is 0.1, the minimum precision of early fusion model is 41.35%. It is significantly larger than the precision of kaistT (30.06%), cocokittikaistT (28.07%), kaistRGB (35.59%) and cocokittikaistRGB (32.02%). The minimum precision of early fusion model (39.96%) is also larger than single-sensor models. This means that with a very low score threshold, the true positive predicted by fusion model is larger than that by a single-sensor model. The reason is that fusion models has higher accuracy in human detection.

It is also observed that the precision interval of fusion models is smaller than single sensor models. Precision interval is defined as the interval between minimum and maximum precision of a model. This means that the precision of fusion models are more stable than single-sensor models.

precision	Score threshold	kaistT	cocokittikaistT	kaistRGB	cocokittikaistRGB	Early fusion	Late fusion
Minimum	0.1	30.06%	28.07%	35.59%	32.02%	41.35%	39.96%
Maximum	0.9	54.55%	61.99%	51.73%	56.99%	52.23%	56.46%

Table 11 minimum and maximum precision of six models

Figure 34 shows the average IOU of six models. Average IOU increases with the score threshold increases. With increasing the score threshold, only the bounding boxes which has large confidence level are remained in the output. High-score bounding boxes are more accurate than the low-score bounding boxes. When score threshold is low, for example 0.1, the average IOU of early fusion model is the highest. This reflects what has been explained in the previous paragraph: early fusion has the largest precision among all six models, and cocokittikaistT has the smallest precision, when score threshold is 0.1.

I would like to compare my fusion models with the fusion models in literature research. However, there are two reasons that makes this difficult. First, those literatures do not provide the numbers of precision or recall. The graph that they show is unclear on what the exact value is. Secondly, the fusion models in literature uses different approaches to implement early and late fusion. Their fusion is executed in the layers of convolutional neural network. For example, their late fusion architecture integrates thermal and

visual subnetworks on a fully convolutional layer. The fusion that I implemented is data integration before training a CNN (early fusion) and decision integration after training a CNN (late fusion).

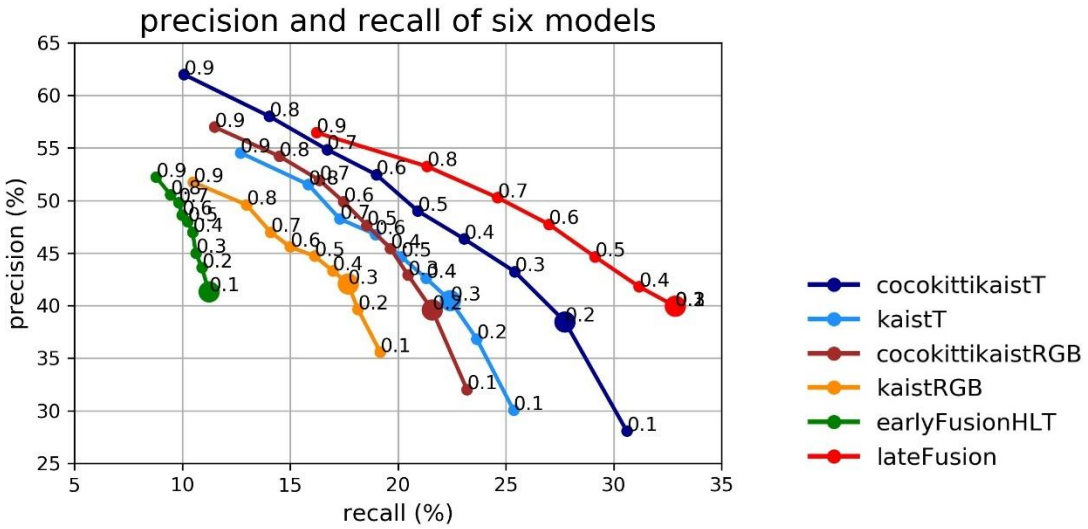


Figure 36 precision and recall of six modles

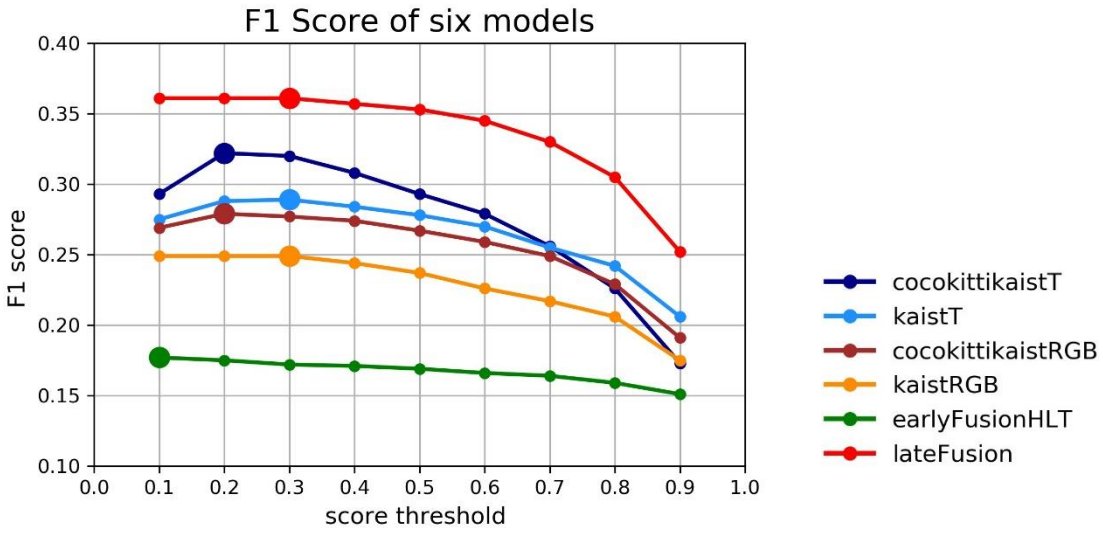


Figure 35 F1 score of six modles

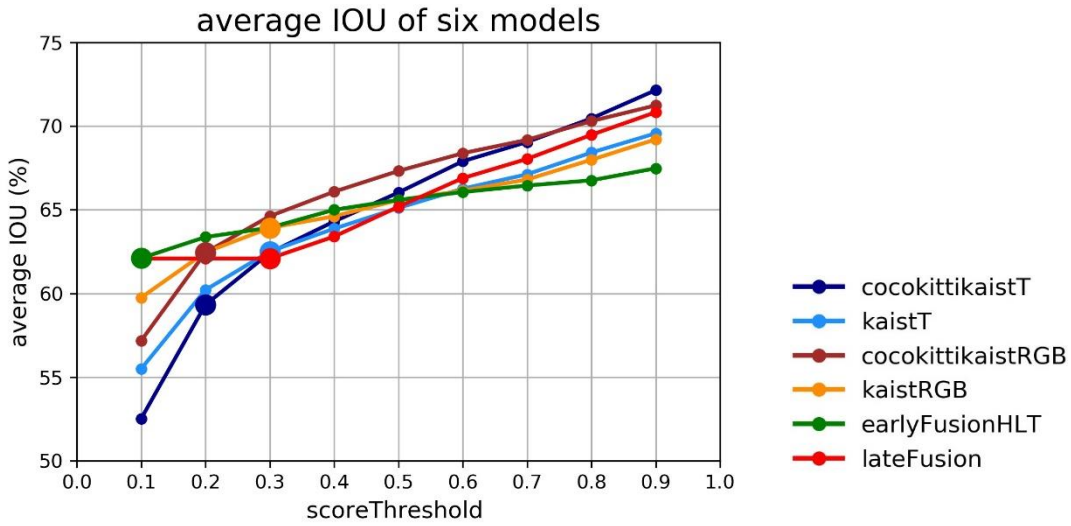


Figure 34 average IOU of six models

8. DISCUSSION

This section discusses the factors that inpact recall and the factors that impact precision. An example is provided for explaining each factors. In all the following the example graphs, the left image is in its the original size. The right image is the zoomed-in images. Right image emphasizes what the problem is.

8.1. Factors that impact recall

The recall in the six models are low. The factors that causes low recall are explained in this section.

8.1.1. Far Distance

When a person is far away from the camera, it is difficult for a model to extract its feature. This problem is very challenging in both visual and thermal images. In visual images (Figure 38), the human shape and colors are unclear. In thermal images (Figure 37), long distance causes the amount of radiation that is captured by thermal camera is low. Both RGB and thermal camera does not provide enough information for human detection. As a consequence, it is hard for a model to detect small and long-distance persons. This problem occurs very frequently in this research.

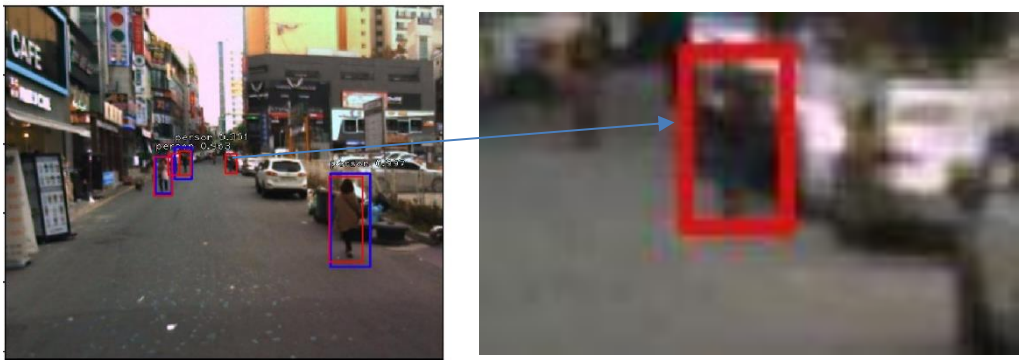


Figure 38 far distance persons on visual images

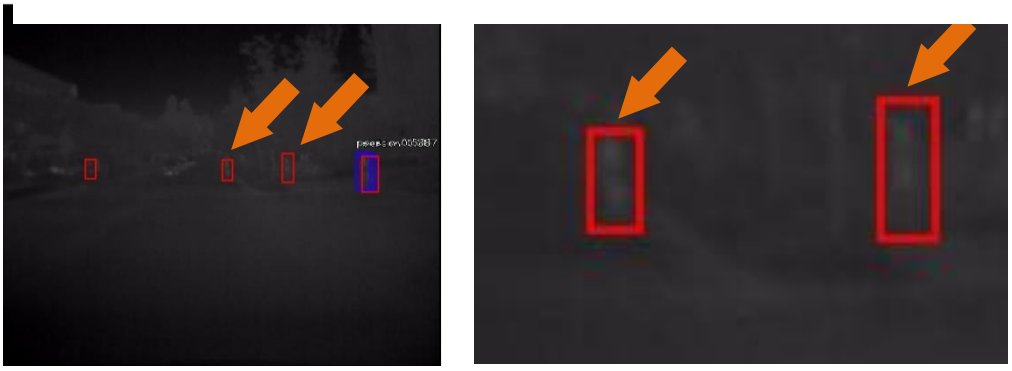


Figure 37 far distance persons on thermal images

8.1.2. Occlusion

When a person is occluded by an object, such as a car, sometimes the model does not detect it. Because human shape is incomplete. This causes recall to be low.

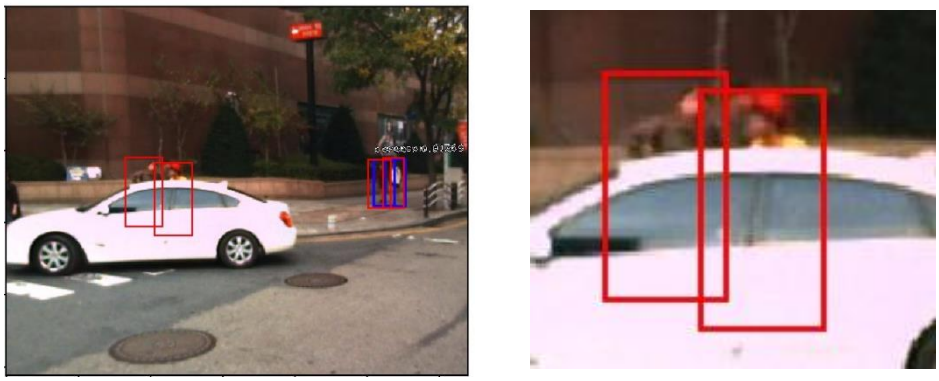


Figure 39 occlusion

8.1.3. Half of A Person

When only a part of a person is visible in an image, it is hard for the model to detect it (Figure 40). This is challenging for both visual and thermal images.

Figure 40 half of a person makes it difficult to detect aperson

8.1.4. Low Resolution

The resolution of images is low. Human features are unclear in the images. This happens in both visual and thermal images, especially during night time. Two sets of visual images has this problem.

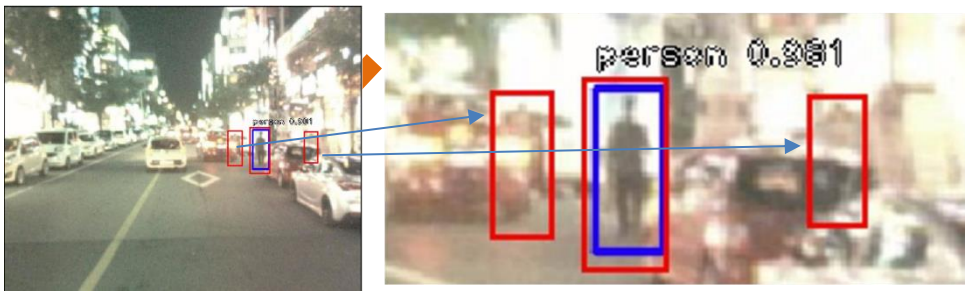


Figure 41 low resolution

8.1.5. Insufficient illumination (Under a tree or in a shadow)

Illumination is insufficient when a person is under a tree or in a shadow.

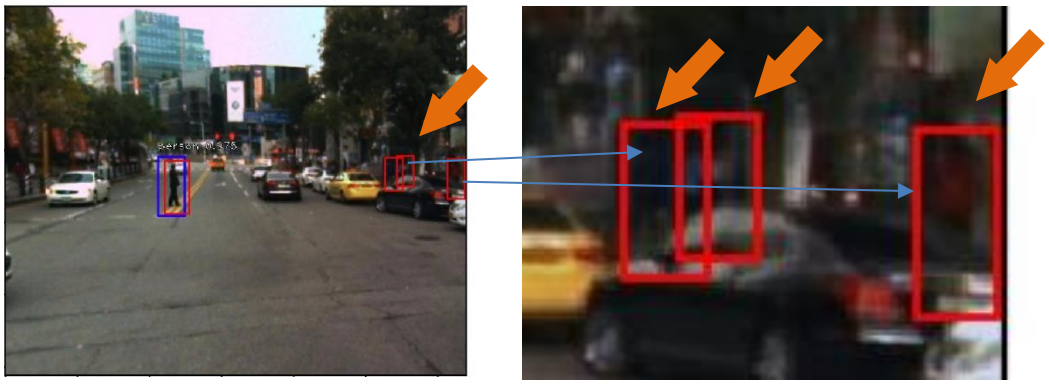


Figure 42 insufficient illumination

8.1.6. Over Explosion

When the illumination is too bright, like in Figure 43, human features would not be extracted.

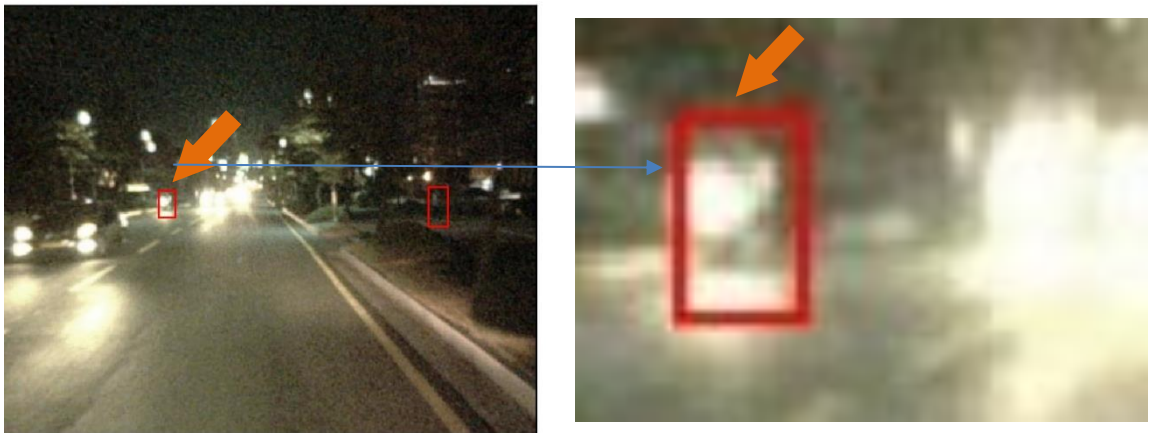


Figure 43 over explosion

8.1.7. Overfitting

In single-sensor models, overfitting occurs on all four models. Overfitting has an influence on the result of test. Take kaistT model as an example. Overfitting problem is found based on the fact that Kaist T convergence value is smaller than cocokittikaistT convergence value, while precision-recall graph shows that Kaist T has lower precision. This implies that KaistT has a problem of overfitting.

8.2. Factors that impact precision

8.2.1. IOU between predicted bounding box and annotation is smaller than true positive threshold.

Several factors could cause the problem that IOU is smaller than the true positive threshold. Firstly, the true positive threshold is set as very high. Secondly, there is a problem in the annotation. In some images, annotation groups several people together in one bounding box. So the annotated bounding box is super large. IOU become very small, even though the model correctly detect a person. Thirdly, annotation is larger than a person.

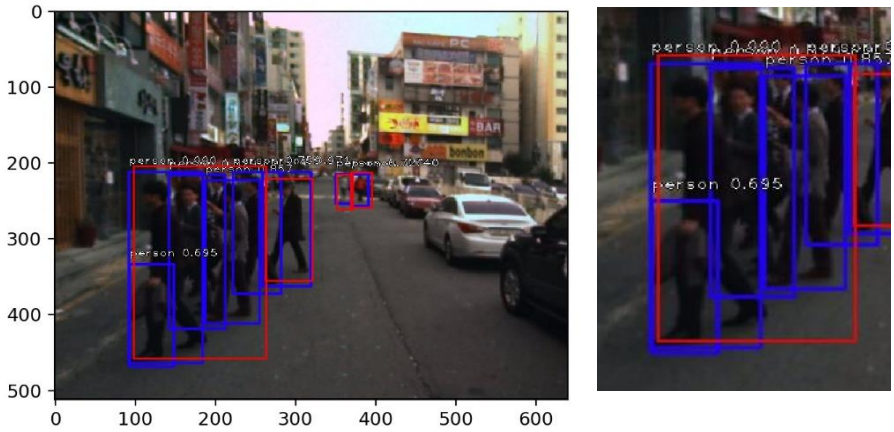
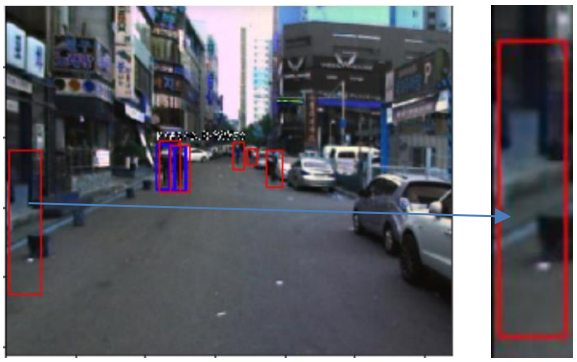


Figure 44 an annotation bounding box is much larger than a person.

8.2.2. A non-human object is annotated as a human.



8.3. Average IOU

Different definition of average IOU causes difference outcome value. This research defines average IOU as the sum of IOU divided by the total number of all predicted bounding box. All the predicted bounding boxes are taken into calculation. This causes a relatively low value of average IOU. Because some predicted bounding box does not intersect with any annotation. Its contribution to total IOU is zero. But it is counted in the denominator.

An alternative way to calculate average IOU is to only take true positives. Therefore, average IOU is equal to the total amount of IOU divided by the total number of all true positive bounding box. It means, among the true positives, the average IOU. It will generate higher value than the value shown in Figure x.

8.4. Computation complexity

It is better to use mathematical functions to calculate the computational complexity. This research has recorded the average time spent on predicting an image. However, two different hard wares are both used in this research. One is a server. The other is my personal laptop. The prediction time is largely influenced by the hardware. So, the computation time is not shown in this thesis.

9. CONCLUSION

The goal of this research is to find out the best deep learning model in human detection, with the input thermal and visual images. In this research, six deep learning models have been implemented. Four models are single-sensor models. They are trained by either thermal or visual images. Two models are multi-sensor fusion models. They are early fusion model and late fusion model. Early fusion model integrates thermal and visual images on a pixel level.

It is concluded that late fusion model surpasses the performance of all single-sensor models. Early fusion does not have significant improvement on single-sensor models, because its F1 score is lower than all single-sensor models. It is also observed that models finetuned with coco and kitti dataset has larger f1 score than the models which has only been trained on ImageNet.

Two fusion models in this research has both advantages and disadvantages. For the early fusion model, the advantage is that the weights of the network are trained. RetinaNet learns the features from both thermal and visual images. The disadvantage is that, information loss occurs when RGBT is converted to HST. For late fusion model, the advantage is that no information is lost. The bounding boxes predicted by single-sensor models are all taken into its consideration. The drawback is that when true positive increases, false positive may also increase. The total amount of prediction is larger than single-sensor models. This makes it very tricky to find the best merge threshold.

This research has addressed three research questions.

Research Question 1: What is the best human detection approach?

Currently, RetinaNet is the best approach in human detection. Based on the literature research, deep learning surpasses the traditional approaches. Among various deep learning approaches, RetinaNet is the best approach. The innovation of RetinaNet is that it solves two problems. Firstly, it applies a special function, called focal loss function to solve the class imbalance problem. Secondly, it has a unique archeticture Pyramid Feature Nework. This network solves the problem that high resolution images contributes to weak features while low resolution images contribute to strong features. This problem is not solved by other deep learning networks. Therefore, RetinaNet is the best approach in human detection.

The reason why deep learning outperforms traditional approaches is the following. Convolutional Neuron Network (CNN) is a highly efficient method, compared with other traditional methods like HOG (histogram of gradient) (Zhu et al., 2017). Because traditional methods use a dictionary to store all the human in the training data (Fischer, Herman, Behnke, & Systems, 2016). It has two drawbacks (Liu, Zhang, Wang, & Metaxas, 2016). Firstly, its speed is slow, because it compares with all the objects in the dictionary with the unknown object in the image. Secondly, if the algorithm sees a human which does not exist in the dictionary, the it is not able to detect it. In contrast, CNN uses feature extraction to capture the core of human shape.

Research Question 2: How to integrate multiple sensors in the state of art?

There are two architectures to integrate multiple sensors: early fusion and late fusion. In the application of human detection, early fusion means to integrate thermal and visual images on pixel level. Late fusion means to integrate on a decision level. Besides, halfway fusion is a third way of fusion. It is not implemented in this research. Because it is a very complicated and challenging task to modify the algorithm of RetinaNet. The methodology to design early and late fusion is the following.

In early fusion, the input RGB+T has been converted to HST (Hue Saturation Thermal). This is achieved by firstly converting RGB to HSL and then replacing lightness band by thermal band. Both of the training and testing process are implemented on HST images. RetinaNet was trained by learning human features from HST images.

In late fusion, prediction is made by incorporating the decision from single-sensor models. The input are bounding boxes predicted by the best model on visual images and the best model on thermal images. The next step is to optimize the value of merge threshold. When a person is detected by two bounding boxes, the largest score decides which bounding box is the final output.

Research Question 3: How much does multi-sensor approach improve the single sensor approach?

Late fusion significantly improves the accuracy of all four single-sensor models. The amount of improvement is 7.2% on kaistT, 3.9% on , cocokittikaistT, 11.2% on kaistRGB and 8.2% on cocokittikaist RGB. The improvement is achieved by integrate the predicted bounding boxes from thermal model and RGB model. Early fusion model does not improve single-sensor models. Both early and late fusion models improve the precision for low score threshold which are 0.1,0.2,0.3,0.4.

Future suggestion:

1) In terms of quality assessment, it is better to separate the assessment by the difficulty level of human detection. Human detection can be classified by difficult detection, middle-level detection and easy level detection. Difficult detection are the challenging cases, for example occlusion and insufficient illumination. Implementing the quality assessment on different difficulty-level of human would remove the outliers of accuracy assessment. It also help identifying which kind of human instances are the most challenging one for a model.

2) In early fusion, it is suggested to try another method to convert RGBT to a 3band images. This means replacing hue band by thermal band or replacing saturation band by thermal band. After training a network, we can compare these three different methods and find out which early fusion method is the best.

LIST OF REFERENCES

Afsar, P., Cortez, P., & Santos, H. (2015). Automatic visual detection of human behavior: A review from 2000 to 2014. *Expert Systems with Applications*, *42*(20), 6935–6956. https://doi.org/10.1016/J.ESWA.2015.05.023

Baek, J., Hong, S., Kim, J., & Kim, E. (2017). Efficient Pedestrian Detection at Nighttime Using a Thermal Camera. *Sensors*, *17*(8), 1850. https://doi.org/10.3390/s17081850

Balani, K., Deshpande, S., Nair, R., & Rane, V. (2015). Human detection for autonomous vehicles. *2015 IEEE International Transportation Electrification Conference (ITEC)*, 1–8. https://doi.org/10.1109/ITEC-India.2015.7386891

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2017). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. https://doi.org/10.1109/TPAMI.2018.2798607

Bharathi.V.S, G. (2005). Alive Human Detection in Disaster Zones using Manually Controlled Robots. *International Journal of Innovative Research in Computer and Communication Engineering*, *3*(2), 11–17. Retrieved from www.ijircce.com

Brunetti, A., Buongiorno, D., Trotta, G. F., & Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, *300*, 17–33. https://doi.org/10.1016/J.NEUCOM.2018.01.092

Correa, M., Hermosilla, G., Verschae, R., & Ruiz-del-Solar, J. (2012). Human Detection and Identification by Robots Using Thermal and Visual Information in Domestic Environments. *Journal of Intelligent & Robotic Systems*, *66*(1–2), 223–243. https://doi.org/10.1007/s10846-011-9612-2

D, G., Manjunath, & Abirami, S. (2012). Suspicious Human Activity Detection from Surveillance Videos. *(IJIDCS) International Journal on Internet and Distributed Computing Systems*, *2*(3). Retrieved from https://pdfs.semanticscholar.org/c3bc/90003193ad9c1973a3529b551ab8857ad589.pdf

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *Conference on Computer Vision and Pattern Recognition (CVPR)*, *1*, 886–893. https://doi.org/10.1109/CVPR.2005.177

Dollar, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1532–1545. https://doi.org/10.1109/TPAMI.2014.2300479

Du, X., El-khamy, M., Lee, J., & Davis, L. (2017). Fused DNN : A deep neural network fusion approach to fast and robust pedestrian detection. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 953–961. https://doi.org/10.1109/WACV.2017.111

Fan, X., Xu, L., Zhang, X., & Chen, L. (2008). The Research and Application of Human Detection Based on Support Vector Machine Using in Intelligent Video Surveillance System. In *2008 Fourth International Conference on Natural Computation* (pp. 139–143). IEEE. https://doi.org/10.1109/ICNC.2008.315

Girshick, R. (2015). Fast R-CNN. *The IEEE International Conference on Computer Vision (ICCV)*, 1440–1448. Retrieved from https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html

Girshick, R., Donahue, J., Darrell, T., Malik, J., & Berkeley, U. C. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *ARXIV*. Retrieved from http://arxiv.org/abs/1311.2524

Guan, D., Cao, Y., Yang, J., & Yang, M. Y. (2018). Fusion of Multispectral Data Through Illumination-

aware Deep Neural Networks for Pedestrian Detection. *ARXIV*. Retrieved from https://arxiv.org/pdf/1802.09972.pdf

Han, J., Zhang, D., Cheng, G., Liu, N., & Xu, D. (2018). Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection. *IEEE Signal Processing Magazine*, *35*(1), 84–100. https://doi.org/10.1109/MSP.2017.2749125

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *ARXIV*. Retrieved from http://arxiv.org/abs/1703.06870

Hosang, J., Omran, M., Benenson, R., & Schiele, B. (2015). Taking a Deeper Look at Pedestrians. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4073–4082. Retrieved from http://arxiv.org/abs/1501.05790

Hwang, S., Park, J., Kim, N., Choi, Y., & Kweon, I. S. (2015). Multispectral Pedestrian Detection : Benchmark Dataset and Baseline. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1037–1045. https://doi.org/10.1109/CVPR.2015.7298706

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale Video Classification with Convolutional Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 1725–1732. https://doi.org/10.1109/CVPR.2014.223

Kim, J. H., Hong, H. G., & Park, K. R. (2017). Convolutional Neural Network-Based Human Detection in Nighttime Images Using Visible Light Camera Sensors. *Passaro VMN, Ed. Sensors (Basel, Switzerland)*, *17*(5), 1065. https://doi.org/10.3390/s17051065

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2017). Scale-aware Fast R-CNN for Pedestrian Detection. *IEEE Transactions on Multimedia*, 1–10. https://doi.org/10.1109/TMM.2017.2759508

Lin, T., Goyal, P., Girshick, R., He, K., & Piotr Dollar. (2018). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2018.2858826

Liu, J., Zhang, S., Wang, S., & Metaxas, D. N. (2016). Multispectral Deep Neural Networks for Pedestrian Detection, 1–13. Retrieved from http://arxiv.org/abs/1611.02644

Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., & Wang, X. (2018). Deep learning for pixel-level image fusion : Recent advances and future prospects. *Information Fusion*, *42*, 158–173. https://doi.org/10.1016/j.inffus.2017.10.007

Mitra, V., Vanhout, J., Wang, W., Bartels, C., Franco, H., Vergyri, D., … Morgan, N. (2016). Fusion Strategies for Robust Speech Recognition and Keyword Spotting for Channel- and Noise-Degraded Speech. *INTERSPEECH 2016*, 3683–3687. https://doi.org/10.21437/Interspeech.2016-279

Moore, D. (2003). A real-world system for human motion detection and tracking. Retrieved from http://www.vision.caltech.edu/~dmoore/dmoore-final-thesis.pdf

Nam, W., Dollar, P., & Hee Han, J. (2014). Local Decorrelation for Improved Pedestrian Detection. *ARXIV*, 1–9. Retrieved from https://papers.nips.cc/paper/5419-local-decorrelation-for-improved-pedestrian-detection.pdf

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal Deep Learning, 1–9. Retrieved from https://people.csail.mit.edu/khosla/papers/icml2011_ngiam.pdf

Niels Gerlif, M. (2013). Visual Detection of Humans in a Disaster Scenario. Retrieved from es.aau.dk

Paisitkriangkrai, S., Shen, C., & Hengel, A. Van Den. (2014). Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features, 546–561. Retrieved from http://arxiv.org/abs/1407.0786

Powers, D. M. W. (2007). Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation. Retrieved from http://david.wardpowers.info/BM/index.htm.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once : Unified , Real-Time Object Detection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. *ARXIV*, 1–14. Retrieved from http://arxiv.org/abs/1506.01497

Reyes-Ortiz, J. L., Oneto, L., Samà, A., Parra, X., & Anguita, D. (2016). Transition-Aware Human Activity Recognition Using Smartphones. *Neurocomputing*, *171*, 754–767. https://doi.org/10.1016/j.neucom.2015.07.085

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *ARXIV*, 1–9. Retrieved from http://arxiv.org/abs/1406.2199

Uijlings, J. R. ., Sande, K. E. . Van De, Sande, T., & Smeulders, A. W. M. (2012). Selective Search for Object Recognition. *International Journal of Computer Vision*, *104*(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5

Wagner, J., Fischer, V., Herman, M., & Behnke, S. (2016). Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. *ESANN 2016 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 27–29. Retrieved from http://www.i6doc.com/en/

Wu, Q., Shen, C., Wang, P., Dick, A., & Van Den Hengel, A. (2016). Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. *ARXIV*. Retrieved from https://arxiv.org/pdf/1603.02814.pdf

Yang, B., Yan, J., Lei, Z., & Stan Z. Li. (2014). Aggregate Channel Features for Multi-view Face Detection. *International Joint Conference on Biometrics*. Retrieved from http://arxiv.org/abs/1407.4023

Zhang, S., Bauckhage, C., & Cremers, A. B. (2014). Informed Haar-like Features Improve Pedestrian Detection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 947–954. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2014/html/Zhang_Informed_Haar-like_Features_2014_CVPR_paper.html

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Zollhöfer, M., Stotko, P., Görlitz, A., Theobalt, C., Nießner, M., Klein, R., & Kolb, A. (2018). State of the Art on 3D Reconstruction with RGB-D Cameras. *Computer Graphics Forum*, *37*(2), 625–652. https://doi.org/10.1111/cgf.13386

APPENDIX

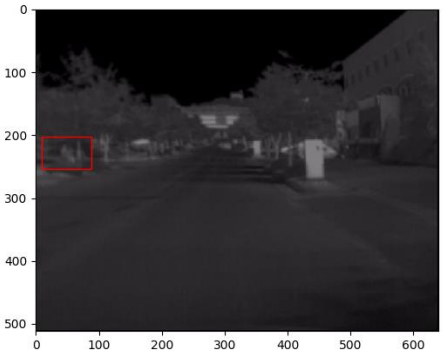
Appendix 1 Confusion matrices of 5-class show accuracy for each class (row) and each row is summed up to 1.

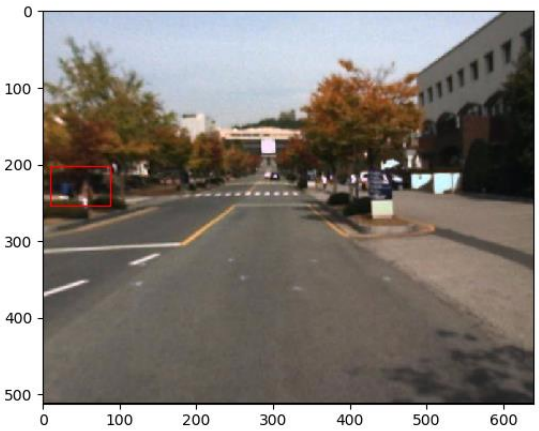
Appendix 1

Four problems in annotation (written by Xinran Wang)
The blue rectangle is model detected people. the red rectangle is annotations. The yellow rectangle is used to show the person
Problem 1 here is a person but annotation doesn't give the bounding box.

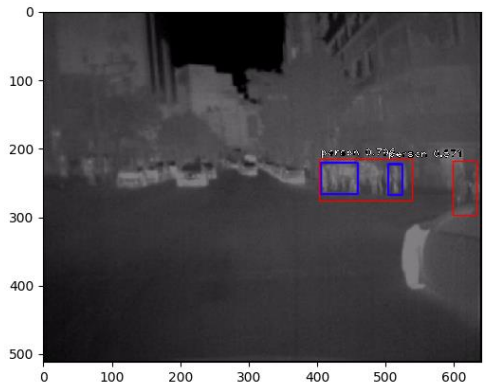
The blue rectangle is model detected people. the red rectangle is annotations. The yellow rectangle is used to show the person

Problem 2 there is no person but annotations give a bounding box

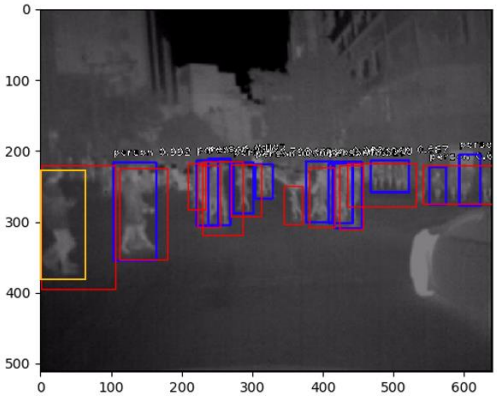




Problem 3 the annotations give a bounding box which contains a group of people rather than a single person.

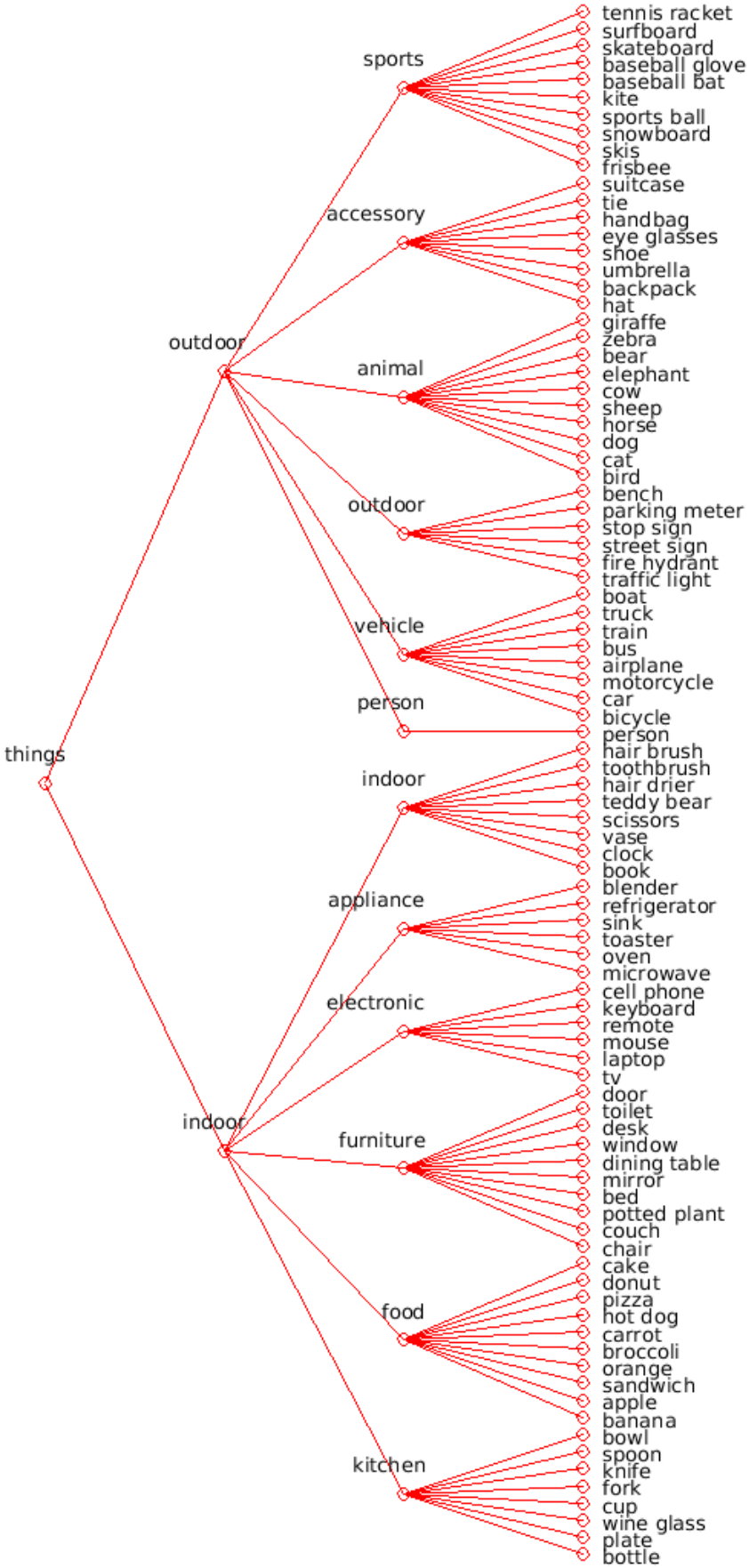


Problem 4 sometimes the annotations give a bounding box larger than the real person which are represented by the yellow bounding boxes.


0
100
200
300
400
500
0
100
200
300
400
500
600


0
100
200
300
400
500
0
100
200
300
400
500
600

Appendix 2


things
outdoor
indoor
sports
accessory
animal
outdoor
vehicle
person
indoor
appliance
electronic
furniture
food
kitchen
tennis racket
surfboard
skateboard
baseball glove
baseball bat
kite
sports ball
snowboard
skis
frisbee
suitcase
tie
handbag
eye glasses
shoe
umbrella
backpack
hat
giraffe
zebra
bear
elephant
cow
sheep
horse
dog
cat
bird
bench
parking meter
stop sign
street sign
fire hydrant
traffic light
boat
truck
train
bus
airplane
motorcycle
car
bicycle
person
hair brush
toothbrush
hair drier
teddy bear
scissors
vase
clock
book
blender
refrigerator
sink
toaster
oven
microwave
cell phone
keyboard
remote
mouse
laptop
tv
door
toilet
desk
window
dining table
mirror
bed
potted plant
couch
chair
cake
donut
pizza
hot dog
carrot
broccoli
orange
sandwich
apple
banana
bowl
spoon
knife
fork
cup
wine glass
plate
bottle

Appendix 3 convert RGB to HSL

R, G, and B are converted to the floating-point format and scaled to fit the 0 to 1 range.

$$V_{max} \leftarrow max(R, G, B)$$

$$V_{min} \leftarrow min(R, G, B)$$

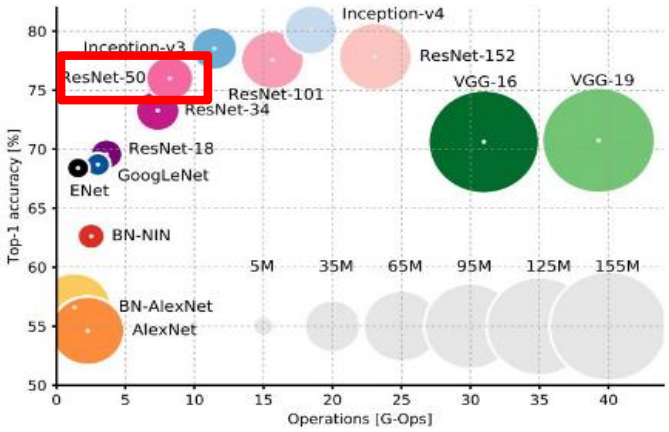
$$L \leftarrow \frac{V_{max} + V_{min}}{2}$$

$$S \leftarrow \begin{cases} \frac{V_{max} - V_{min}}{V_{max} + V_{min}} & \text{if } L < 0.5 \\ \frac{V_{max} - V_{min}}{2 - (V_{max} + V_{min})} & \text{if } L \geq 0.5 \end{cases}$$

$$H \leftarrow \begin{cases} 60(G - B)/(V_{max} - V_{min}) & \text{if } V_{max} = R \\ 120 + 60(B - R)/(V_{max} - V_{min}) & \text{if } V_{max} = G \\ 240 + 60(R - G)/(V_{max} - V_{min}) & \text{if } V_{max} = B \end{cases}$$

If $H < 0$ then $H \leftarrow H + 360$. On output $0 \leq L \leq 1$, $0 \leq S \leq 1$, $0 \leq H \leq 360$.

Appendix 4



Appendix 5 mean average precision

$$AP = \sum_{i=1}^{n} Precision_i \cdot \Delta Recall_i$$

Equation 1equation of average precision [30]. $Precision_i$ is a percentage of correct items among first i recommendations. $\Delta Recall_i$ equals $1/n$ if i^{th} item is correct and 0 otherwise.

[30] https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

Equation 2 equation of mean average precision[31] . Q is the number of queries

[31] https://www.quora.com/How-can-I-measure-the-accuracy-of-a-recommender-system